

AIR POLLUTION AND MORTALITY: AN INVESTIGATION INTO THE LAG STRUCTURE BETWEEN EXPOSURE TO AIR POLLUTION, AMBIENT TEMPERATURE AND MORTALITY FROM PNEUMONIA, CHRONIC OBSTRUCTIVE PULMONARY DISEASE, & ISCHAEMIC HEART DISEASE.

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy (PhD) in the Faculty of Biology, Medicine and Health.

2016

Matthew Paul Gittins

School of Health Sciences

Supervisors: Prof Roseanne McNamee & Prof Raymond Agius

TABLE OF CONTENTS

LIST OF TABLES	8
LIST OF FIGURES	11
GLOSSARY.....	14
ABSTRACT.....	18
DECLARATION	19
COPYRIGHT STATEMENT	20
PRESENTATIONS RELATING TO THIS THESIS	21
PUBLICATIONS RELATING TO THIS THESIS	22
ACKNOWLEDGEMENTS	23
1 INTRODUCTION	24
1.1 Chapter Introduction	24
1.1.1 Thesis structure	24
1.2 Background	26
1.3 Exposure data – Atmospheric air pollution.....	30
1.3.1 Fixed pollution monitoring networks.....	31
1.3.2 Potential pollutants of interest.....	32
1.3.3 Pollutants their measurements and influence.....	33
1.4 Exposure data – Meteorological factors.....	39
1.4.1 Manual & automated meteorological monitoring.....	39
1.5 Pollution & meteorological fixed site monitor details.....	42
1.5.1 Background stations	42
1.5.2 Hotspots or kerbside stations	43
1.6 Improving individual exposure estimates	43
1.7 Adverse outcomes – Influence of exposure in humans.....	45
1.7.1 Potential biological mechanisms - Pollution.....	45
1.7.2 Potential biological mechanisms - Temperature	47
1.7.3 Communicable Diseases - Pneumonia.....	48
1.7.4 Non-communicable disease – COPD.....	50
1.7.5 Non-Communicable Disease – Ischaemic Heart Disease (IHD)	53
1.8 Defining the cause of death.....	55

2	INTRODUCTION TO ANALYSIS METHODS.....	60
2.1	Chapter introduction	60
2.2	Traditional study design.....	60
2.3	Traditional study design vs Time series study design.....	63
2.4	Data characteristics	64
2.4.1	Confounders	65
2.5	Time series designs	66
2.5.1	Poisson regression model.....	66
2.5.2	Generalised Additive Models (GAM).....	69
2.5.3	Generalised Linear Models (GLM).....	72
2.6	Case-crossover design.....	74
2.6.1	Control/referent selection strategies.....	75
2.6.2	Unidirectional control selection	77
2.6.3	Bidirectional control selection	78
2.6.4	Semi-symmetric bidirectional controls	80
2.6.5	Overlap bias	81
2.6.6	Time stratified control selection	82
2.7	Choosing the time-series model or the case-crossover design?.....	83
2.8	Modelling exposure 1: The exposure-response relationship.....	85
2.9	Modelling exposure 2: The lagged relationship.....	88
2.9.1	Defining the ‘lag’ period.....	88
2.9.2	Unconstrained distributed Lag	89
2.9.3	Lag stratified models.....	90
2.9.4	Constrained distributed lag	91
2.9.5	Penalised spline.....	93
2.10	Modelling exposure 3: Exposure-response and lagged relationship.....	94
2.10.1	Distributed lag non-linear models.....	94
2.11	Harvesting/mortality displacement	95
2.11.1	Distributed lag methods in harvesting	98
2.12	Missing data	100
2.12.1	Introduction.....	100
2.12.2	Exposure data characteristics & missing data implications	100
2.12.3	Missing data processes & definitions.....	102

2.12.4	Missing data characteristics and patterns in time-dependent data	104
2.12.5	Modelling with missing data present - Complete Cases	105
2.12.6	Modelling with missing data present - Imputation	106
2.12.7	Modelling with missing data present - Mean imputation	106
2.12.8	Modelling with missing data present - regression model	107
2.12.9	Modelling with missing data present - regression model with random variation	107
2.12.10	Modelling with missing data present - Multiple imputation	108
2.12.11	Multiple imputation methods – Univariate ‘imputation model’	111
2.12.12	Multiple imputation methods - Multivariate ‘imputation model’	112
2.12.13	Multiple imputations by chained equations (MICE)	112
2.12.14	Multiple imputation - Joint modelling.....	113
2.12.15	Multiple imputation considerations.....	114
2.12.16	Multiple imputation – with multilevel data.....	117
2.13	Aims and objectives of the thesis.....	119
2.13.1	Overall aim.....	119
2.13.2	Specific objectives	119
3	STUDY METHODS AND MATERIALS	120
3.1	Introduction to methods and materials.....	120
3.2	Study population and study period.....	120
3.3	Participant data recruitment, administration, and acquisition.....	121
3.3.1	Participant consent	122
3.3.2	Ethical & security considerations	122
3.3.3	Exclusion criteria	124
3.3.4	Expected sample size	124
3.3.5	Power calculation	125
3.4	Independent exposure data.....	126
3.4.1	Accessing exposure data	126
3.4.2	Defining the exposure data.....	127
3.4.3	Monitor identification criteria.....	128
3.4.4	Exposure parameters - Construction	130
3.5	Study outcome or dependent variable	135
3.6	Study design.....	135

3.6.1	Data manipulation	136
3.7	Analysis with missing exposure data	138
3.7.1	Imputation model – The data structure and primary variable	138
3.7.2	Imputation model - Method	140
3.7.3	Imputation model – Primary model predictors	141
3.7.4	Imputation model – Additional covariates.....	143
3.7.5	Imputation model - Summary	144
3.8	Simulation study - Missing exposure data	145
3.8.1	Simulation study - Introduction	145
3.8.2	Simulation study – Data.....	146
3.8.3	Simulation study – Missing data characteristics	146
3.8.4	Simulation study – Procedure & performance	148
3.9	Statistical analysis of the main study	151
3.9.1	Analysis of the hospital admission moderator	153
3.9.2	Analysis of outliers and missing data	153
3.9.3	Analysis procedure of the main data.....	154
4	RESULTS – INTRODUCTION.....	156
4.1	Data background - Summary of mortality data.....	156
4.2	Summary of pollution exposure data	159
4.3	Summary of temperature exposure data.....	162
5	RESULTS – INITIAL PNEUMONIA INVESTIGATION	164
5.1	Introduction and differences in the methods.....	164
5.2	Results associated with Pneumonia mortality analysis.....	165
6	RESULTS - SIMULATION STUDY	170
6.1	Introduction to the simulation datasets and the true effect estimates	170
6.2	Simulating the missing data characteristics	171
6.3	Simulation study results - MCAR.....	175
6.4	Simulation study results - MAR.....	177
7	RESULTS – MODELLING TEMPERATURE	184
7.1	Investigating temperature – mortality in Pneumonia, COPD, IHD	184
7.2	Bespoke temperature modelling for Pneumonia, COPD, and IHD	186
7.3	Comparing the temperature effects between causes of death	190

7.4	Influence of hospital admission moderator on temperature effect.....	192
7.4.1	Adjusting for hospital admission exposure – Pneumonia mortality	192
7.4.2	Adjusting for hospital admission during exposure – COPD mortality	194
7.4.3	Adjusting for hospital admission during exposure – IHD mortality.....	196
8	RESULTS – MAIN STUDY ANALYSIS OF POLLUTION.....	200
8.1	Investigating pollution–mortality: Pneumonia, COPD, and IHD	200
8.2	Modelling lagged pollution effect: Pneumonia, COPD, and IHD	205
8.3	Accounting for hospital admission during pollution exposure	210
8.3.1	Pollution-pneumonia mortality: accounting for hospital status.	210
8.3.2	Pollution-COPD mortality: accounting for hospital status.	213
8.3.3	Pollution-IHD mortality: accounting for hospital status.....	216
8.4	Influence of outliers and missing data	219
8.4.1	Outliers and missing data - Pneumonia.....	220
8.4.2	Outliers and missing data - COPD.....	223
8.4.3	Outliers and missing data – Ischaemic Heart Disease	226
8.5	Sensitivity analysis – Analysis and data checks	230
9	DISCUSSION.....	238
9.1	Introduction	238
9.2	Strengths.....	240
9.3	Reducing bias	245
9.3.1	Misclassification of exposure due to hospital admission.....	245
9.3.2	Missing data	254
9.3.3	Outliers – High pollution episodes.....	262
9.4	Results summary	266
9.4.1	Temperature	266
9.4.2	Pollution	273
9.5	Limitations - Additional causes of bias.....	285
9.5.1	Misclassification in the exposure assignment.....	285
9.5.2	Misclassification in the cause of death assignment	291
9.5.3	Further confounding – Multiple pollutants	293
9.5.4	Multiple testing	294
9.5.5	The exposure metric.....	295
9.6	Potential future work.....	296

9.7	Conclusion	298
	APPENDIX.....	333
Appendix A	Full description of the data collection procedure	333
Appendix B	Published original research paper.....	336
Appendix C	Background details regarding the exposure data.....	357
Appendix D	Supplementary analysis results	362
Appendix E	Example simulation study stata code	370

Word Count: 78,857

LIST OF TABLES

Table 3.1 - Harrell suggested percentiles for knot positions within a data range. ³²⁸	131
Table 3.2 - Generated example of the data structure	137
Table 3.3 - Example of the exposure dataset in preparation for multiple imputation analysis.....	139
Table 3.4 – Sample size calculation outlining the No. simulations required given the percentage bias in the missing data simulation study	149
Table 4.1 – Number of deaths from Pneumonia, Chronic Obstructive Pulmonary Disease (COPD), and Ishaemic Heart Disease (IHD), split by location, age, gender, and prior number of exposure days in hospital.....	157
Table 4.2 – Number of cause specific deaths observed in the data (1980-2011) compared with expected number of deaths based on ISD reported years (1991-2011).	158
Table 4.3 - Summary statistics for pollution monitors used within the study	160
Table 4.4 – Pairwise correlation coefficients comparing the five pollutants and average temperature modelled in the study.	161
Table 4.5 – Comparison statistics for available temperature monitors in Glasgow and Inverness based on 9am measurements.....	162
Table 4.6 – Daily (7am – 11pm) summary statistics for the temperature monitors used in this study within each city.....	163
Table 5.1 - Descriptive statistics of exposure and subjects for black smoke (BS) air pollution within Edinburgh between Jan 1981 and March 1996.	166
Table 5.2 – Results of the lag stratified black smoke air pollution percent change in risk for Pneumonia mortality, for all subjects, and split by community based only subjects and non-community based subjects only.	168
Table 6.1 - Description of the number (%) missing data for each pollutant and monitor in the study	171
Table 6.2 – Number (%) missing days for each pollution and monitor split by season (Winter/Summer)	172
Table 6.3 - Summary of the MAR and MCAR scenarios applied in simulation study	174
Table 6.4 – Bias reported in pollution simulation results for complete cases (CC) and multiple imputation (MI) analysis under missing completely at random (MCAR) characteristics.....	176

Table 6.5 – Bias reported in pollution simulation results for complete cases (CC) and multiple imputation (MI) analysis under missing at random (MAR) characteristics season and blocking	178
Table 6.6 – Bias reported in pollution simulation results for complete cases (CC) and multiple imputation (MI) analysis under combined missing at random (MAR) characteristics season and blocking	179
Table 6.7 - Temperature simulation results for Complete Cases (CC) and Multiple Imputation (MI) under Missing at Random characteristics season and blocking.....	181
Table 6.8 - Temperature simulation results for Complete Cases (CC) and Multiple Imputation (MI) under combined Missing at Random characteristics season and blocking.....	182
Table 7.1 – Percentage relative risk (%RR) associated with an increase in 1°C within each temperature zone associated with the lag stratified analysis bespoke for each cause of death.....	187
Table 7.2 – Cause of death specific percentage relative risk (%RR) associated with 1°C increase within each temperature zone (identified by all-cause of death), with comparison test result of three causes of death by lag period.....	191
Table 7.3 – Pneumonia specific percentage relative risk (%RR) and comparison test of hospital admission status during exposure (Zero, 1-29, all 30 days) associated with an increase in 1°C within each temperature zone	193
Table 7.4 - COPD specific percentage relative risk (%RR) and comparison test of hospital admission status during exposure (Zero, 1-29, all 30 days) associated with an increase in 1°C within each temperature zone.	195
Table 7.5 – Ischaemic heart disease specific percentage relative risk (%RR) and comparison test of hospital admission status during exposure (Zero, 1-29, all 30 days) associated with an increase in 1°C within each temperature zone.....	197
Table 8.1 – Cause of death specific percentage relative risk (%RR) associated with 10µgm ⁻³ increase per pollutant reported by lag stratified analysis, with comparison test result between three causes of death by lag period.....	206
Table 8.2 – Percentage relative risk (%RR) associated with an increase in 10µgm ⁻³ within each pollutant associated with the lag stratified analysis for Pneumonia, split by hospital admission status during exposure (Zero, 1-29, and All 30 days).....	211

Table 8.3 - Percentage relative risk (%RR) associated with an increase in $10\mu\text{gm}^{-3}$ within each pollutant associated with the lag stratified analysis for COPD, split by hospital admission status during exposure (Zero, 1-29, and All 30 days).	215
Table 8.4 – Percentage relative risk (%RR) associated with an increase in $10\mu\text{gm}^{-3}$ within each pollutant associated with the lag stratified analysis for IHD, split by hospital admission status during exposure (Zero, 1-29, and all 30 days).	217
Table 8.5 - Percentage relative risk (%RR) associated with an increase in $10\mu\text{gm}^{-3}$ within each pollutant associated with the lag stratified analysis for Pneumonia, results of the analysis investigating outliers and missing data.	221
Table 8.6 – Percentage relative risk (%RR) associated with an increase in $10\mu\text{gm}^{-3}$ within each pollutant associated with the lag stratified analysis for COPD, results of the analysis investigating outliers and missing data.	224
Table 8.7 – Percentage relative risk (%RR) associated with an increase in $10\mu\text{gm}^{-3}$ within each pollutant associated with the lag stratified analysis for IHD, results of the analysis investigating outliers and missing data.	227
Table 8.8 – Results (%RR) associated with of a sensitivity analysis attempting to replicate previous work (Carder et al (2008)) done on the similar dataset.	232
Table 8.9 –Percentage relative risk (%RR) associated with increase in black smoke within each individual monitor used in the main analysis.	234
Table 9.1 – The UK legislated air quality standard, ¹⁴ the DEFRA recommended limits, ¹⁷ and the 99th percentile position for each pollutant within the study.	263

LIST OF FIGURES

Figure 1.1 - Thesis structure	25
Figure 1.2 – Basic graphical representation of the comparison of mortality rates and pollution levels for the London Smog of 1952 (Wilkins E.T. 1954) ⁵	27
Figure 1.3 - Example cause of death statement from the ICD-10 Manual ¹⁶⁹	56
Figure 2.1 - Traditional study design vs Time dependent study design	64
Figure 2.2 – Basic representation of the case-crossover design over time	74
Figure 2.3 – Graphical description of the unidirectional and full stratum unidirectional case-crossover designs where X is the case day.	78
Figure 2.4 – Example of how long-term trends would cause bias in a unidirectional case-crossover design when exposure is consistently different in control days compared to the case day.	78
Figure 2.5 - Graphical description of the bidirectional and full stratum bidirectional case-crossover designs where X is the case day.	80
Figure 2.6 - Example month of a subject identifying their case day (X) on their ‘day of death’ and control days (O) under the time-stratified case-crossover design.	83
Figure 2.7 - A simple ‘harvesting’ model showing continuous transition of subjects from health general population to frail sub-group leading up to death	96
Figure 2.8 - A ‘harvesting’ model showing transition of subjects from health general population to frail sub-group (with the possibility of return) leading to death.....	97
Figure 2.9 - The likely change in mortality rate from the Expected Mortality Rate (Red dashed Line) when in the presence of short-term harvesting conditions leading to the rebound effect.	98
Figure 2.10 - Histogram of black smoke air pollution measurements from Edinburgh 1979-2009	101
Figure 3.1 – Fabricated example of monitor ‘active periods’ for seven monitors in Edinburgh (Ed) and Glasgow (Glas).....	128
Figure 3.2 - Example of the 30 day exposure period (0=day of death, X=exposure period)	130
Figure 3.3 - Graphical representation of several example threshold models with the knot point moving along the temperature exposure range.	133
Figure 3.4 - Example representations of exposure data distribution.....	139

Figure 3.5 – Example of the change in the pollution levels for differing monitors over time.....	141
Figure 4.1 - Graphical representation of the periods when each black smoke monitors employed in the study are active.	161
Figure 5.1 – Plot of the quadratic distributed lag model for all pneumonia (AP) and community deaths from pneumonia (CDP).	169
Figure 6.1 - Graphical display illustrating the distributional properties of cumulative days with missing data designated as ‘blocks’ of missing days. Note all monitors combined.	174
Figure 7.1 – Simultaneous plotting across the temperature range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields in a distributed lag non-linear model.....	185
Figure 7.2 – Simultaneously plotting across the temperature range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields using a cubic distributed lag model.	189
Figure 8.1 – Simultaneously plotting across the ‘black smoke’ range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields in a distributed lag non-linear model.....	201
Figure 8.2 – Simultaneously plotting across the ‘Particulate Matter’ range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields in a distributed lag non-linear model.....	203
Figure 8.3 – Simultaneously plotting across the ‘Gaseous Pollutants’ range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields in a distributed lag non-linear model.....	204
Figure 8.4 – Plotting the change in mortality risk described by a cubic distributed lag model associated with a $10\mu\text{gm}^{-3}$ increase in black smoke on pneumonia, COPD, and IHD mortality reported in any cause of death fields.	207
Figure 8.5 - Plotting the change in mortality risk described by a cubic distributed lag model associated with a $10\mu\text{gm}^{-3}$ increase in particulate matter concentrations on pneumonia, COPD, and IHD mortality reported in any cause of death fields.....	208
Figure 8.6 - Plotting change in mortality risk described by a cubic distributed lag model associated with a $10\mu\text{gm}^{-3}$ increase in gaseous pollutant concentrations on pneumonia, COPD, and IHD mortality reported in any cause of death fields.....	209

Figure 8.7 – Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on pneumonia split by hospital admission during exposure (zero, 1-29, and all 30 days).....212

Figure 8.8 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on COPD split by hospital admission during exposure (zero, 1-29, and all 30 days)..... 214

Figure 8.9 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on IHD split by hospital admission during exposure (zero, 1-29, and all 30 days)..... 218

Figure 8.10 – Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on pneumonia mortality, repeated for analyses investigating outliers and missing data.222

Figure 8.11 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on COPD mortality, repeated for analyses investigating outliers and missing data.225

Figure 8.12 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on IHD mortality, repeated for analyses investigating outliers and missing data..... 228

Figure 8.13 – Monitor specific cubic distributed lag plots for increases in black smoke on risk of cause specific mortality (Pneumonia, COPD, and IHD) using all study monitors.235

Figure 8.14 - Monitor specific cubic distributed lag plots for increases in black smoke on risk of cause specific mortality (Pneumonia, COPD, and IHD) using four largest study monitors only..... 236

GLOSSARY

%RR	The percentage change in Relative Risk
AFCOD	Any ‘cause of death’ field including primary and all secondary cause of death data fields from the mortality data
AIC	Akaike Information Criteria.
Alveoli	Small sacs located at end of branching system within the lungs.
AP	All Source pneumonia, i.e. CAP and HAP combined.
APHEA	Air Pollution and Health: a European Approach. A large multi-city European study
Atherosclerosis	Thickening or hardening of the artery walls.
AUC	Area under the Curve
BADC	British Atmospheric Data Centre
BIC	Bayesian Information Criteria
BS	Black smoke air pollution
CAP	Community Acquired pneumonia
CDP	Community deaths from pneumonia, a subgroup of CAP defined as those subjects in the community for the entire exposure period
CO	Carbon monoxide
Coagulation	Process by which the blood begins to clot.
COD	Cause of Death
COEH	Centre for Occupational and Environmental Health
Confounder	Secondary variable that influences both the independent and dependent variables
COPD	Chronic Obstructive Pulmonary Disease, formerly chronic bronchitis and emphysema
DEFRA	Department for Environmental Food and Rural Affairs
Epithelium	Cell that line surfaces of the human body, such as the lung wall
FEV	Forced Expiratory Volume
FMI	Fraction of Missing Information
Frail Group	A vulnerable sub-group of the population
FVC	Forced Vital Capacity
GAM	Generalised Additive Model
GIS	Geographic Information Systems

GLM	Generalised Linear Model
GRO	General Registry Office
Haemostatic	Processes that stop bleeding.
HAP	Hospital acquired pneumonia
Harvesting	Deaths occurring earlier than expected due external influence
HES	Health Episodes Statistics
HSCIC	The Health and Social Care information Centre
ICD	WHO International Statistical Classification of Diseases and Related Health Problems
IHD	Ischaemic (Ischemic) Heart disease also known as Coronary Heart Disease
Immediate COD	The final cause of death in the chain that led to death
ISD	The Information Services Division of Scotland
JM	Joint Modelling
Lag	Period after exposure with increased risk
Lag (1-n)	Period 1-n days post exposure
Lag 0	Same day of exposure
Lag 1	Day after exposure
LOESS	Local regression modelling, a non-parametric smoothing technique using k-nearest neighbour weighted data points to fit a regression model to predict n distinct data points
Macrophages	Type of cell that removes unwanted material in the body
MAR	Missing At Random
MCAR	Missing Completely at Random
MCMC	Monte Carlo Markov chains
NHS	National Health Service
MI	Multiple Imputation
MICE	Multiple Imputation with Chained Equations
MIDAS	Meteorological Office Integrated Data Archive System
MNAR	Missing Not at Random
MSE	Mean Square Error
NIGB	NHS National Information Governance Board
NMMAPS	The National Morbidity, Mortality, and Air Pollution Study

NO	Nitrogen monoxide, also termed nitric oxide
NO ₂	Nitrogen dioxide
non-CDP	Non community dependent pneumonia
NO _x	Nitrogen Oxides (including NO and NO ₂)
O ₃	Ozone Gaseous pollutant
ONS	The Office of National Statistics for England and Wales
Oxidative Stress	Tissue damage due to an imbalance within the cells between the level of free radicals and antioxidants
PAC	Privacy Advisory Committee, Scotland
PAHs	Polycyclic aromatic hydrocarbons
PCBS	Polychlorinated biphenyls
PCDFs	Polychlorinated dibenzo-furans
Phagocytosis	The process dealing with foreign material in the body.
PM _{0.1}	Ultra Fine Particulate Matter with aerodynamic diameter $\leq 1\mu\text{m}$
PM ₁₀	Particulate Matter with aerodynamic diameter $\leq 10\mu\text{m}$
PM _{2.5}	Fine Particulate Matter with aerodynamic diameter $\leq 2.5\mu\text{m}$
PMM	Predictive Mean Matching
PFCOD	Primary 'cause of death' field corresponds to only the underlying cause of death field in the mortality dataset.
Pulmonary Epithelium	The lung wall, in particular the cell wall within the alveoli.
Rate Ratio	Rate at any particular time if exposure = x+10/Rate if exposure=x
REC	NHS Research Ethics Committee
Relative Risk	No of deaths if exposure = x+10 / No of deaths if exposure = x
Risk	Probability that an event will occur
%RR	Percentage Relative Risk
SDEM	Site Dependent Effect Method
Sensitivity	Proportion of positive results which are correctly identified
SO ₂	Sulphur dioxide
Specificity	Proportion of negative results which are correctly identified
Thrombosis	Blood clots within the blood vessel, obstructing flow.
TSP	Total Suspended Particles
Underlying COD	Cause of death initiating the chain of events leading to death
Vascular Plaque	A Lesion located on the artery wall.

VIF	Variance inflation factor
WHO	The World Health Organisation
μgm^{-3}	Micrograms per cubic metre of air. A unit for describing the concentration of air pollutants in the atmosphere, as a mass of pollutant per unit volume of air. A concentration of $1 \mu\text{g m}^{-3}$ means that one cubic metre of air contains one microgram of pollutant.

ABSTRACT

Submitted by Matthew Gittins-April 2016 - For the degree of Doctor of Philosophy

Air pollution and mortality: an investigation into the lag structure between exposure to air pollution, temperature and mortality from pneumonia, chronic obstructive pulmonary disease, & ischaemic heart disease.

Introduction: The association between daily air pollution exposure and risk of mortality is well established. Few studies have investigated in detail the associations beyond a seven day lag. The aim of this thesis was to investigate the change in risk across longer (30 day) periods post exposure for three specific causes of death: pneumonia, chronic obstructive pulmonary disease (COPD), and ischaemic heart disease (IHD).

Methods: Daily Scottish mortality data (1980-2011) was matched to measurements from local fixed site pollution (Black smoke, PM10, PM2.5, SO₂, & NO₂) and temperature monitors. Exposure on subjects' 'day of death' was compared with control days in a time-stratified case-crossover analysis. Exposure effects on 30 days prior to day of death were modelled using distributed lag non-linear, lag stratified, and cubic distributed lag models. Matching hospital admissions data inferred subject location during exposure, further analyses investigated extreme outliers and missing data using multiple imputation techniques. The analysis accounted for several confounders including accurately modelling temperature relationships unique for each cause of death.

Results: Of the 919,301 deaths, 20% were classified as being caused by pneumonia, 9.5% as COPD, and 30% as IHD in the 'any' cause of death field. Non-linear effects for temperature and linear effects for the pollutants were present across all 30 days. Temperature-mortality was observed to be U-shaped at shorter lags. Consistently increased risk occurred for longer in cold temperatures with 1°C increase (30 days lag) = %RR -0.35% Pneumonia, -0.62% COPD, and -0.26% IHD. PM_{2.5} on all three outcomes, and all pollutants on COPD showed the greatest effect sizes. In general, COPD risk only occurred after a delay, peaking between 12-18 days. COPD risk due to PM_{2.5} was immediate (%RR (95% C.I.) = 1.05% (0.14%,2.01%)) and lasted the full 30 days. Pneumonia risk often reported the shortest lag of 10-15 days, whereas IHD risk occurred 2 days after exposure but lasted the remaining 30 days. There was some evidence especially for pneumonia of a smaller association between air pollution on mortality when subjects included were present in hospital. A simulation study indicated slight improvement in accuracy when 'multiple imputation' was performed compared to 'complete cases' analysis; though both techniques reported similarly underestimated effect estimates. Extreme outliers in the main analysis of pollution exposure did not appear to have a strong influence on the risk. However, large variability between monitor measurements of pollution exposure was present and appeared to be influencing the results.

Conclusion: This study provides additional evidence on the link between air pollution, and temperature, and acute mortality. Particular focus was on three causes of death (pneumonia, COPD, and IHD) that are shown to be influenced by air pollution in subtly different ways. Results also indicated that the 'true' effect of air pollution on mortality might be greater than shown by mortality studies which do not use hospital admission location during exposure into account.

DECLARATION

No portion of the work referred to in the thesis has been submitted in the support of an application for another degree or qualification of this or any other university or other institute of learning.

COPYRIGHT STATEMENT

- i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes..
- ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.
- iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
- iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property University IP Policy (see <http://documents.manchester.ac.uk/display.aspx?DocID=24420>), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see <http://www.library.manchester.ac.uk/about/regulations/>) and in The University’s policy on Presentation of Theses

PRESENTATIONS RELATING TO THIS THESIS

- Lane Lecture series (Manchester) 2011 Have short-term effects of environmental exposure on Pneumonia mortality been under-estimated because hospitalisation is ignored? [oral presentation]
- ISEE International Society for Environmental Epidemiology (Barcelona) 2011 Have short-term effects of environmental exposure on Pneumonia mortality been under-estimated because hospitalisation is ignored? [poster presentation]
- Cranfield Outdoor air pollution meeting (Cranfield) 2012 Have short-term effects of environmental exposure on Pneumonia mortality been under-estimated because hospitalisation is ignored? [poster presentation]
- 6th Occupational & Environmental Epidemiology Meeting (London School of Hygiene and Tropical Medicine) 2012 Have short-term effects of environmental exposure on Pneumonia mortality been under-estimated because hospitalisation is ignored? [poster presentation]
- Institute of Population Health Showcase (Manchester) 2013 Have short-term effects of environmental exposure on Pneumonia mortality been under-estimated because hospitalisation is ignored? [poster presentation]
- Institute of Population Health Showcase (Manchester) 2015, *“Modelling delayed mortality risk associated with potentially non-linear air pollution effects across a 30 day exposure period.”* [poster presentation]
- Centre for Biostatistics Seminar series (Manchester) 2015, *“Modelling delayed mortality risk associated with potentially non-linear air pollution effects across a 30 day exposure period.”* [oral presentation]
- ISCB International Society for Clinical Biostatisticians 2015 (Utrecht), *“Modelling delayed mortality risk associated with potentially non-linear air pollution effects across a 30 day exposure period.”* [oral presentation]

PUBLICATIONS RELATING TO THIS THESIS

- Matthew Gittins, Roseanne McNamee, Melanie Carder, Iain Beverland, Raymond Agius. 2013 'Has the short-term effect of black smoke exposure on pneumonia mortality been underestimated because hospitalisation is ignored: findings from a case-crossover study'. (published) Environmental Health DOI: 10.1186/1476-069X-12-97

ACKNOWLEDGEMENTS

This has been a long time coming, a part-time PhD is not something anyone should chose to do lightly. Juggling what only ends up being a full time job and a part-time PhD has been a challenge at the best of times, exhausting for the rest. Six years of early mornings, late nights, weekends, and working holidays are the only way it can be done no matter what people say. I can therefore only thank those who have been patient, supportive and provided guidance along the way.

Firstly my supervisors, I can only express my gratitude to both Professors Roseanne McNamee and Raymond Agius. Thank you for your continued patient and thorough support, advice and guidance over the last six years that have enabled me to keep on the right track achieve the goals we set out and see the work to a satisfactory completion.

Thank you to my line manager Prof Chris Roberts and the two leads of the Centre for Biostatistics; Prof Graham Dunn and Prof Andy Vail. Who have kept me in employment, sponsored my PhD and provided additional support and guidance.

Six years feels even longer when all around you are completing their PhDs. This however has meant I have made a number of new friends all of whom have contributed strongly to both working and social life through both wisdom and laughs and of course proof reading. There are many, and I am sure I will forget someone important but I will always be grateful to: Antonia Marsden, Matthias Pierce, Lesley-Anne Carter, Clare Flach, Kim Hannam, Ian Jacob, and Hazel England. Hopefully, I will one day return the favour in the mean time I wish you the best of luck no matter what you are doing or where you end up.

Finally, and most importantly my family, the last few months have not been easy for any of us and just like any of us I am sad Gran is not here to see this come to completion. This is not just for her but all my grandparents Iris and Harry Duke, Muriel and Harold Gittins. In reality they were just the beginning of this and I hope I have made them and the rest of my family mum - Carol, dad – Malcom, brother - Martin, uncles and aunts – Steve, Ruth, Mavis, John, all proud. Thank you all for everything.

1 INTRODUCTION

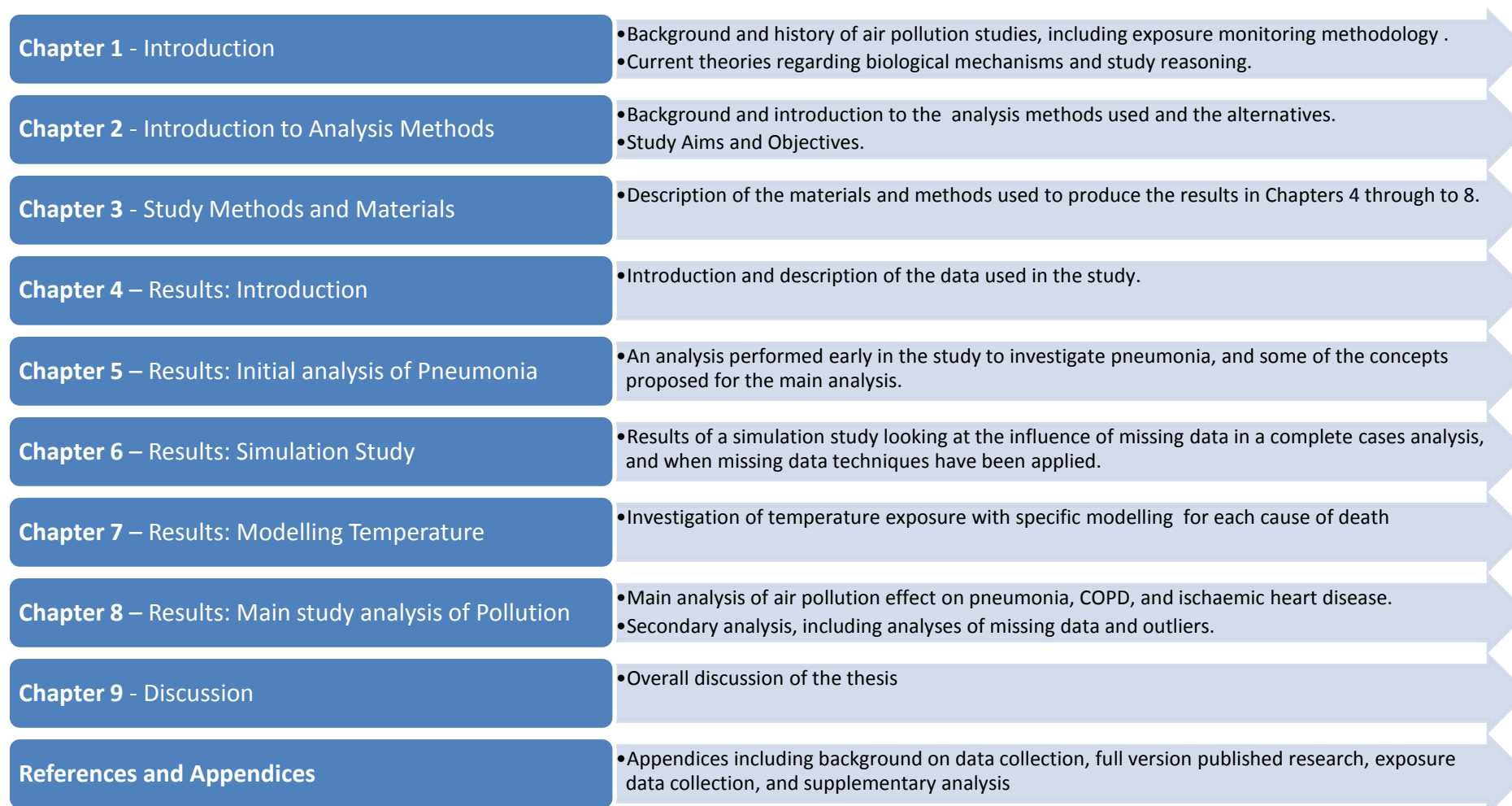
1.1 Chapter Introduction

This chapter introduces the thesis with a brief history and background to air pollution and its influence on human health. It outlines the early attempts to estimate the risk between exposure to air pollution and mortality. Information is provided regarding both potential air pollutants and the main confounding factors including temperature, and how they link to specific causes of death. The current evidence within the latest literature and the plausible biological mechanisms connecting exposure to air pollution will be outlined for the specific causes of death: pneumonia, chronic obstructive pulmonary disease, and ischaemic heart disease.

1.1.1 Thesis structure

This thesis has been written in a ‘traditional format’ structure. An overview of the thesis structure can be found in Figure 1.1, with a brief explanation of the contents within each chapter. The first three chapters of the thesis introduce the background and reasoning behind air pollution studies, the analysis methods likely to be employed, and the materials and methods actually used in the analysis in this thesis. Chapters 4 to 8 report the results of the analysis in distinct sections, where 5, 6, and 7 are results relating to early investigations leading up to the main analysis in Chapter 8. An overall discussion is then given in Chapter 9 that will aim to bring the results together. A paper was published during the course of comparing indoor and outdoor personal particulate matter exposure this thesis in the *Journal of Environmental Health*. A large amount of the paper has been reported in the results Chapter 5, and the full paper can be found in Appendix B.

Figure 1.1 - Thesis structure



1.2 Background

In March 2014 the World Health Organization estimated that 3.7 million deaths a year are linked to outdoor air pollution, 580,000 of which are located within Europe.¹ Specific causes of death such as pneumonia, Chronic Obstructive Pulmonary Disease (COPD) and Ischaemic Heart Disease (IHD) are strongly linked to air pollution, with ischaemic heart disease and COPD thought to represent 40% and 11% of all outdoor air pollution related deaths respectively.

The effect of air pollution on the population has been a concern for many years but a clear link was not confirmed until a series of air pollution disasters in the mid-20th century.²⁻⁵ These initially included events such as:

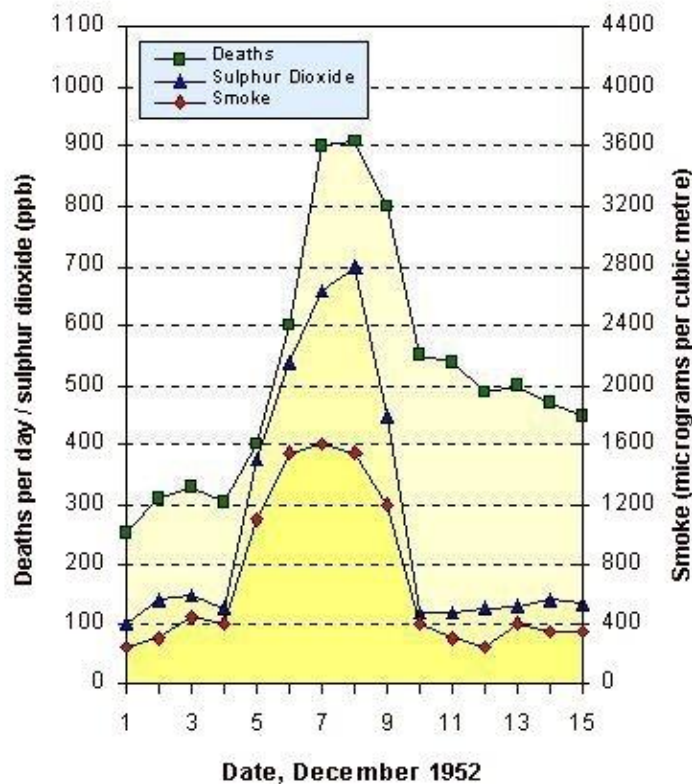
- Meuse Valley, Belgium – where industrial pollution from local steel works in was thought to have contributed to 60 deaths in December 1930.²
- Donora, Pennsylvania – where approximately twenty deaths occurred in October 1948 that were associated with coal fired home and industrial facilities,³
- London, England - 300 died in a severe London fog in November 1948.⁴

However, these incidences were minor in scale compared to the four day London fog of December 1952 which was accredited with a minimum of 4,000 deaths within the first two weeks, and approximately 12,000 deaths by the end of February 1953.⁵ A lack of pollution measurements for Meuse Valley and Donora meant that confirmation of the association with air pollution was not possible until London in 1952.⁵ These episodes however, clearly highlighted the connection between substantial increases compared to normal in mortality rates with substantial increases in concentration levels of air pollution.^{2,4,6}

Using basic graphical representations (Figure 1.2) and simple comparisons between mortality rates and pollution levels, these incidents of the early 20th century clearly indicated a link between high levels of air pollution and mortality, and acted as a catalyst for improvements in air quality standards. Despite the rise of the motor vehicle as the major source of pollution, improvements in air quality came about as a result of

implementing pollution limiting technology, more stringent air quality legislation, and the move from coal fire to electric heated homes. The challenge then was to determine if low level concentrations of pollution affect mortality. In particular, do increases in lower concentration levels affect mortality at a similar rate to increases at higher concentration levels? Is the relationship the result of a steady increase or is there a threshold, a concentration level at which mortality significantly increases?⁷ How long does a change in air pollution level contribute to mortality and morbidity rates? How do different ambient pollution compositions effect and interact with each other with regards to health effects.

Figure 1.2 – Basic graphical representation of the comparison of mortality rates and pollution levels for the London Smog of 1952 (Wilkins E.T. 1954)⁵



These are just some of the questions that have been addressed since the 1980s. Better pollution monitoring techniques and improvements in computational technology and power have become available, making more refined comparisons possible and aiding analysis where effects experienced were small, if seen at all. These improvements have led to further development in analysis techniques, and improved quality of data during the 1990s. Many studies using various statistical models have tried to account for the

relationship between pollution exposure and mortality such as the generalised additive regression model (GAM)⁸ and more recently the case-crossover design with conditional logistic regression.⁹ A large number of these studies have indicated a relationship between short-term changes in air pollution (e.g. same day or within 2-3 days) and increases in daily deaths.^{10,11}

In 2000 Joel Schwartz proposed that the effect of an increase in exposure on any day may continue over a longer period of time.¹² One might expect this may show a rapidly increasing risk level in the early days that peak several days after exposure before it begins to dissipate. The length and shape of the delayed risk will illustrate important characteristics regarding the change and length in risk of all-cause and cause specific mortality, when exposed to air pollution. However, few studies have investigated in detail the relationship when the delay in risk, denoted as the 'lag period', is extended to 30 days.¹³ By omitting these days, underestimation of the effects may be incurred. If the effect period is in fact less than 24 hrs then the model can be constrained to the same day exposure. However, if this takes longer or large delay (greater than 3 days) is present between initial infection and serious symptoms then exposure levels on days greater than 3 days prior to the event will be important. The lag shape and length is likely to vary depending on the cause of death and perhaps the pollutant. The time between initial onset of the disease and final outcome is dependent on how the pollutant interacts with the different stages of a disease progression; the induction, the period prior to first symptoms, and the final outcome. These issues, along with many more, need to be accounted for before adequate effect estimates are produced.

In the UK, the current air quality standards were made into law in 2010,¹⁴ and were based on a European Union directive from 2008.¹⁵ For each pollutant these directives outline limits on the average annual pollutant, and a maximum number of days per year that a daily average should exceed. These, along with guidelines from experts at The World Health Organisation (WHO), and The committee on the Medical Effects of Air pollution (COMEAP), have gradually improved the air quality in the UK over the last 30 years.^{16,17} However, even with the implementation of these recommendations, the general downward trend in background exposure, and the presence of a large amount of

literature on the short-term and long-term links, exposure to air pollution and its influence on ill-health in the population is still a major concern.

As recently as February 2016 the Royal College of Physicians released a report outlining the latest guidance and recommendations regard life-long exposure to air pollution.¹⁸ The report went to great lengths to highlight the public health burden of air pollution and for air pollution to be taken more seriously by the public, industry, and policymakers. Improving our understanding of the relationship between air pollution and short-term outcomes is clearly important. This may be particularly true for cause specific mortality such as pneumonia, COPD, and Ischaemic Heart Disease (IHD), where underlying biological mechanisms and pathogenesis are still unclear. Increasing the information regarding the change in acute risk over longer time periods may improve our understanding of the underlying pathogenesis, but also create insight into individual and population level prevention measures and preparation of health service to account for peak periods of risk. The following thesis will hopefully help improve our understanding of the short-term changes in population risk associated with three specific causes of death when under the influence air pollution.

1.3 Exposure data – Atmospheric air pollution

Air pollutants are atmospheric materials that have been identified to be harmful to human health by either increasing the risk of illness or mortality rates. Absolute increases and changes in the mixture of air pollution primarily result from human activity via the burning of fossil fuels for production of energy, home heat sources, or through traffic emissions. Over the last 30 years the levels of certain pollutants have improved greatly due to implementation of pollution reduction laws and technology.

Pollutants have been crudely classified into four groups 1) particulate matter e.g. Black smoke, 2) gaseous e.g. Sulphur Dioxide or Carbon Monoxide, 3) heavy metals (which are also present in particulates) e.g. lead, nickel, or manganese, and 4) persistent organic pollutants also known as dioxins, e.g. polychlorinated biphenyls (PCBS), Polychlorinated dibenzo-furans (PCDFs), and Polycyclic aromatic hydrocarbons (PAHs).¹⁹ Persistent organic pollutants are primarily produced via the burning of chlorine usually found in plastics or pesticides. Heavy metals are naturally found in the earth's crust and tend to be related to geographical sources such as volcanoes or sand storms. Both heavy metals and persistent organic pollutants are resistant to being degraded or broken down and tend to enter the human body via food and water sources where, as they move up the food chain, they increase in quantity. Here the focus is on the routinely measured particulates and gaseous pollutants which have a more direct influence on health via inhalation.

A major challenge in any observational epidemiological study such as this, is to accurately measure exposure.^{20,21} Misclassification of the true exposure can cause biased and invalid effect estimates. This is particularly important in studies of air pollution where effect estimates are small to begin with.²² A prospective study where participants wear personal monitors would provide the most accurate assessment of the individual subject's exposure.²³ Unfortunately studies involving personal monitors are impractical since the monitors can be expensive, cumbersome, and would need to be worn at all times for a number of days. In time these issues may be solved, however they mean studies of this type have currently been limited in their sample size and have tended to focus on specific sub populations.²⁴⁻²⁷ Small effect sizes require greater sample

sizes in order to gain enough statistical power to be confident of the effect size estimates. To improve sample size, most studies have retrospectively matched fixed monitor pollution estimates to individuals based on the proximity of the monitor to home postcodes and date of the event of interest e.g. date of hospital admission.²⁸⁻³¹ Fixed site monitor estimates provide a comparatively simpler method for providing exposure data for a large number of subjects, over an extended time period, for a number of different pollutant types. However, the limited number of fixed site monitors even in urban areas means a limited ability to account for any spatial variation in exposure. The assumption is that anyone whose home address is within a certain radius of the monitor is present there during the time period and experiences the same exposure level as the monitor records. This is highly unlikely, with pollution hotspots, major roads, or street canyons experiencing greater exposure levels that may, depending on the pollutant, disperse rapidly with increased distance from the source, creating within community differences.³²⁻³⁴ The following gives background information relating to fixed site pollution and meteorological exposure data collection for the UK and with focus on Scotland since 1980.

1.3.1 Fixed pollution monitoring networks

The series of extreme pollution events occurring in the 1950's resulted in the UK Government introducing its first Clean Air Act in 1956. In 1961 the UK established a national air pollution monitoring network called the National Survey, which has been monitoring black smoke and sulphur dioxide in the UK for the last 60 years.

The switch in the primary air pollution source during the 20th century from industrialised or home fires to traffic emissions, meant monitoring techniques needed to adjust to the rise of gaseous pollutions leading to the UK establishing in 1987 an automatic urban monitoring network.³⁵ In 1992 the then Department of Environment established an Enhanced Urban Network (EUN), where in 1995 all urban monitoring was consolidated into one comprehensive programme. Finally in 1998 the previously separate UK urban and rural automatic networks were combined to form the current Automatic Urban and Rural Network (AURN).

The Black Carbon Network and its predecessor, The Black smoke & Sulphur Dioxide monitoring network, are a part of the non-automatic networks with automatic methods introduced post 2008. Samples collected in person were subjected to a form of chemical analysis from which the final pollutant concentrations are calculated. Measurements may be recorded less frequently compared to the automatic networks. However, because the network has been running since the early 1960s, it provides a useful dataset that allows for a comparison across periods that include changes in pollution abatement technology, policies, primary sources, and population lifestyle factors.

Started in 1987, the AURN is the most comprehensive automatic national monitoring network in the country. AURN aims to measure and provide information on air quality with respect to short-term changes, long-term trends, and the influence of interventions. It covers 127 sites located across the UK and produces hourly measurements of nitrogen oxide, nitrogen dioxide, sulphur dioxide, ozone, carbon monoxide and both particulate matter aerodynamic sizes.³⁵

1.3.2 Potential pollutants of interest

Atmospheric pollution can be designated into several distinct compositions. Determination of the health effects of specific pollutants has been an important consideration. Most studies have focused on the 24 hour average ambient airborne particles which are a measurement of the amount of inhalable particles in the air. Initially, measures were of Total Suspended Particles (TSP) which included a significant percentage of overly coarse material that, due to their size, are thought to have little impact. TSP was quickly replaced by more sophisticated measurement techniques that gauged particles with an aerodynamic diameter less than or equal to 10µm (PM10), or fine particles with diameter less than or equal to 2.5 µm (PM2.5) that are thought to penetrate further into the lungs. Alternatively there is a slightly more crude measurement in 'black smoke' (BS) particulate matter. Black smoke is slightly different from PM as the colour can indicate the toxicity of the measurement. Darker colour may relate to a more harmful makeup (e.g. predominantly PAHs), whereas a colourless result may represent less harmful components such as sodium chloride,

sulphate, or ammonium.^{36,37} The source and composition of the particles involved varies depending on location, though they are commonly made up of coal smoke, soot, dust, and wind-blown soil with the burning of fossil fuels and vegetation.³⁸

Pollutants emitted directly into the atmosphere such as CO & SO₂ are 'primary' pollutants, whereas those resulting from chemical reactions (e.g. Ozone) are 'secondary' pollutants, and in some cases (NO₂ & PM) are classed as both. Gaseous pollutants commonly include sulphur dioxide (SO₂), carbon monoxide (CO), nitrogen oxides (NO & NO₂), and ozone (O₃). Major sources of these pollutions tend to be coal and oil based power stations, industrial plants, and most recently petrol powered vehicles. Nitrogen oxides will inter-convert depending on certain atmospheric conditions. Ozone is classed as a secondary pollutant that forms as a result of other emissions.³⁸

In research studies the presence of several co-pollutants is a concern when determining valid individual pollutant effect estimates; as if not dealt with adequately in the analysis, there will be bias. However, the highly correlated nature of pollutant variables makes it difficult to determine separate effects when including multiple pollutant covariates in a statistical model. Those who have attempted multi-pollutant models have fitted pollutants individually at first followed by two pollutant models,³⁹⁻⁴¹ and then in some cases extended to four pollutant models.^{42,43} Results tended to indicate significant associations with the pollutants when included individually in the model. As other pollutants were included as covariates the effects seen for the non-particulate pollutants such as SO₂, CO, NO₂ and O₃ tended to disappear, whereas the effect of PM₁₀ and PM_{2.5} either did not significantly alter, or in the PM_{2.5} case, became stronger.

1.3.3 Pollutants their measurements and influence

To estimate the effect of pollution on health it is important to accurately measure pollutant concentrations. Pollutants emitted directly into the atmosphere such as CO & SO₂ are primary pollutants, whereas those resulting from chemical reactions (e.g. Ozone) are secondary pollutants, and in some cases (NO₂ & PM) are classed as both. Fine particles for example, can be emitted directly from a number of natural and manmade sources as well as indirectly from oxidation of SO₂ and NO₂.

The type of pollutants recorded in urban areas tends to depend on the dominant fuel composition, the atmospheric conditions such as temperature and pressure, and the employment of any pollutant abatement technology. The two major sources are currently power stations and road transport, meaning urban areas will typically have higher contributions from road traffic sources than the predominantly rurally located power stations. Measurements are much more difficult to take when they are coming from a large number of small sources e.g. households or vehicles. Vehicle emissions are particularly difficult to accurately measure as it can depend on driving style, traffic conditions, type of car, and local geographical characteristics. Traffic related emissions tend to be emitted lower to the ground whereas the monitors are often located a distance from the road or even high up on a building.

1.3.3.1 Black smoke (BS)

Along with ‘black carbon’ measurements, ‘black smoke’ formed one of the earliest measures of air pollution in the UK and across Europe with limited measurements beginning in the early 1970s.^{44,45} Black smoke is commonly measured by the number of particles collected on a filter whose blackness is determined by light reflection or transmission. In the reflection method, light reflected from the filter is measured in the form of a percentage.⁴⁶ A 100% reflection is an unexposed filter; in the exposed filter the percentage of incident light extracted by the filter is quantified. The reflectance is then related to the concentration by using a standardised calibration curve.⁴⁷ The number of airborne particles determined by the BS reflection method depends heavily on monitor location and season and is a rather crude measure of airborne particles reflecting primarily the carbon content and not the mass of the particles. The reflectance method was originally calibrated based on the primary source of pollution being coal burning fires therefore making them potentially less suitable to measuring traffic emissions.⁴⁸ The monitors are considered unrestrictive regarding size and composite, however in practice the size measured is thought to be approximately equivalent to particulate matter with an aerodynamic diameter of $4.5\mu\text{m}^{-3}$.^{49,50} Black smoke can also be considered as a useful measure of the overall pollutant level in the form of a particulate and gaseous composite. Black Smoke is therefore thought to be a useful

measure of traffic emissions and a useful general atmospheric pollutant marker in pollution exposure studies.⁵¹

1.3.3.2 Particulate Matter (PM)

In addition to the 'black smoke' measure, particulate matter is also measured and categorised based on its aerodynamic size. Particulate matter, like black smoke, is not strictly a single pollutant rather a varying mixture of all particles composed of different sizes and chemical compositions.¹⁹ Though particulates may come from a wide range of sources such as windblown soil, street dust, and sea spray, the primary urban sources tend to be road traffic and re-suspended road surface dusts and soils. Particulate matter measurements are based on their aerodynamic size formally designated as PM10 or PM2.5. PM10 relates to the fraction of particles passing an inlet with a 50% cut-off efficiency at an aerodynamic diameter of 10 μ m. The PM10 fraction includes the majority of atmospheric particles excluding only the upper end of the coarse range (10-100 μ m). Particulates are believed to have significant health implications.⁵²⁻⁵⁴ PM10 corresponds to particles small enough to reach the thorax region of the lung when inhaled. PM2.5 or fine particles with $\leq 2.5\mu$ m aerodynamic diameter are capable of penetrating deeper into the alveolar region of the lung with several studies indicating that fine particles are linked to increased risk.⁵⁵⁻⁵⁹

Several methods of measuring particulate matter exist. One common method, is the use of a high volume sampler fitted with size selective inlet to collect PM10. Particulates are collected on a pre-weighed filter paper made of glass fibre or quartz, which is then weighed again after 24hrs at a rate of 1m³ per min. Hygroscopic salts in the air require that the filters are kept at a constant temperature and low humidity. Alternatively, a dichotomous sampler draws air in and splits at a virtual impactor. One stream continues straight onto a coarse filter and contains coarse particles due to a natural inertia, the second stream moves right carrying fine particles and collects on a fine particle filter. Any fine particles in the coarse filter are subtracted by a measurement based on the fine filter. If fitted with 10 μ m inlet and a 2.5 μ m virtual impactor then fine ($\leq 2.5\mu$ m) and coarse (2.5-10 μ m) can be recorded.⁶⁰ Thirdly, a continuous measurement of PM10 or

PM_{2.5} can be acquired through the Tapered Element Oscillating Microbalance (TEOM).⁶¹ Air is drawn in via a 10/2.5 µm inlet and heated to 50°C to dry the particles. The particles collect on a filter and are vibrated using an oscillating microbalance where the frequency changes depending on the accumulation of particles. This generally gives lower values due to loss of semi-volatile material in the heating stage, though it has been shown to produce exposure measurements comparable to other monitoring methods.⁶²

1.3.3.3 Sulphur Dioxide (SO₂)

The primary source of SO₂ is the combustion of fossil fuels containing sulphur, particularly coal and heavy fuel oil power stations. The introduction of the flue gas desulfurization technology to limit SO₂ release means that sources tend now to be released in highly concentrated plumes leading to highly elevated ground level concentrations that last for quite short periods. Concentrations do not tend to build up efficiently at ground level and so are thought to have a smaller impact.

The most common SO₂ measurement technique uses a gas ultra-violet (UV) fluorescence.⁶³⁻⁶⁵ In a gas fluorescence instrument, air is drawn through a cell which is irradiated by light at a particular wavelength. The intensity of the resulting fluorescence can be related to the amount of SO₂ concentration in the air. Interference from fluorescence formed by water vapour or hydrocarbons is removed by including diffusion dryers and hydrocarbon scrubbers. This technique can measure 1 part per billion (ppb) or less in a time of 1 minute.³⁸

1.3.3.4 Carbon Monoxide (CO)

Carbon monoxide is mostly generated from the internal combustion engine. The majority of carbon present in fuel is oxidised into carbon dioxide, with a small fraction left that is incompletely oxidized in the form of carbon monoxide.

Carbon monoxide monitoring tends to use non-dispersive or gas filter correlation infrared methods.⁶⁶ The gas filter involves broad infrared band radiation passing alternately through gas cells containing carbon monoxide and molecular nitrogen as they rotate on a spinning wheel. The pulse of radiation travels through a multi-pass optical cell, through which ambient air is drawn. The sample (nitrogen cell) and reference (carbon monoxide cell) beams are separated. The reference beam is then absorbed by all components in the sample cell except carbon monoxide. The difference in beams is due to carbon monoxide absorption in the air sample.

1.3.3.5 Nitrogen Oxides (NO, NO₂, NO_x)

Nitrogen oxides (NO_x) are formed in high-temperature combustion of nitrogen & oxygen, which are already present in the air or are from oxidation of the nitrogen present in fuels. Ozone in the air immediately interacts with traffic emissions initially in the form of Nitrogen Oxide (NO) converting into Nitrogen Dioxide (NO₂). Measurements can be difficult to perform as nitrogen and oxygen are already common atmospheric components. Due to the composition and process generating NO₂ all but the most polluted environments contain nitrogen dioxide, even though its considered more toxic.

The chemiluminescent analyser method measures light emission from a chemiluminescent reaction between Nitric Oxides in the air and Ozone generated within the instrument.⁶⁷ The monitor can function in two stages. In the first, air is passed directly into a reaction chamber to mix with ozone and the NO is measured directly. In the second mode the air is passed through a heated stainless steel/molybdenum

converter before reaching a reaction chamber. The idea is to convert the NO₂ in the air to NO; NO₂ is then measured by calculating the difference between the two modes.

An alternative method, an inexpensive diffusion tube, can also measure NO₂. A straight hollow tube of 7cm by 1 cm with one end closed is hung vertically with the open end at the bottom. Inside the closed end is a metal grid coated in triethanolamine, which acts as an absorber and sink for the NO₂. At the end of 1-2 weeks the nitrate collected by molecular diffusion up to the grid is analysed to determine the NO₂ concentration.⁶⁸ This method can be inaccurate and may overestimate by approx. 22-24%,⁶⁹ and so are often deployed in multiples but can still be affected by windy conditions, changes in the NO-NO₂-O₃ chemical reactions and location of monitor.⁷⁰

1.3.3.6 Ozone (O₃)

Ozone when found at ground level is a secondary pollutant as it forms in the atmosphere from pollutants already present and not directly from a source. Ozone is produced from interactions between nitrogen oxides and volatile organic compounds under the influence of hot temperatures and ultra-violet light.^{51,71} Due to the complex atmospheric chemistry its concentrations tend to be formed close to fossil fuel sources, particularly road traffic emissions. In cooler months, ozone tends to be destructed in urban areas when large amounts of nitrogen oxide are present to produce nitrogen dioxide, meaning relatively higher concentrations in rural areas where nitrogen dioxide is reduced.⁵¹ In warmer summer months ozone has a longer timescale to form resulting in more high pollutant episodes in urban areas.

Early chemical methods of measuring ozone were shown to be unreliable until the increased availability of the chemiluminescent analysis which is based on ozone-ethylene reactions. This was subsequently followed by the ultra-violet photometric analysis which is now the common measurement method. Ozone has strong UV absorption at 254nm which is measured in a long path absorption cell where ambient air is drawn in. Before entering the cell some of the air is passed through an ozone scrubber to remove other UV absorbing material. The difference in absorption between the air scrubbed and the untouched air is then due solely to ozone.³⁸

1.4 Exposure data – Meteorological factors

A primary factor in air pollution measurements and studies investigating health are the meteorological conditions most prominently temperature. Meteorological data is available from the British Atmospheric Data Centre (BADC)⁷², who allow access to data recorded under the Met Office Integrated Data Archive System or MIDAS Land Surface Observation database.⁷³ The Met Office was established in 1854 with the aim of researching the possibility of forecasting the weather. The MIDAS dataset, also started in 1853, now records at least some information for each county in the UK. Daily and hourly weather measurements are recorded in the forms of air temperature, soil temperature, wind patterns, sunshine duration and radiation levels, rain measurements, and climatology data. In 1974, this information was extended to a global database covering the rest of Europe, Africa, Asia, South and Central America, West Pacific, and Antarctica.

Certain meteorological factors particularly temperature, are important predictors of ill-health in their own right but also confounding factors in pollution-mortality studies and so it is important that the process behind recording these exposures is clear.

1.4.1 Manual & automated meteorological monitoring

Until recently meteorological observations were performed manually; the observer read the instruments and made visual estimates of visibility, cloud cover, current weather, hourly wind, hourly sunshine, and hourly rainfall. Regular attendance, maintenance, and calibration reduced the presence of missing data and errors. However, errors could occur due to miss-reading the measurements or failure to follow correct procedures to ensure accurate subsequent measurements.

The introduction of automated systems reduced the role of a human observer and any corresponding errors. Even if the station remains manned 24 hours a day the process of reading the instrument, processing the result, and transmitting the data is fully automated. An observer can modify the automated output if they consider there to be an

error. Automated stations, first introduced in the 1970's, now cover the entire network. Unfortunately, by converting to the automatic system, an increase in the presence of missing data, by up to 5% in some cases, was observed. A human observer would have noticed instrument failure earlier would have and been able to improvise when needed, reducing the chance of missing data or error. The introduction of automated stations also meant a change in instruments, with the liquid in a glass thermometer being replaced by an electrical resistance thermometer. Because of this change, there may be bias between differing instruments in the data if the crossover occurs during the study period. This is likely to be a minor source of bias in studies carried out from 1980 onwards as monitor changes occurred in the 1970s.

The primary meteorological measurements of interest in an air pollution study have tended to be temperature and relative humidity.

1.4.1.1 Temperature

All temperature measurements in the MIDAS dataset were converted to Celsius (prior to 1961 temperature was recorded in Fahrenheit) and are recorded with a precision of 0.1°C. Instruments are calibrated on receipt from the manufacturer and at intervals after deployment with the aim of ensuring that the accuracy of the measurement at least to 0.2°C.

The standard measurement method since the 19th Century was the liquid-in-glass thermometer. Either mercury or ethanol liquid was used in order to accurately measure minimum temperatures. Once the temperature is recorded (initially on paper) the minimum and maximum thermometers are reset by shaking or tilting the device. Since the early 1980's, Electrical Resistance Thermometers (ERT) have been employed to record temperature data by measuring the resistance in platinum, dependent on the temperature.

1.4.1.2 Humidity

There is not direct measure of humidity instead a relative humidity has commonly been recorded based on a function of the dry bulb temperature relative to either a wet bulb temperature or dew point temperature. Where wet bulb temperature is the temperature after the cooling effect of evaporation has occurred and dew point temperature is the temperature at which water in the air starts to condense.^{74,75} The standard exposure method for thermometers for measuring the dry, wet-bulb, maximum and minimum temperatures is to locate 1.25m above ground a louvered white screen wooden box. This allows for free circulation of air around the thermometers whilst shielding from precipitation and external radiation. Most observations of surface humidity have been made using a wet bulb thermometer with some recent automatic station observations being made by a relative humidity sensor. The wet bulb thermometer is exposed alongside a dry bulb in the meteorological screen with the only ventilation provided by a natural flow of air, and the difference in temperature converted into a humidity score.

1.5 Pollution & meteorological fixed site monitor details

Monitor location is an important factor in accurately recording the pollution level, as monitors need to be well exposed to air from all directions. The primary aim of pollution monitoring is to estimate of the true pollution level in an area. However, some monitors are designed to determine specific pollution associated with traffic or industrial hotspots. Unless the main interest is in street canyons, monitors should not be located near large buildings. Instead they should be in positioned in open areas or on a roof top to achieve good atmospheric exposure. Sites can be broken down into the following types of sampling location.

1.5.1 Background stations

Background stations are located such that they are not strongly influenced by one pollution source and can give a good estimation of the typical general pollution level in town or urban centres. They tend to be in an open space and are ideally located approximately 100 meters from a busy road with a sampling height of least 2-3 meters. Background monitor estimates should not be influenced by extreme concentrations such as one high pollutant producing source but the general level contributed to by a sum of multiple sources. The majority of people live in or experience background type areas hence these estimates should represent the concentrations that most people experience at home, shopping areas, or pedestrian precincts. Background stations are also better at reflecting long term time trends and seasonal changes in concentration than hotspots or traffic specific monitors. Rural background sites, tend to be located in open fields and so are more representative of the non-manmade background pollution levels.⁷⁶

1.5.2 Hotspots or kerbside stations

At some point during most people's lives they will spend some time in a high concentration area, possibly close to a busy road or near an industrialised location. Pollution networks often have hotspot monitors at a major road, busy intersection or in street canyons. Such monitors represent a small area only, often with only a radius of 20 metres. These locations represent the extremes of concentration and can be useful for pollutants that are fast acting with high, acute exposures. In a rural setting, hotspot monitors tend to measure the pattern of pollutant up wind from a major source (e.g. a city or power station).^{76,77}

1.6 Improving individual exposure estimates

Using a fixed site exposure monitor to estimate an individual person's exposure measurement can be limited due to spatial variation in exposure and lifestyle factors which will be discussed in detail later (Chapter 9.5.1). However, it may be possible to improve the relationship. Often a deceased individual's exposure is inferred from information regarding their place of residence with little or no attempt to take account of subject's actual location, circumstances or activities. Exposure is typically assumed to be the same for all subjects living within a given distance from a single pollution monitor or an average of multiple monitors within the area.^{78,79} Recent studies trying to improve exposure estimates have taken into account traffic density and other geographical information regarding the subject's neighbourhood or city.^{56,80} The presumption still remained that the deceased was in the geographical location of residence during the exposure period.

In fact, it is common for people to die in non-residential locations with 65.3% estimated to die in a NHS hospital/Hospice in England and Wales.⁸¹ Even if the hospital is located close to the place of residence one might reasonably suppose that a patient's exposure to outdoor pollution might be reduced when confined indoors.⁸² Epidemiological observations have shown that deaths associated with air pollution, specifically TSP and Particle Matter with an aerodynamic diameter less than 10µm (PM10), are

disproportionately increased outside of hospital.^{83,84} In addition, Jansen et al. 2002 found that the health effects of PM10 on cardiovascular disease and COPD in 14 U.S. cities decreased significantly as the proportion of homes with air conditioning increased.⁸⁵ Previous attempts at comparing risks in and out of hospital, such as Téllez-Rojo et al. 2001 and Zeka et al. 2006, have shown significant increased risk of death from respiratory or cardiovascular causes outside hospital with some cases almost a threefold increase. These studies have primarily used location at time of death without confirming location during exposure.^{86,87} Failure to take account of hospitalisation during exposure could lead to further effect underestimation if a substantial fraction of the population experience reduced exposure in air-conditioned hospitals. A large proportion of subjects hospitalised during exposure might explain why some observational epidemiology studies based on routinely collected data may struggle to replicate previously demonstrated associations between pollution and pneumonia caused mortality.⁸⁸

1.7 Adverse outcomes – Influence of exposure in humans

1.7.1 Potential biological mechanisms - Pollution

The underlying biological mechanisms between air pollution and mortality are not fully understood. Fine and ultrafine particulate matter (particles with aerodynamic diameter $\leq 2.5\mu\text{m}$ & $\leq 1\mu\text{m}$, respectively), are thought to have a stronger influence on biological mechanisms than larger particles for three reasons:

- 1) smaller particles are able to penetrate further into the lung potentially reaching deeper into the alveoli region
- 2) smaller particles mean it is easier to inhale a larger total number
- 3) the smaller the individual particles the greater the total surface area related to the total volume inhaled, meaning an increase in the contact area.

The increased contact of the inhaled foreign material in the alveoli region of the lung is thought to cause inflammation in the epithelium (membrane tissue of the lung walls) and interfere with the alveolar macrophages (cells designed to remove unwanted material). Inhibiting the lung's ability to remove unwanted material (a process called Phagocytosis) causes the lung to become overloaded. Contact with epithelial cells causes an increase in inflammatory cytokines that are proteins or peptides produced by cells associated with the immune system.⁸⁹ An the macrophages produce reactive oxygen and nitrogen species (molecules that contain oxygen/nitrogen) along with other cytokines. The resulting oxidative stress increases inflammation further in the surrounding pulmonary tissue. Inflammation in the lungs can lead to, or aggravate, respiratory lung diseases such as pneumonia, COPD, Bronchitis, Fibrosis, or Asthma.⁹⁰

Foreign material may further enter circulation into the blood stream by directly crossing the pulmonary epithelium and capillary endothelium, or by damaging the epithelium due to pulmonary oxidative stress and inflammation. Small particles can bypass removal processes and continue to circulate. They are then thought to affect the cardiovascular events in three ways:⁹¹

- 1) pro-inflammatory and vasoactive mediators (i.e. agents that affect the diameter of blood vessels) are released.
- 2) foreign material influences the autonomic nervous system (i.e. system that regulates major organs) through receptors in or near the lung
- 3) movement of foreign material into and around the blood stream.

The ultrafine material may enter directly into the cell in the heart affecting the mitochondria (energy production and stores) within the heart cells impairing the cardiac contractions. Similarly coarse material may interact with the certain macrophage cells causing them to aid entry into pulmonary or vascular cells through Phagocytosis where they may trigger inflammation through interaction stimulating reactive oxygen species (molecules that contain oxygen).⁹¹ Contact with the lung cells can cause inflammatory protein production to be increased that increase vascular inflammation and atherosclerosis (fatty deposits on the inner artery wall).⁹¹

Short-term (e.g. hours) increases in foreign material may also cause abnormalities in the body's haemostatic system (system keeps the body's functions in balance) with increased particle matter causing increased platelets, coagulation and subsequently thrombosis formation (blood clots).⁹² Other potential causes of thrombosis are due to increased production and circulation of plasma macrovesicles (fragments of plasma membrane) which can cause coagulation. Those particles that have entered the blood stream are also thought to directly activate endothelial cells and cause platelet accumulation causing atherothrombotic events (disruption of plaque or lesion within circulatory system).

The particles are also likely to influence the autonomic nervous system that regulates unconscious bodily functions in this case heart rate and breathing. Any imbalance in the autonomic system may increase hypertensive vasoconstriction and chance of arrhythmias (irregular heart beat).⁹³ There has been suggestion, though with some scepticism to the causal chain, that increased exposure may also reduce the heart rate variability which has in turn been linked to increased risk of cardiovascular events.⁹⁴

Ultrafine particles may continue to enter secondary organs such as the liver, kidneys or even the nervous system if the blood brain barrier has been breached.^{92,95,96}

1.7.2 Potential biological mechanisms - Temperature

As with pollution, though theories have been proposed the underlying biological mechanisms associated with risk of mortality due to temperature are currently not fully understood,⁹⁷ partly because they can vary in different susceptible sub-populations and different causes of death.⁹⁷⁻⁹⁹

To maintain the core body temperature under acute exposure to elevated ambient temperatures, the internal thermoregulation system causes blood to flow towards the skin in an effort to cool down via perspiration. The flow of blood away from internal organs is thought to cause excess stress on vital organs such as the heart and lungs.¹⁰⁰ If increased core temperature persists and the body reaches its thermoregulatory threshold i.e. the point where core temperature cannot be controlled by the internal cooling system, then an increase in the heart rate can occur along with increased blood viscosity, and cholesterol levels.^{101,102} Increased body temperature may also increase dehydration causing an imbalance in the ratio of fluid to electrolytes which may affect responses to heat in those already susceptible which may be due to general level of health or to a chronic disease already present.¹⁰³ Many of these factors will influence cardiovascular disease. However, other than exacerbating an already present condition such as may be the case in COPD, it is not clear how increased heat may directly cause respiratory disease.¹⁰⁴

Though associated with both, cold temperatures have been linked to respiratory diseases more than cardiovascular diseases. This may largely be due to changes in lifestyle meaning increased contact and spread of communicable respiratory diseases during colder seasons of the year. In addition to the standard response mechanism of shivering and constriction of blood vessels, a drop in core body temperature is thought to suppress the immune system. It is thought that an increase of infection occurs when a drop in core body temperature suppresses the production of important cells related to protecting or clearing the body of unwanted material (e.g. Phagocytosis).¹⁰⁵ More specifically, inhaled cold air is thought to increase the likelihood of acquiring an infection by compromising the body's natural immunological and mucociliary clearance by decreasing the temperature of the respiratory epithelium.⁹⁷ Colder temperatures may

increase the likelihood of cardiovascular related mortality due to a number of influences on the circulatory system. This includes increased heart rate, raised blood pressure, increased chance of a blood clot, constriction of both the arteries and blood vessels, a greater thickening of blood viscosity, and increased red cell count.^{102,106,107}

Regardless of exposure, once disease onset has commenced the induction period gives way to the promotion period between onset and first symptoms. This is followed by an expression period where symptoms are observed and eventually lead to a final outcome in this case this would be death. With the potential for wide variation in lag periods it is reasonable to suggest that the relationship with exposure may be different depending the type of cause of death, hence the need for investigation.

1.7.3 Communicable Diseases - Pneumonia

Many studies worldwide have demonstrated an association between air pollution and all-cause mortality,^{39,78} with the majority focusing on the link with respiratory mortality.^{7,108,109} Respiratory diseases such as pneumonia are communicable diseases or infectious diseases, as they are passed from person to person via contact or through airborne pathogens in the form of bacteria or viruses. In healthy humans, most pathogens are dealt with by the body's internal filter or cleaning systems. Once the pathogen has entered the alveoli region of the lung and made contact with epithelial surface of the lung wall, the body's innate defence mechanisms are initiated. This primarily comes in the form of pulmonary macrophages which are cells that find, engulf, and remove the foreign material a process called phagocytosis. If this fails or is impeded the epithelium cells can become inflamed and the alveoli are filled with fluid, impeding the region of the lungs responsible for the transfer of oxygen into the blood stream.

Respiratory tract infections particularly pneumonia are, according to the 2010 Global burden of disease study, the 4th most common infectious disease in the world.¹¹⁰ In 2011, pneumonia was the 6th leading cause of death in England and Wales in males at 10,824 deaths and 4th at 14,872 for females.¹¹¹ Community acquired pneumonia (CAP),

that is pneumonia caused by non-hospital based pathogens, is considered to be the leading cause of death due to infection within Europe.^{112,113} In the UK many underlying causes of CAP, both bacterial and viral have been identified. The bacterial causes leading to a hospital visit due to pneumonia are most commonly *Streptococcus pneumoniae* (35-48%), with less common cases being caused by *Mycoplasma pneumoniae* (2-17%), *Legionella pneumophila* (6-8%), and *Chlamydia pneumoniae* (3-6%).¹¹⁴ Approximately one third of CAP admissions to hospital are thought to result from a viral infection.¹¹⁵ Most investigations have identified Influenza A and B as the common viral causes,¹¹⁶⁻¹¹⁸ though in some studies Rhinovirus or Respiratory Syncytial Virus have been the most common.¹¹⁵ Hospital acquired pneumonia is primarily caused by *Staphylococcus aureus* or Gram-negative enterobacteria, and CAP is most commonly caused by *Streptococcus pneumoniae* (35% of CAP cases).¹¹⁹

Much of the evidence for the association between air pollution and general mortality has focused on exposure in a short time period – less than 40 days prior to death^{39,79}. This focus on short to medium exposure is appropriate for pneumonia as a cause of death. Pneumonia is generally an acute condition which progresses rapidly. Within 24 hours it can presents symptoms such as coughing, chest pain, shortness of breath, and fever and can generally be diagnosed reliably through medical consultation and a chest radiography.¹²⁰ The relatively quick onset of the disease, short diagnosis period, and the time varying nature of air pollution exposure satisfies all conditions required for comparing air pollution with pneumonia mortality.¹²¹ However, only a few studies - with limited findings - have specifically investigated associations between pneumonia related deaths and ambient air pollution. Schwartz and Dockery, indicated an increase in pneumonia mortality of 11% (95% C.I -3%,27%) per 100 μgm^{-3} increase of Total Suspended Particles (TSP).¹²² Halonen et al. demonstrated a percentage increase in pneumonia mortality in Finland of 3.16% (95% C.I -2.64%,9.32%) per increase in interquartile range of a 5 day Coarse Particle Matter ($\text{PM}_{10-2.5}$) mean.¹²³ Zanobetti et al. proposed that air pollution may be a predisposing factor to community acquired pneumonia (CAP) and that subjects with CAP rather than hospital acquired pneumonia, may be more susceptible to the effects of air pollution¹²⁴. Studies such as Neupane et al. have indicated a relationship between long-term exposure to air pollution and emergency visits to hospital with community acquired pneumonia.¹²⁵ However, so far

no study has attempted to investigate effect of pollution on deaths from community acquired pneumonia only.

The large number of causative agents may make identifying the true relationship between pollutant and outcome difficult as the many bacterial and viral causes may interact with the pollutions differently. Another question is whether the pollutant itself influences the likelihood of initial infection or reduces the body's ability to combat it. Additional confusion may be caused by the differing incubation periods. Streptococcus pneumonia has an incubation period of 1-3 days, shorter than other pathogens such as Haemophilus influenzae and Mycoplasma pneumonia with incubation periods of 2-4, and 6-32 days respectively.¹²⁰ To identify if any of these characteristics influence the underlying relationship between air pollution and pneumonia mortality, a comparison between different causes of death with different underlying mechanisms will be useful. Careful choice of comparison cause or causes of death is required. Ideally the comparison cause would be similar to pneumonia in that they have an acute onset, have a logical underlying biological mechanism linked to air pollution but, unlike pneumonia has a reduced number of potential causative agent's i.e. fewer than the bacterial, viral, and chemical associated with pneumonia.

1.7.4 Non-communicable disease – COPD

Chronic Obstructive Pulmonary Disease (COPD) is characterised by an irreversible restriction in air flow in the lungs due to inflammation or other damage to the airways which may also be accompanied by parenchymal lesions. Common symptoms include increased sputum production, shortness of breath, wheezing, chest tightness, and a chronic cough. Physical signs may include an overinflated chest, prolonged expiration, reduction in normal breath sounds, and sometimes the presence of abnormal sounds audible through a stethoscope. COPD is formally diagnosed through spirometry observing a ratio between a post-bronchodilator force expiratory volume in one second (FEV₁) to forced vital capacity (FVC) of less than 70%. Once diagnosed COPD patients are classified into four groups A, B, C, and D based on the patients air flow limitation severity, exacerbation risk, and symptoms. Air flow limitation is measured using a post-

bronchodilator FEV₁ and categorised into four groups GOLD 1-4 relating to FEV₁ ≥80%, 50-80%, 30-50%, and <30%. Exacerbation risk based on recent history indicating an exacerbation required change in day to day treatment (mild), medical intervention (moderate), and hospitalisation (severe). Symptoms are assessed using either the modified British Medical Research Council questionnaire with a score 0-1 (mild) or ≥2 (severe) or COPD assessment test (CAT) of <10 or ≥10. The patients are then categorised into A (low risk, less symptoms), B (low risk, more symptoms), C (high risk, less symptoms), or D (high risk, more symptoms).¹²⁶ Sufferers from COPD experience sudden increase in symptoms and these periods of acute exacerbation may occur at any time and may last for several days. A cure for COPD does not currently exist instead treatment focuses on reducing the severity and frequency of exacerbations. Management strategies have been proposed that are dependent on disease severity and the influence of any comorbidities.¹²⁶ The frequency and severity the exacerbations increase as the general level of COPD severity increases.¹²⁷

Genetic factors strongly influence the risk of developing COPD.¹²⁸ The most prominent causative factors, especially when in combination with the genetic risk factors, relate to foreign chemical agents entering the lung due to smoking, occupational exposure to dust or fumes, other respiratory infections, and air pollution.¹²⁹ Indoor air pollution produced by cooking or heating is a particular problem in the developing world where ventilation in properties is poor.¹³⁰ Other factors associated with increased risk and exacerbations of COPD are socio-economic status, gender, diet, co-morbidities, and airway hyper-responsiveness which is an abnormally sensitive bronchial airway.¹³¹

Rather than a direct causative agent, short-term exposure to outdoor air pollution is thought to influence COPD sufferers by increasing the severity of symptoms already present and consequently the rate of mortality. The majority of studies have looked at the influence of particulates. Sunyer et al (2000) showed an 11.2% (0.17%, 21.5% - 95% Confidence Interval) increased odds of mortality for a 20ugm⁻³ (IQR) increase in black smoke in those who already had COPD.¹³² Subjects previously diagnosed with COPD were observed to have an increased risk of 0.58% (-0.82%, 2.00%) per 10 ug m⁻³ increase in PM10 average over the same day and the day prior to death.¹³³ Similar percentage risks of 0.62% and 0.84% under similar circumstances i.e. PM10 increase,

have also been observed in those already suffering from COPD.^{134,135} With respect to the influence of particulates on mortality from COPD, in 2013 a meta-analysis was performed on 31 studies published across the developed world between 2000 and 2011, that indicated a pooled effect risk of 1.1% (0.8%,1.4%) per $10\mu\text{g}\text{m}^{-3}$ increase in daily PM10 with no indication of publication bias.¹³⁶

The influence of the gaseous pollutants on COPD has tended to be a secondary concern. In single pollutant models the effect of an increase in $10\mu\text{g}\text{m}^{-3}$ in SO₂ in the 48 hours prior been related to an increase of 1.0% (-1.0%, 2.9%), 1.6% (1.1%,2.5%), and 1.38% (0.92%,1.85%) in COPD mortality, and these effects hold when adjusted for other pollutants including particulates, but note they were performed in China where pollution is generally higher.^{43,137,138} . Similarly NO, NO₂, O₃, and CO have all been associated with an increase risk on COPD mortality though in some studies the effects weaken when other pollutants such as SO₂ and PM10 were modelled simultaneously.¹³⁷⁻¹⁴⁰

COPD was the 3rd most common cause of death in 2010¹¹⁰ and is a respiratory disease with proposed similar underlying exposure mechanisms to pneumonia. However, it is a non-communicable disease that has fewer causative agents most of which may take many years before leading the person to progressing to stages 3 or 4. With respect to acute exposure to air pollution it is more likely to exacerbate COPD symptoms rather than initiate it.¹⁴¹ This makes it a potential comparison with pneumonia, where exposure could initiate or exacerbate the onset of pneumonia. The differences will hopefully be seen in the delay and change in rate of mortality between exposure to pollution and the two causes of death. If pollution is exacerbating a condition already present then it would be expected that the rate of mortality should increase immediately before returning to the base mortality rate. If pollution is instigating the condition then there should be a delay whilst the underlying bacterial, viral, or chemical mechanism develops before seeing the rate of mortality increase more gradually to a peak before returning to baseline.

1.7.5 Non-Communicable Disease – Ischaemic Heart Disease (IHD)

Mortality from cardiovascular disease has been steadily declining in economically developed countries since the 1970s.^{142,143} Even so, the most common cause of death worldwide is still Ischaemic Heart Disease (IHD) accounting for approximately 13% of all deaths.^{110,144,145} This is in part due to the fact that increased fast food, smoking, drinking, and sedentary lifestyle in traditionally less developed countries have become closer to those in economically developed countries.¹⁴³ Ischaemic heart disease, also known as coronary heart disease, is typically characterised as reduced blood flow to the heart causing the heart cells (myocardial cells) to starve of oxygen resulting in a heart attack (myocardial infarction). Risk factors include genetic predisposition, smoking, obesity, poor diet, lack of exercise, stress, and the presence of comorbidities such as diabetes and hypertension. Prevention can be obtained through careful management of the lifestyle factors i.e. reduced smoking, improved diet, and increased exercise.

Higher levels of air pollution have also been identified as a risk factor for acute cardiovascular mortality in the form of cardiac arrhythmias, blood clots, and myocardial infarction.^{92,95,96} As mentioned earlier, small particles, after entering the cardiovascular system either directly through a healthy or an already damaged lung wall, may influence the heart in two ways. Either the foreign material impairs the blood flow itself adding strain to the heart and reducing its ability to provide oxygenated blood to the heart and other organs, or it directly impairs the autonomic nervous system.¹⁴⁶ The autonomic nervous system regulates, unconsciously, the internal organs and bodily functions such as digestion, breathing, and heart rate. It may be expected that air pollution will increase risk of mortality due to IHD very quickly within hours of exposure if the autonomic nervous system is responsible, or within several days if due to strained blood flow. This means unlike the respiratory diseases, there may be a double peak risk for IHD; once in the first few hours or days, and then again several days later.

The more general classification, ‘cardiovascular’ mortality is a major cause of death more commonly studied than the more specific IHD. The relationship between air pollutants, both particulate and gaseous, and cardiovascular mortality has been investigated extensively.^{31,79,92,147-149} Braga (2001) modelled across a lag of six days

(lag 0-5 days) and compared to same day mean (lag 0) and two day mean (lag 0-1), for an increase in $10\mu\text{g}\text{m}^{-3}$ of PM10 for 10 US cities.¹⁵⁰ Cardiovascular and myocardial infarction were both compared, with cardiovascular and myocardial infarction showing an immediate effect which dissipated quickly from lag 1 and lag 0, respectively. Kim et al 2003 similarly investigated cause specific mortality for 5 days including same day this time in Seoul, Korea (1995-99).¹⁵¹ In this study, effects of PM10 were expressed in relation to an increase equal to the interquartile range (IQR). The cardiovascular deaths had a slight delayed effect peaking 1-2 days after exposure before dropping away.¹⁵¹ This is in contrast to the results given by Braga 2001, which Kim explains may be due to differences in location and pollution levels, but might also be due to differing underlying modelling processes. Extending the lag period in studies helped to determine if cardiovascular mortality might show a double peak risk. Zanobetti et al. investigated the lag distribution across a 40 day lag period, and compared this with a mean lag 0-1 for cause specific mortality¹⁵² in 10 (APHEA-2) European cities.¹⁵³ The total percentage increase in cardiovascular mortality was 4.21%, with an immediate effect that drops rapidly over the first 10 days followed by a gradual rise to a peak at around 25 days.¹⁵² A recent meta-analysis comparing 19 suitably similar studies from Asia, North America, Australasia, and Europe which indicated a significant immediate 2.7% increase in risk of Myocardial Infarction for PM2.5 at lag 0 that dropped to 0.8% and 0.2% in lags 1 and 2.¹⁵⁴

Ischaemic heart disease has also been investigated,⁹² particularly with respect to fine particles (PM2.5) which are thought to enter deeper into the alveoli region. In 1996 Schwartz et al. showed a 2.1% increase in IHD mortality for $10\mu\text{g}\text{m}^{-3}$ same day change in PM2.5.¹⁵⁵ Effect sizes since then have varied. In 2006, with respect to 9 counties in California between Jan 1999 to Dec 2002, a $10\mu\text{g}\text{m}^{-3}$ of PM2.5 average over lag 0-1 was associated with a non-significant increase of 0.3%.¹⁵⁶ Yet Pope et al. indicated a significant 4.5% increase in risk associated with a same day change in PM2.5 that dropped rapidly in the next few days. Similarly PM10 also showed a significant though smaller, 2.0% increased same day risk that dissipated.¹⁵⁷ However, in both cases sample sizes were relatively small. More recently, studies from Beijing and Denver have indicated smaller effect sizes of 0.27% (95% C.I. 0.10%, 0.40%) and approximately

0.01% for PM_{2.5} increase at lag 0 with even a decreasing percentage relative risk observed for lags up to 4 days.^{158,159}

A limited number of studies have investigated the association between gaseous pollutants and IHD. In Seoul, Korea (January 1991 to December 1997) the risk of ischemic heart disease mortality was shown to significantly increase by 2.7% for each 1 ppm increase in CO and 1.5% for each 10 ppb increase in O₃, with no increase for TSP, NO₂, and SO₂.¹⁶⁰ A study in Vienna, Austria (2000-2004) showed a significant average effect over 7 days, with a small non-significant increase during lag 0-1 of 0.4%. Even so the strongest effect was still seen in PM_{2.5}.¹³⁹ Similar results were seen when the study was extended to two other cities in Austria Graz (0.5%) and Linz (1.0%).¹⁶¹ The influence of an average 10µg^m⁻³ increase in ozone (lag 0-1) and PM₁₀ (lag 0) on IHD in Moscow during a 3-year period (2003-2005) was calculated to be 1.61% (1.01-2.21) and 0.66% (0.30-1.02), respectively. A significant 1.6% increase is of note considering ozone production is related to air temperature and Moscow has a mean temperature of 9°C with a minimum of -25°C.¹⁶²

Being the dominant cause of death in the developed world and increasingly the developing world as lifestyles change, Ischaemic heart disease is clearly an important cause of death. The increased potential sample size, the differing underlying biological mechanisms and potential conflicting delayed risk structure all means it is of interest to compare results for IHD with those for the respiratory cause's pneumonia and COPD.

1.8 Defining the cause of death

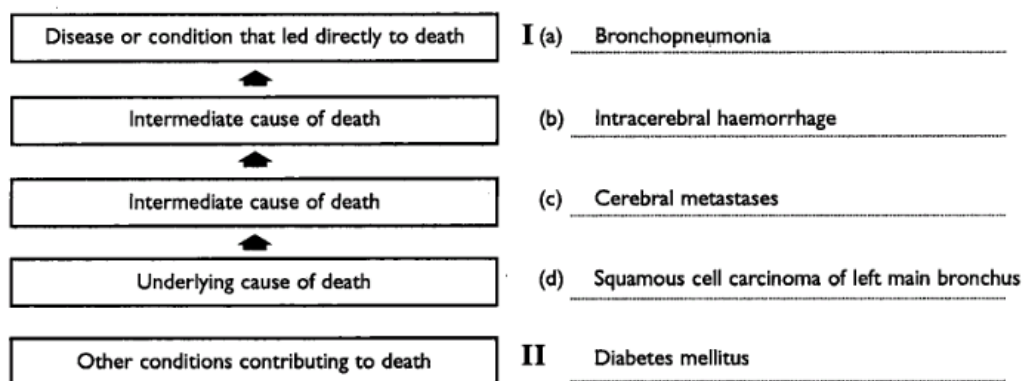
More than one cause can contribute to death and yet air pollution and temperature studies have commonly employed, when stated explicitly, the underlying cause of death identifying the subjects cause of death.^{30,80,163-165}

In the event of a death, the attending doctor fills out a medical certificate cause of death form, which along with demographic details, contains cause of death information. Further guidance has been provided to the doctor by the NHS, Office of National

Statistics, and the Scottish government regarding accurately completing the certificate, particularly with regards to the cause of death.¹⁶⁶⁻¹⁶⁸ In short, the attending physician is expected to fill out accurately and to the best of their knowledge a cause of death statement. An example of a certificate can be found in Figure 1.3, which has been extracted from the International Statistical Classification of Diseases and Related Health Problems (ICD) by the WHO.¹⁶⁹ The certificate comes in two parts, the first part relates the primary cause of death chain leading up to death, and the second relates to secondary conditions that may have contributed but were not directly thought to have caused death.

For Part 1, the physician fills out the chain of events directly leading to death starting with the ‘immediate’ cause of death i.e. the final cause directly leading to death. The immediate cause of death is written in the first field (I(a) of Figure 1.3) with each subsequent field representing the previous chain of events directly leading to death. The final field then represents the ‘underlying’ cause of death or the cause of death that began the chain of events. The example in Figure 1.3 reports four causes in the chain, designating Bronchopneumonia to be the final ‘immediate’ cause of death, and Squamous cell Carcinoma as the final ‘underlying’ cause of death. If only one cause of death was present then 1(a) Bronchopneumonia would be filled in, and the rest left blank. Bronchopneumonia would then be designated as the ‘underlying’ cause of death.

Figure 1.3 - Example cause of death statement from the ICD-10 Manual¹⁶⁹



Once completed, the death certificate is passed to the General Registry Office (GRO) in Scotland, or elsewhere, where it is processed and confirmed. If incorrect, which of the

stated causes is the true ‘underlying’ cause of death is identified based on the WHO recommendations in the ICD guidelines. Once processed, cause of death fields are set as ‘primary’ and ‘secondary’ where primary corresponds to the ‘underlying’ cause of death and the all other causes in no particular order are set as secondary causes of death. This makes the immediate cause of death in the dataset unclear.

In pollution studies investigating acute effects (e.g. within 30 days of exposure), the immediate cause of death field is likely to be most relevant. This may be particularly true for pneumonia which has a short incubation period between exposure and event. Pneumonia is often referred to as “the old man’s friend”, as it will commonly occur in already vulnerable subjects who are unable to fight off the infection. This means that pneumonia will often be the final ‘immediate’ cause in the days before death in subjects already suffering from another condition, for example cancer, that may have begun months or years earlier. This means that investigating mortality based only on the primary cause of death field may cause misrepresentation of the true relationship as we are excluding a number of events. This may be less likely in COPD and ischaemic heart disease, as COPD is more likely to be a long-term condition¹⁷⁰ and ischaemic heart disease is less likely to have preceding events in a chain. However, both are likely to have co-morbidities present, which may be included in the chain of events. In order to investigate the link between air pollution in the short period prior to death and these causes of death the immediate cause of death field would be appropriate.

With that in mind, one aim proposed at study conception was to compare the pollution effect on cause of death when cause of death was identified by both the ‘underlying’ and the ‘immediate’ cause of death designation. Mortality datasets that identify the ‘underlying’ cause but not the ‘immediate’ cause of death appears to be standard procedure when following the ICD coding procedure. However, an attempt was made to acquire an equivalent dataset for three English cities from The Office of National Statistics (ONS). ONS indicated that they would be able to provide a dataset with ICD coded ‘primary’ and ‘secondary’ causes before and after the manipulation that would normally remove the identification of the ‘immediate’ cause. Meaning this second dataset would allow for a comparison between the ‘underlying’ and ‘immediate’ cause of death. Even though every effort was made to acquire this dataset, unforeseen

administration delays at the Health and Social Care Information Centre (HSCIC) meant it was not possible to obtain the English data within time. A more detailed description of the procedure undertaken to acquire the data from Scotland and the attempts to acquire the equivalent English data have been provided in Appendix A.

With respect to environmental epidemiology, only one paper in 1990 is known to have separately examined 'immediate' and 'underlying' causes of death and two papers have included associated/contributing causes of death i.e. they used the primary and secondary fields. Katsouyanni et al. investigated cause specific mortality in Athens between 1975 and 1982.¹⁷¹ The study compared 199 high pollution days ($\text{SO}_2 > 150 \mu\text{g m}^{-3}$) with match 2*199 low pollution days ($< 150 \mu\text{g m}^{-3}$). Two blinded medical practitioners identified the underlying and immediate causes of death from their death certificates, identifying cardiac, respiratory, and other causes. The excess percentages of deaths in high pollution days were similar, with cardiac causes decreasing slightly from 0.58% to 0.35%, and respiratory increasing slightly from 1.99% to 2.37%, when labelled as 'underlying' and 'immediate' cause of death respectively. Schwartz (1994), similarly compared the number of cause specific deaths for underlying and contributing causes on the highest and lowest 5% TSP concentrations days between 1975-1979.⁸³ In reality only one paper which has focused on heat related mortality is known to have compared underlying only, and underlying and associated, cause of death for circulatory and respiratory disease in Sydney Australia.¹⁷² Circulatory effect sizes and standard errors changed very little between underlying only and underlying and associated together whereas respiratory effects and standard errors were smaller when underlying and associated were combined. This may be a reflection of change in respiratory sample size from underlying on (7.5% of all cause) to underlying and associated (22.6% of all cause) whereas circulatory was a 10% increase.

Even with limited evidence available, careful choice of the cause of death field was required, especially in a study investigating the short-term relationship on specific causes of death.

2 INTRODUCTION TO ANALYSIS METHODS

2.1 Chapter introduction

Chapter 2 provides the background of the methods that may be applied in this thesis, beginning with a reminder of traditional study designs, an introduction to the time-dependent study designs, and the data characteristics and confounders. This is followed by a description of the most common methods used to model the dependent variable; in this case number of deaths per day or day of death. These modelling methods can be categorised into two groups, time-series models and the case-crossover design. The independent variables, in this case pollution and temperature exposure, are introduced and their relationships with the dependent variable are discussed. This begins with the dose-response aka exposure-mortality relationship and then outlines methods of modelling the short-term time delayed effect expressed as a ‘lagged’ effect. Missing exposure data will likely impact on both the dose-response and the lagged-response effect, hence the final section of chapter 2 will provide information regarding the analysis of data containing missing observations. Finally, the study objectives will be outlined.

2.2 Traditional study design

Epidemiology is concerned with monitoring the occurrence and identifying the causes of ill-health or disease in a population. Confirming a causal link between exposure and a disease outcome is extremely difficult with many confounding factors present. A confounding factor is a secondary characteristic that influences both the exposure and outcome. It would ideally be accounted for through randomly allocating subjects to an exposure group and then monitoring the outcome. This is often not practical or ethical and so observational epidemiological research studies are employed. These most commonly compare the exposure in two separate population groups at the same time or during a common time period. In epidemiology, three of the most commonly employed study designs are; a cross-sectional design, a cohort design, or a case-control design.

A cross-sectional study involves observing and collecting data at one specific time-point on an entire population or sub-sample containing the same characteristics as the population. Cross-sectional designs lack a time varying component and the main focus is to determine the prevalence of a disease rather than the incidence. If the event of interest is rare or requires subject participation through a questionnaire, surveying the entire population may not be practical or cost effective except when the data is routinely collected.

A cohort is a group of individuals chosen because they share similar characteristics such as place of residence, gender, ethnicity, or exposure level. The subject cohort is followed over a set period of time during which the occurrence of an event is measured or counted. Multiple cohorts that differ by at least one characteristic are followed at the same time and compared by calculating the risk or incidence rate. Cohorts can be defined either by circumstance, called a 'natural' experiment, or by design through random assignment called a 'true' experiment. Subjects may leave the cohort when the event occurs or if the subject is lost to follow up due to an unrelated cause. If the size of the cohort is fixed at outset the design is considered 'closed', but if any subject who leaves is replaced then it's an 'open' design. Incidence rates and risk ratios are calculated directly by dividing the number of events by the amount of person time at risk which is reported as rate per year, or more commonly by fitting a Cox-proportional hazards model.¹⁷³ Cohorts allow for multiple events to be recorded at once or multiple exposures for a single event type.¹⁷⁴

In observational epidemiology, the case-control design samples subjects from the same population who have the event of interest (cases), and without the event of interest (controls). Within the two groups the numbers of subjects exposed and unexposed are then compared. Ideally controls are randomly sampled from the population or with stratified random sampling matching to cases on confounding characteristics. The odds ratio is then calculated using a logistic regression model or by calculating $(a*d)/(b*c)$ where a=number of cases exposed, b=number of cases unexposed, c=number of controls exposed, and d=number of controls unexposed. The odds ratio can be equivalent to the risk ratio when the outcome of interest is rare or when the case and controls are

representative of the exposure history in all people in the population with or without the disease respectively. The accuracy of the estimate is improved by increasing the number of controls. As only subjects with the event or their equivalent controls are recruited, case-control studies typically have a reduced cost meaning a larger initial population can be used increasing the number of potential cases and controls.¹⁷⁴

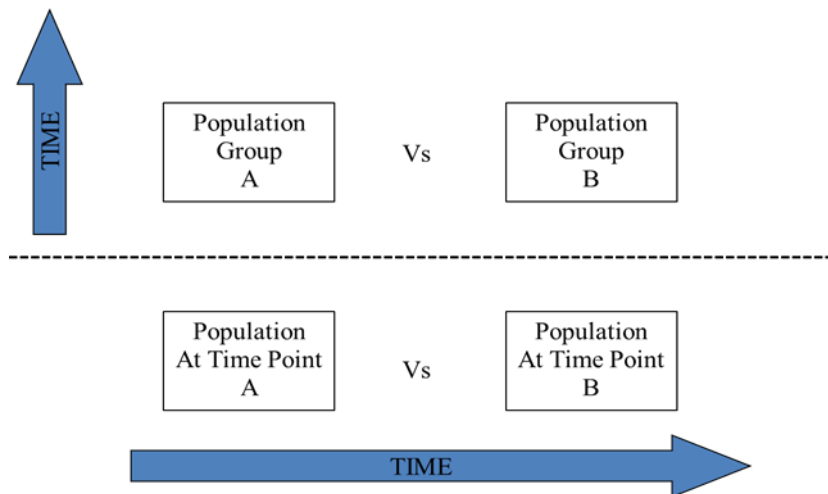
Cohort studies and to a lesser extent case-control studies have been employed in the air pollution context to determine the effect of pollution over long time periods e.g. a 30 year case-control study.¹⁷⁵ Early cohort studies by Abbey et al. (1991), Dockery et al. (1993) and Pope et al. (1995) reported the effect of pollutants in 6000, 8000 and 500,000 subjects across 10, 16 and 9 year periods, respectively.¹⁷⁶⁻¹⁷⁸ Dockery and Pope found significant associations for PM_{2.5} whereas Abbey only found an association with PM₁₀ within females. The large sample size, long time periods required and cost in time and money means only results of extensions or reanalysis of these cohorts have been reported.¹⁷⁹⁻¹⁸¹ Cohort studies have a large impact on public policy and so it been argued that the cost of implementing air pollution policies would justify the cost of a cohort study.¹⁸² Hence European studies with similar results have more recently been reported.^{45,183,184}

2.3 Traditional study design vs Time series study design

As illustrated in the first half of Figure 2.1 confounding factors in traditional studies tend to be subject specific differences between the two populations groups. Once identified and recorded these confounding factors need to be accounted for in the model as covariates or through weighting.¹⁸⁵ The traditional study designs have tended to focus less on time-dependent factors though adjusting for change in baseline covariates over time in conjunction with the outcome has become more prominent.^{186,187} However, including them as covariates has its own challenges that are yet to be fully addressed.¹⁸⁸ Even so, time-varying factors tend to be short-term within subject changes and so subject specific confounding has been the priority.

In this thesis we are studying the short-term effects of air pollution where time varying confounding factors are present.¹⁸⁹ To investigate the change over time the data is presented in terms of change in exposure on the population over a period of time, in this case short-term (<30 days). So called ‘time-series’ studies differ in that they compare the same population group at two or more different time points (time point A and time point B) as illustrated in the second half of Figure 2.1. If the same subject group is studied at each time point any subject specific factors are no longer the primary confounding issue. Instead non-subject specific time-varying factors become important. These factors may relate to short term changes such as day to day changes and day of the week differences, or long-term changes such as monthly, seasonal, or long term time trends over years. Identifying and adequately controlling for time varying confounders is an important part of any epidemiological study employing time-series data but is particularly important in air pollution studies.¹⁹⁰

Figure 2.1 - Traditional study design vs Time dependent study design



2.4 Data characteristics

The primary outcome in air pollution studies relates to the number or count of deaths (mortality studies) or hospital admissions (morbidity studies) on each individual day within the study period, the whole dataset being called a time-series. The proportion of the overall population who die or enter hospital on any individual day is very small and so classed as a rare event. The number of daily deaths/hospital admissions is a non-negative integer count, Y_t , that can be modelled using a Poisson process with an expected mean that varies over time t . Under standard Poisson distributional properties the variance is equal to the mean, but due to heterogeneity in the population mortality datasets in reality often have a variance that is greater than the mean; this is called over-dispersion. Data measured on consecutive days are not considered independent of each other as a high mortality day is likely to be surrounded by similarly high mortality days. This correlation in the data is called serial correlation or autocorrelation.

2.4.1 Confounders

A confounding factor is a secondary characteristic that influences both the exposure and outcome. Identifying and accounting for confounders are important in any study. Due to the study designs employed subject specific confounding is largely matched out, and instead comparisons between different time points make time-dependent confounding more important. The time-dependent confounders include:

- *Long-term time trends*; Any two variables that contain a long-term trend irrespective of a causal link will be correlated so it is important to allow for this. Improvements in public health, pollution abatement policies, and technology may influence the exposure level and mortality rate over long time periods.
- *Seasonal variation*; Mortality counts and pollution levels will vary over the course of a year due to seasonal weather changes, and increases or decreases of population activity levels at different times of the year. Seasonal variations tend to be periodic but may also alter between years.
- *Day of the week & Public holidays*; daily differences in population activity differences e.g. rush hour traffic means pollution and mortality rates are affected.

In addition to time-dependent factors, variations in meteorological conditions are important confounders. Short-term meteorological variation has a significant impact on both daily mortality/morbidity and pollution levels. Combinations of covariates representing temperature and humidity are commonly included in statistical models. Weather variables such as these may have a non-linear relationship with mortality; in particular temperature research has suggested a V, U or a reverse J shape relationship indicating increased risk at the extreme temperature ranges.¹⁹¹⁻¹⁹³ Alternatively variables have been included that identify extreme temperature and humidity days. To be confident that short-term variation in mortality is due to changes in an air pollutant these confounding factors must first be accounted for.¹⁹⁴

2.5 Time series designs

Mortality is a rare event¹⁹⁵ that, except in a few early studies¹⁹⁶⁻¹⁹⁸, has traditionally been modelled using the Poisson distribution:

$$Y_i \sim \text{Poisson}(\lambda_i)$$

where Y_i represents the daily number of deaths, and λ_i is the mean or expected mortality count $E(Y_i)$.

In early studies the analysis assumed Gaussian distributional properties for Y .¹⁹⁶⁻¹⁹⁸ Moving averages removed seasonal and long-term time trends from the mortality data with daily outcome replaced with the mean count of days either side, known as a moving average window.¹⁹⁷ In data where the mean number of deaths is large a Gaussian approximation is viable, however most mortality studies have a mean number of daily deaths that is too small. Gaussian distributed data also assumes a variance that is independent of the mean, whereas mortality data contains characteristics (Chapter 2.4) that means a Gaussian approximation is not practical,⁷ and so Poisson models were developed.

2.5.1 Poisson regression model

The Poisson Regression model fits the log expected mortality count Y_i for the i^{th} day, such that:

$$\text{Log}(E(Y_i)) = X_i \beta$$

where X_i represents a matrix of covariates on the i^{th} day. Two concerns associated with this data are over-dispersion and serial correlation. These concerns can be accounted for by using generalized estimating equations to estimate the Poisson regression

model.¹⁹⁹ In these models a changeable covariance structure is assumed such that the covariance matrix is:

$$\alpha A^{1/2} R A^{1/2}$$

where $A_{ij} = E(Y_i) \delta_{ij}$, δ_{ij} is = 0 when $i \neq j$, and 1 otherwise, α is the over-dispersion parameter and R is an auto regressive matrix. This will give more efficient results as it includes the covariance structure when estimating the regression coefficients, and variances. Also, by calculating robust variance estimates unbiased hypothesis tests are given even if the covariance is incorrectly specified.¹⁹⁵

In air pollution studies, a model is often fit such that as much variation due to confounding as possible is accounted for prior to adding in any pollution variables.^{200,201} This results in pollution effect estimates dependent on the method employed to account for confounding that has been driven by the data and does not account for the pollution effect. Without a common modelling method any combination of within study sites or comparison between study sites becomes compromised. The pollution effect estimates themselves may be conservative as when pollution and other variables are correlated accounting for as much variation as possible in the other variables may inadvertently remove some of the effect that is due to pollution.

How the model accounts for confounding factors defines the type of model that has been used. Early attempts, such as Schwartz (1991) fitting Total Suspended Particles (TSP) to daily mortality in Detroit, included a continuous time variable alongside categorical variables identifying the seasons and years as covariates. Weather variables have included; same day 24hr mean temperature, dew point temperature, along with identifiers of extreme weather days.²⁰⁰ Interaction variables between season and weather allowed for evaluation of different weather effects in each season. Once non-pollutant variation in the data were accounted for, the pollution variables were added and indicated a relative risk of 1.04 per $100\mu\text{g}\text{m}^{-3}$ increase. At this time it was policy to collect TSP readings once in every six days. Even after imputing the missing with a regression model, missing data means limited information particularly for key confounders such as day of the week.

The main concern here and in similar studies¹⁹⁵ is the likelihood of biased estimates of relative risk as indicator variables of year and season are unlikely to be adequately representing the existence of a more complex relationship over time and across seasons in mortality rates and pollution levels. These factors, whether or not they significantly predict mortality need to be accounted for adequately before an unbiased pollution effect estimate can be calculated.

Hence, Schwartz (1993) fitted more complex Poisson regression models to mortality data from Birmingham, Alabama and Philadelphia.^{7,202} In both cases the model controlled for long-term trends and seasonal waves in cycles of 2 year and one month periods, by including 24 sine and 24 cosine terms alongside other time and meteorological confounders similar to those describe in the previous Schwartz (1991) paper. This fully parametric method of accounting for seasonal trends was described further within the APHEA project,^{203,204} the sinusoidal waves included took the form:

$$\alpha \sin\left(2\pi \frac{k}{365} t\right) + \beta \cos\left(2\pi \frac{k}{365} t\right)$$

here t represents the day of the study, α and β the amplitude of the wave, and $2\pi t k/365$ affects the frequency of the wave with k representing the wave cycles taking values of 1,2,3,etc. Models that have included trigonometric waves to account for wave cycles in terms of 12, 6, 4, 3 and 2 months have also included time in days as a linear and quadratic variable to account for long-term trends.^{43,205} Determining the value of k - the frequency of wave cycles - is important in accounting for seasonal patterns in the data. The disadvantage of using trigonometric waves is that they assume that the wave is the same height and located at the same time point every year. This is unlikely in practice, particularly in studies that span several years during which changes in public policies regarding pollution emissions, improvements in emission reducing technology, and possible influenza epidemics will alter patterns in both mortality rates and pollution levels.¹⁹⁴ Though variation in the data were accounted for, to improve the representativeness of the true variation a more flexible method was required.

2.5.2 Generalised Additive Models (GAM)

As before, the Poisson regression version of this model fits the covariates X_i against the log expected mortality count Y_i for the i^{th} day. The Generalised Additive Model (GAM), originally proposed by Trevor Hastie and Robert Tibshirani,⁸ replaces the linear predictor terms $\sum_i \beta_i x_i$ with an unspecified non-parametric function $s_i(x_x)$. A GAM is then defined as containing at least two non-parametric smoothing components. Confounding factors X_1, X_2 , that have a non-linear relationship with daily mortality are here modelled as unspecified non-parametric smoothing functions, such that the model becomes:

$$\text{Log}(E(Y_i)) = X_i \beta + s_1(X_1) + s_2(X_2) + \dots$$

where $s_1(X_1), s_2(X_2), \dots$ represent smoothing functions. These factors might include seasonality, long term trends using calendar time, or non-linear weather variation in temperature, humidity, barometric pressure, etc.

Air pollution time-series data has two characteristics that are suitable for applying a generalised additive model (GAM). The data has confounding factors that require being modelled by at least two non-parametric smoothing functions, and the data will produce small regression coefficients.²⁰⁶ Using multiple smoothing parameters meant the GAM can suitably account for both short and long term trend variation, unlike the previously described parametric methods that made too many inappropriate assumptions to be able to flexibly account for the true relationship.

The first attempt to fit a GAM in air pollution was in the already mentioned Schwartz (1993) paper using Birmingham, Alabama data. Using a LOESS smoother, the GAM was used to model the non-linear relationship between the weather terms and mortality.⁷ A LOESS smooth or locally weighted regression²⁰⁷ is similar to a weighted moving average in that it fits a low degree polynomial regression, usually linear, to a subset of the data called the 'span' around a mid-data point, the data closest to the mid-point

being weighted higher than those further away. A weighted regression model is fitted to each mortality data point (day within the study period) and its span, which produces a fitted value to replace the raw value. Here, LOESS smoothing was fitted to the residuals of the model once all other covariates were fitted; a linear relationship with PM₁₀ is then subsequently estimated. The model result was compared with the fully-parametric model described in the final paragraph of Chapter 2.5.1. Results indicated that the GAM did detect a non-linear relationship between weather and mortality; however the improved fit made little difference to the PM₁₀ effect size estimated by the fully parametric model.

The amount of smoothing is controlled by the width of the span with a wide span increasing the level of smoothing compared to the true nature of the data. Ideally the span will adequately remove seasonality and non-linearity in the weather covariates without removing variation due to the pollutant. The span is usually reported in terms of the fraction of the data they represent. To control for season and time trends across calendar time the span will depend on the length of the study but typically has 2 or 3 degrees of freedom per calendar year. The exposure range (min to max) in weather data tends to be similar across studies and so to control for non-linear weather variation, spans of 0.5-0.8 of the data range are often applied^{147 208}.

As an alternative to the LOESS smooth, studies such as Kelsall et al. (1997) have fitted a GAM using interpolating smoothing splines.¹⁰ Here the following equation representing a penalised residual sum of squares is fitted to the data such that $f(x_i)$, a generic function from the class of functions called splines, is minimised.

$$\sum_{i=1}^n [Y_i - f(x_i)]^2 + \lambda \int [f''(x_i)]^2 dx$$

Where λ is a fixed constant value greater than 0 that controls the level of smoothing: the greater the value the larger the amount of smoothing. The first part of this equation compares the fitted function to the data (sum of the squared difference) and the second restricts the curvature of the function to better represent the relationship with the outcome.⁸ Smoothing splines are fitted using an iterative estimation based on a back

fitting algorithm to give an approximate solution to the penalised residual sum of squares.²⁰⁹

The smoothing spline is a special case of a penalised-spline,²¹⁰ where in a penalised spline the number of knots and their position are important considerations regarding smoothing. The effect of the knot location can be minimised by increasing the number of knots used.²⁰⁹ In a smoothing spline the degrees of freedom defines how much smoothing occurs a greater number the more flexible the smoothing with one degree of freedom representing a linear regression and all variation removed. As an example Kelsall et al. applied 160 degrees of freedom to calendar time this was equivalent to including an indicator variable in the model for every 31 days.

The flexibility of the non-parametric smoothers to account for non-linear relationships over time and within weather covariates has meant that the generalised additive model has been a popular method of modelling the pollution-mortality^{211,212} and morbidity relationship.^{147,213} However, the GAM has come under some criticism. Initially this was due to concern that a lack of suitable constraints in the default convergence criteria used by a GAM procedure in S-Plus meant a failure to produce unbiased estimates. Work on a simulated dataset followed by a reanalysis of the NMMAPS dataset indicated that a GAM using a LOESS smooth and the default convergence criteria produced an overestimated effect size.²⁰⁶ The estimate was improved when more stringent convergence criteria were employed. However bias may still exist especially when concurvity exists in the covariates. Concurvity occurs when a non-parametric function of a covariate can be approximated by a linear combination of other non-parametric functions of covariates. Concurvity is equivalent to multi-collinearity in a standard linear regression. If concurvity is present, as with multi-collinearity, effect estimates can be unstable. However, this does not show itself in the variances produced by the S-Plus GAM procedure, where variances are underestimated and hence increasing the likelihood of a false positive result. Under the default and proposed constrained convergence criteria the standard error was shown to be 23% and 16% lower than the true estimate.²¹⁴ This lead to the Generalised Linear model being proposed as an alternative.

2.5.3 Generalised Linear Models (GLM)

Generalised Linear Models (GLM) are well known in statistics. Here we refer to the particular GLMs applied within the air pollution field. A fully parametric version of the GAM, the non-parametric smoothing terms are replaced with regression splines where a piecewise polynomial is fit to the data. Individual polynomials are fit to multiple regions of the data which are then smoothly connected at predefined locations or *knots* within the range of the dataset by constraining the first and second order derivatives to be continuous.²¹⁵ The flexibility to represent the data, and the level of smoothing, defined as degrees of freedom, are controlled by the number and location of the knots; the smaller the number of knots the smoother the data fit. Knots can be located anywhere within the data range but are usually placed at equally spaced quantiles.

One popular choice is to use natural cubic splines, where 3rd degree polynomials are fitted between each knot (k =number of knots) of the variable X_i

$$\sum_{h=0}^3 \beta_{0h} X_i^h + \sum_{j=1}^k \beta_{i3} (X_j - X_{ij})_+^3$$

where if $(X_j - X_{ij})_+$ is greater than zero it is then set to equal zero. Further constraints are applied such that a linear relationship is set at the very ends of the data ranges, and the second and third derivatives are set to equal zero.⁸ The advantage of using regression splines is that standard linear model estimation procedures used in usual generalised linear models can be employed, hence the parametric aspect of the model⁸.

As with all of the models described, the main concern in a GLM is determining the number and location of knots to employ, as the optimum representation of seasonal and long-term confounding is important. The level of smoothing employed to account for confounding is an important consideration in all models including the trigonometric functions, GAM and GLM. Studies have taken different approaches to find the optimum smoothing parameter, such as minimising the Akaike information criteria (AIC),²¹⁶ removing any autocorrelation in the residuals,²¹⁷ or using previously established information in earlier studies.^{39,218} In the majority of cases the smoothing

parameter for time is determined using residual diagnostic plots to remove autocorrelation, and for the weather variables it was determined by minimising the AIC.^{153,219} Even so, it is still difficult to assess whether or not the model has adequately controlled for confounding. A model that has adequately accounted for confounding in one dataset may not when repeated in another based on a different location even if that location is within the same country. Effect estimates are sensitive to the model design, creating a potentially unique model for particular location that is not applicable elsewhere.²²⁰

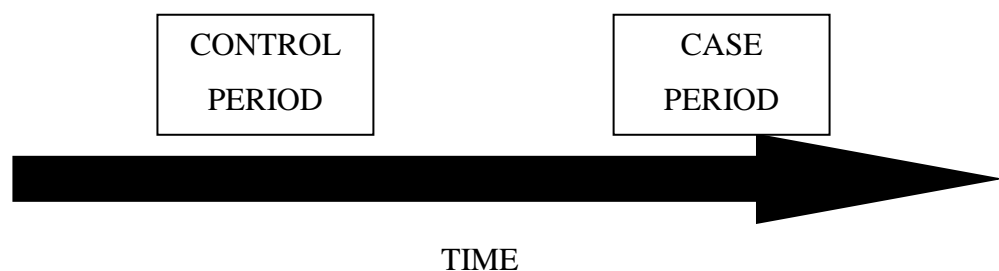
Unlike traditional epidemiological designs where comparisons are made between groups of individuals, these time series models do not allow for direct analysis of effect modification by, for example age or gender. This technique is relatively cost free as it allows the use of large publicly available datasets of independently collected mortality and morbidity data, and atmospheric weather and pollution data.²²¹

2.6 Case-crossover design

In 1991, Malcolm Maclure introduced an alternative to time-series modelling in the case-crossover design.²²² Originally he proposed using it to assess the immediate causes of a myocardial infarction during the morning, and explained that it was conceived from necessity due to low availability of suitable control subjects. It became clear that the most suitable control subject was in fact the subject themselves. The main advantage of the case-crossover design is that by careful choice of the controls a number of confounders, both subject specific and time-dependent can be accounted for within the design rather than through manipulation of covariates in the model.

The case-crossover design is an amalgamation of the matched case-control design (see Chapter 2.2) and the crossover design.⁹ In a crossover design two subject groups are defined and exposed to differing levels of a factor of interest. After a predefined period of time the two groups switch, usually after a 'washout' period, and are compared. In the case-crossover design, the subject acts as both case and control, where the case and control periods of time defined for each subject (Figure 2.2). The exposure levels leading up to the case and control time periods are then compared.

Figure 2.2 – Basic representation of the case-crossover design over time



The case-crossover design is ideally suited when the event is considered to be rare, has an acute onset and the exposure of interest is intermittent with a transient effect.²²³ So it is ideally suited for studies investigating short-term air pollution effects, comparing the varying levels of ambient air pollution on health effects that occur and are diagnosed

quickly and commonly result in death. The case-crossover design is therefore a useful tool and has been applied in many situations^{11,163,224,225}.

The case-crossover design can be analysed in terms of a proportional hazards model as proposed by Navidi 1998,²²⁶ such that the probability a subject i fails at time t_j is given by;

$$\Lambda_i(t_j) = \lambda_i \exp(\beta^T X_{ij})$$

where, X_{ij} is the vector of covariates corresponding to the i^{th} case at the j^{th} time point, and the vector β represents the log relative risk associated with a unit increase in corresponding components of X_{ij} . To compare within the same person at different time points the design assumes that the baseline hazard λ_i is constant over time for each individual but that it varies randomly between individuals.²²³ This can be formally analysed using a conditional logistic regression model, and as before adequate adjustment for confounding is important. Where smoothing techniques are employed, the case-crossover can adjust seasonal and long term trends by choosing an appropriate set of controls time periods.

2.6.1 Control/referent selection strategies

As noted previously the subject acts as both case and control, where ‘case’ is a time period during which the event of interest took place and the ‘control’ is a similar time period during which the event did not take place. Maclure originally proposed two methods of control selection.²²⁷ Exertion levels one hour prior to myocardial infarction (case) were compared with the subject’s usual frequency of exertion over the year prior to the event, and the level of exertion in the same hour one day prior (control). Using the same subject for both case and controls the design naturally accounts for any subject specific confounding factors whether or not they can be measured. In this case the level of physical exertion is determined by interviewing the patient, meaning there is the potential for error due to recall bias. Recall bias is not an issue in the air pollution studies as exposure levels are independently measured. Time-dependent confounding²²⁸

can be accounted for through careful choice of controls. Increasing the number of controls will improve efficiency but bias will be incurred by choosing unrepresentative controls, resulting in a trade-off between maximising efficiency and reducing bias.²²⁹ For example, choosing all available days within the same week or month will increase the sample size and improve power yet it will also increase bias due to overmatching as pollution on adjacent days contains autocorrelation.²³⁰

Several control sampling strategies have proposed, each with varying statistical properties. Janes et al. 2005 classified these methods based on statistical characteristics of the control sampling strategy in particular their ability to account for bias.²³¹ The control strategies are assigned into; localizable versus non-localizable, and ignorable versus non-ignorable. A design is considered to be 'localizable' if the estimating equation, or likelihood, of the case time period conditional on the control time period contains unbiased information about the effect size. This is a desirable property as the estimates are unbiased within the time period that the controls are obtained. A 'non-localisable' control design is one where the estimating equation is considered to be biased and the likelihood conditional on the control time period does not provide information regarding the effect size. This may occur if the position of the case time relative to the control time does not provide any information, for example the case is set to be a fixed specific distance from control time for all subjects. The localizable control designs can then be further split into two groups, 'ignorable' and 'non-ignorable'. In an 'ignorable' design, the conditional logistic regression will give unbiased effect estimates when the control sampling strategy is ignored. In a 'non-ignorable' design the control sampling strategy cannot be ignored and the estimating equation will depend on the strategy used. In order to gain unbiased estimates the strategy must be included in the likelihood equation.

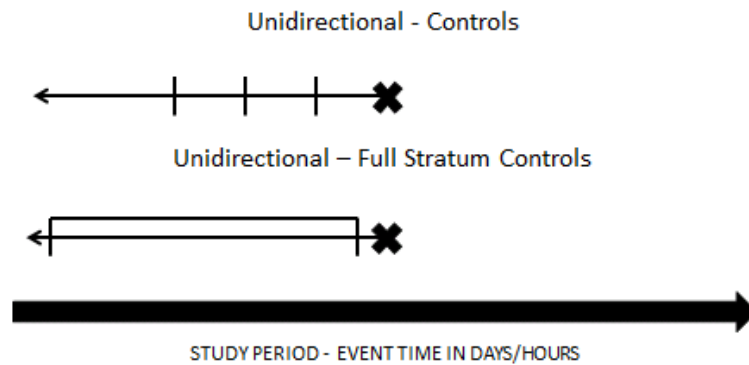
2.6.2 Unidirectional control selection

In a ‘full stratum retrospective’ or ‘total history’ unidirectional design controls are either an average exposure level of all days in the study period prior to the case day,²²⁷ or every day prior to the event set as a separate control.²²⁶ Alternatively, controls can be selected at fixed time points before the case day such as the day before,²²⁷ two, three or four days before²³² or even the 60th, 90th, 180th or 365th day before.²³⁰ Figure 2.3 gives a graphical representation of the unidirectional and full stratum unidirectional case-crossover design where X represents the case time point (usual a particular day or hour) and | the control time point.

As with GAM and GLM long-term trends and seasonal differences in exposure mean that time distant controls and controls chosen at large intervals such as the 180th or 365th day will bias the result. If cases tend to occur in high air pollution seasons choosing controls in other seasons would mean their exposure levels are likely to be lower.²³⁰ Whereas in GAM/GLM these issues were accounted for by covariates in the model in the case-crossover design they can be accounted for by restricting the control periods. Janes 2005 suggested that controls are restricted to short time windows around the case day.²³¹ Further, to avoid autocorrelation in exposure on adjacent days to the case day and ‘day of the week’ effects, a 7 day separation between controls is suggested i.e. 7, 14 or 21 days.²²⁴

The full stratum and fixed unidirectional design are non-localisable and are susceptible to bias if there are time trends in the exposure level. A unidirectional control design was first applied in air pollution by Lee and Schwartz (1999) to analyse the TSP effect on mortality in Seoul, Korea,²²⁴ which had previously been analysed using a GAM.²³³

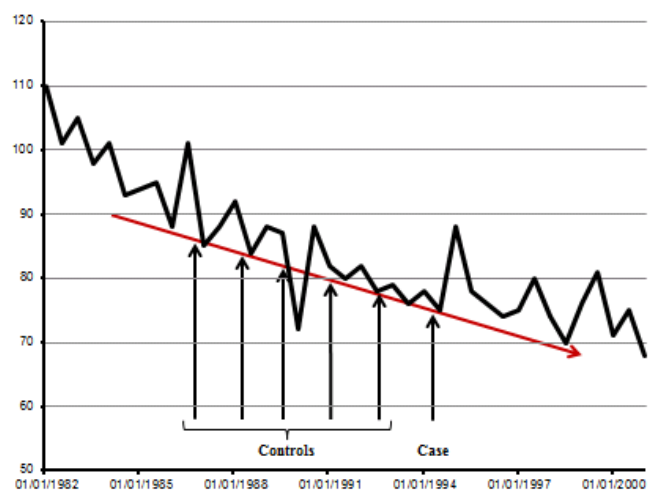
Figure 2.3 – Graphical description of the unidirectional and full stratum unidirectional case-crossover designs where X is the case day.



2.6.3 Bidirectional control selection

If time trends in exposures are present, the unidirectional design is susceptible to bias due to a constant difference in the exposure prior to the case than during the case period.²³⁴ A fabricated example of this can be seen in Figure 2.4, where a downward trend in exposure causes all (preceding) ‘controls’ to experience a greater exposure than the ‘case’. At the very least this would result in underestimated effect sizes.

Figure 2.4 – Example of how long-term trends would cause bias in a unidirectional case-crossover design when exposure is consistently different in control days compared to the case day.

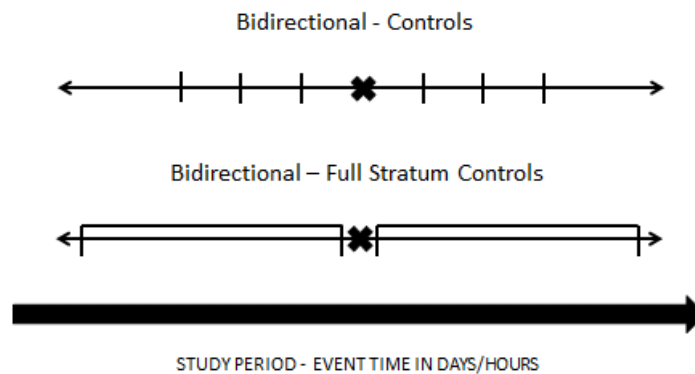


To account for this bias a bidirectional case-crossover design was proposed, where controls are sampled pre and post the case day.²²⁶ Sampling control days from post the event in studies where the event of interest is mortality might not be considered logical. Navidi argued that independently measured exposure data for the entire population meant that we know the exposure level regardless of whether or not the subject died, that exposure level would not be affected by the event of interest. This is acceptable in a study regarding mortality. However there is a chance that it may not be valid when morbidity is the event of interest, as subjects might spend their time post the event indoors at home or in hospital experiencing reduced exposure. Lumley and Levy 2000, showed that the bias caused by including post event controls is negligible when the case-event is rare (as with mortality), and that the bias from unidirectional design with exposure time-trends would be more substantial.²²³

The bidirectional design proposed by Navidi was a full stratum bidirectional design where controls were every pre and post the case day within the study period.²²⁶ In this situation the full stratum bidirectional design is regarded as a localizable, ignorable design²³¹ which controls by design for time trend bias. However, in large studies controls in a full stratum design will not take into account seasonal confounding.²³⁵

A popular alternative is the symmetric bidirectional design,²³⁶ where controls are chosen at identically spaced fixed time intervals pre and post the case day.^{132,237-241} At short intervals e.g. 7, 14 and 21 days^{122,163} this should reduce bias caused by time trend, season, and day of the week so long as the time between case and controls are multiples of seven.²³⁶ However, bias was observed when the seasonal wave cycle was incomplete or not symmetrical such as would be the case if near the beginning or end of the data time frame.²⁴² As the case day is a fixed distance at the centre of the control days it provides no additional information for the effect estimate, hence the design is non-localisable.²³¹ Figure 2.5, reports a graphical example of the common bidirectional case-crossover designs where X represents the case time point (usual a particular day or hour) and | represents the control time point.

Figure 2.5 - Graphical description of the bidirectional and full stratum bidirectional case-crossover designs where X is the case day.



2.6.4 Semi-symmetric bidirectional controls

In a semi bidirectional design, a single control day is randomly picked from days within a fixed time period pre or post the case day.²⁴³ If only one day is available, for example if the case day is at the study beginning or end then the control day available is chosen. This is not classed as a time-stratified design; however it can become one by continuing the randomisation of the control day even at the beginning or end of the study period with ‘no’ control day possible.

The semi-symmetric design is a localizable, non-ignorable design, and hence the control strategy would need to be included in the likelihood when calculating the effect estimates. Currently, only one paper has applied a semi-bidirectional case-crossover design in a clinical setting.²⁴⁴

2.6.5 Overlap bias

A further form of bias present within certain case-crossover designs is ‘overlap bias’.²²³ This is not related to overlapping control windows, but is caused when exposure data is not considered to be exchangeable across the exposure series. Whether the control sampling in a case-crossover design leads to non-exchangeability can be linked to whether or not the control sampling is considered to be in the form of a case-control sampling design or a cohort sampling design (Chapter 2.2).²⁴⁵ Overlap bias is present if controls are set to be a fixed distance from the case such as in unidirectional or symmetric bi-directional controls no matter the size of the window.²³⁵ This can be shown in the score function associated with the sampling design. If the design is unbiased then the score function will equal zero. If the function of exposures is dependent on the individual, as occurs when the control sampling follows the case-control sampling design, then the score function does not equal zero.²⁴⁵ In short, the changes in control windows invalidate an assumption of independent sampling and cause bias due to the likelihood, conditional on the control window, not providing additional information regarding the effect size, hence it is also non-exchangeable²²³.

If cohort sampling design is used, as in full stratum bi-directional and time-stratified control designs, then the function of exposures is dependent of the time period chosen and not the subject. Hence, the score function will equal zero and results are produced that are asymptotically unbiased as the sample size tends towards zero.²⁴⁵ These designs also provide information related to the location within the time-series with regards to the seasonal and time trends²³¹.

2.6.6 Time stratified control selection

The time-stratified control design, introduced in 2000 by Lumley & Levy splits the study period up into distinct, separate cross-sections of most commonly months.²²³ The case day determines the cross-section, with free days within the cross-section either specifically assigned or randomly sampled to be controls. One example is to define the controls days to be, the same day of the week, for every week of the same month.²⁴⁶ For example, if the case day is the second Wednesday of the month then the control days are the 1st, 3rd, 4th and potentially 5th Wednesday of the same month (see Figure 2.6 as an example). A small cross-section will mean fewer suitable controls and less information, a large cross-section will increase the number of controls, yet at the same time increase the potential for confounding.²²³ In a suitably sized cross-section, days within the cross-section can be considered as matched on important confounders such as long-term time trends, season and day of the week are controlled for by the design.²³¹ As the position of case relative to control days varies depending on the case day, the design adds information about the effect estimate and so the design does not need to be included in the analysis; hence the design is localizable and ignorable, and a conditional logistic regression would give unbiased results.²³¹

The full stratum bidirectional design can be considered to be a special case of the time-stratified design with only one cross-section encompassing the entire time-series. though time-dependent factors are not accounted for and would need to be included in the model. The majority of studies have stratified by month and chose controls to be the same day of the week for all weeks within the month.^{148,247,248} Though all days in the same month not including the days adjacent to the case day²⁴⁹ or all days spaced two days apart have previously been employed.⁵⁶ Cardiovascular disease studies have controlled for the same hour,²⁵⁰ or same days in the month with the equivalent temperature.²⁵¹

Figure 2.6 - Example month of a subject identifying their case day (X) on their ‘day of death’ and control days (O) under the time-stratified case-crossover design.

Our Case and Controls

September

SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
					1 ○	2
3	4	5	6	7	8 ○	9
10	11	12	13	14	15 ○	16
17	18	19	20	21	22 ✗	23
24	25	26	27	28	29 ○	30

2.7 Choosing the time-series model or the case-crossover design?

The case-crossover design is an attractive method to gauge the effect of air pollution with the increase in its usage; comparisons have been made with results from the Poisson regression models. Time-series models adjust for confounding via inclusion of extra modelling terms that try to optimise the model fit. This can make them complex and highly data specific. The case-crossover method can control by design confounding by season, long-term time trends, autocorrelation, and day of the week changes. This reduces the required number of terms in the regression model to just the pollution and meteorological terms resulting in a simpler model. Comparisons of the number of deaths on each day can make it difficult to perform time-series analysis in small populations, whereas case-crossover studies use data on individual subjects allowing the use of smaller sample sizes. The inclusion of interaction variables to represent effect modification by subject factors is possible in case-crossover designs. Case-crossover designs are associated with having a lower statistical power than the time-series

models,²²¹ though it should be noted that an accurate power calculation is not currently available (A crude attempt will be described in Chapter 3.3.5)

It has been shown that the results are equivalent when the choice of control strategy of a case-crossover design can be matched by the method chosen to control for time varying factors in a Poisson regression model.²²³ The conditional logistic regression applied in most case-crossover studies assumes that all subjects are independent. In air pollution studies when subjects share a common pollution exposure, the conditional logistic regression is equivalent to assuming the count of deaths on any day within the study follows a Poisson distribution.²⁵²

Navidi originally showed that effect estimates produced in the bi-directional time-stratified design are equivalent if time-series Poisson regression model includes indicator variables identifying certain confounders, e.g. month and day of the week.²⁵³ Fung et al (2003). performed simulations to compare the time-series regression techniques with the case-crossover design.²⁵⁴ The results of the bidirectional case-crossover analysis gave effect estimate closest to the true estimate, however the standard error was larger than the time-series results. The larger standard error may be an indicator of the lower power associated with the bidirectional case-crossover design due to time-dependent confounders being ignored in the design.²⁵⁴ Lu and Zeger (2007), confirmed that a case-crossover design can be equivalent to a time-series model.²⁵² Alongside the symmetrical bi-directional control design and a time-series model with time as LOESS smooth, a time-stratified control design to a time-series model containing indicators of the time-strata.²⁵² These equivalences could then be used to perform model-checking with previously applied log-linear model methods; such as predicted responses, residuals, and influence statistics (e.g. Cooks distance, Dffits or Dfbeta) to determine outliers, autocorrelation, cyclical effects or show the degree of over-dispersion in the data.²⁵⁵ The example given in Lu et al. (2008) confirms the same effect estimates between designs; however a greater standard error occurred in the time-series model due to accounting for over dispersion a factor that conditional logistic regression does not account for.²⁵⁵

2.8 Modelling exposure 1: The exposure-response relationship

A log-linear dose-response or exposure-response models have commonly been fitted in air pollution studies. This is where, as the concentration level of air pollution increases by one unit the logged number of daily deaths increases at a constant rate. If this is not true, the rate of change is different depending on the concentration and a non-linear relationship exists. Effect estimates based on a linear relationship would be biased as in this case they will represent the average of multiple slopes.²⁵⁶ Any change in relationship is an important factor particularly if it's a sudden increase, and should be taken into consideration when making public policy regarding safe air pollution levels. Sudden increase risk at the higher concentrations would mean public policy should concentrate on reducing the number and level of high pollution days, whereas greater increases risk in at low level changes would me the average daily pollution level should be reduced.²⁵⁷

Some early 'threshold' modelling assumed no effect existed until concentration reached a threshold point at which point a linear relationship was applied.²⁵⁶⁻²⁵⁸ Schwartz (2000), initially assumed a $50\mu\text{g}\text{m}^{-3}$ threshold point but observed results that indicated either the threshold level was significantly lower than $50\mu\text{g}\text{m}^{-3}$ or the gradient of the slope was greater at lower concentrations than at higher concentrations. Most such as Daniels et al (2000), fitted a zero effect below a threshold level and a linear relationship above.²⁵⁸ This can be fitted in the any model but if assuming the Poisson regression model;

$$\text{Log}(E(Y_i)) = X_i\beta + \text{confounders}$$

where "confounders" represents adjustment for confounders in a GAM or GLM the pollution variable X_i is replaced with:

$$\text{Log}(E(Y_i)) = (X_i - h)^+ \beta + \text{confounders}$$

such that h is the proposed threshold point and;

$$\begin{aligned}(X_i - h)^+ &= X_i - h \text{ if } X_i \geq h \\ &= 0 \quad \text{if } X_i < h\end{aligned}$$

where the effect coefficient β no longer represents a unit increase in pollution starting at zero, instead from the threshold level h . To determine h , the optimum model defined by maximising the likelihood was chosen from differing h e.g. 5,10,15 $\mu\text{g}\text{m}^{-3}$, etc. A final comparison with a linear (no threshold) relationship is often performed.

Kim et al. (2004) followed a similar procedure except pollution below the threshold was not assumed to have an effect equal to zero.²⁵⁹ The model, also called a B-mode spline model was then defined as:

$$\text{Log}(E(Y_i)) = \beta_0 + X_i\beta_1 + (X_i - h)^+ \beta_2 + \text{confounders}$$

where, $(X_i - h)^+$ is the term representing the threshold representing two linear lines forced to meet at the threshold point h . As before, the threshold h is found through an incremental search that maximises the likelihood and the AIC values are compared with the linear model.

Alternative methods have been employed to determine if a non-linear dose-response effect exists. Daniels (2000) and Kim et al. (2004) both fitted alongside the threshold models a smooth function of air pollution in the form of natural cubic splines (see Chapter 2.5.3). Allowing for more flexibility in the data, the knots were set at the 25th and 75th percentage air pollution concentrations in both studies. Unrestricted spline models have also been used. Unlike the natural cubic spline the cubic functions are not constrained to connect or be linear in the extreme zones.

Schwartz and Zanobetti (2000), proposed a meta-smoothing technique,²⁵⁷ involving non-parametric smoothing for each dataset separately and then predicting the log relative risk of the number of daily deaths increments of the pollution. A meta-analysis then combined each increment using an inverse variance weighting. Shown to work adequately in simulated data, it showed no evidence for a threshold in the real population data here and in other papers.²⁶⁰⁻²⁶²

Others have also used threshold models, natural cubic splines, unrestricted cubic splines, and dose-response polynomial curves, where in 'same day' or 'day before' exposures a logged response-linear relationship has generally been observed.²⁶²⁻²⁶⁴

2.9 Modelling exposure 2: The lagged relationship

The majority of studies modelling air pollution exposure have primarily been concerned with modelling the effects over a short period of time following exposure. Regardless of the analysis model, early studies looked at the effect of same day changes in ambient pollution levels on same day changes in all cause total mortality^{10,122,193} or cause specific mortality,^{202,265} which have been confirmed in multi-city studies.^{204,266} Later studies extended the exposure periods to first determine the effects of the previous days average pollution level,^{195,264} moving to 3 day and 5 day means.²⁶⁷

2.9.1 Defining the ‘lag’ period

Schwartz (2000) proposed that the effect of an increase in exposure at time point zero may influence the population over an extend period of time before returning to baseline risk. If the effect of an increase in exposure lasted for only a few hours then a model fitting same day exposure would be appropriate. If the effect lasted longer than 24 hours, a simple same day mean would not be sufficient to measure the effect of exposure.¹² This delay between exposure and change in risk was designated as the ‘lag’ period or ‘lagged’ effect has been adopted for most short-term air pollution studies and this thesis.

However, this definition of a ‘lag’ or ‘lagged effect’ is in slight contrast to the definition used in other areas of epidemiology. For example, if the event of interest may take many years to produce symptoms (e.g. cancer) then in a cohort study, outcomes that have only occurred after 10 years after first exposure may be chosen.¹²¹ Alternatively, in a case-control design the exposures in the previous 10 years would be ignored.²⁶⁸ In the two designs, ignoring the 10 years would be considered as ‘lagging the effect’. In other words there would be a 10 year lag period followed by a period of observing the effect called ‘effect duration’. The corresponding general epidemiological definition of ‘lag’ would, in Schwartz definition, correspond to be a zero day lagged period followed by three days of effect duration. Instead here, the entire effect duration is defined as the lag period.

The implementation of ‘lag’ periods used in air pollution studies has ranged from single day lags e.g. ‘same day’ (lag 0), previous day (lag 1), or two days prior (lag 2), to combined lag periods of 3 days (lag 0-2), 5 days (lag 1-5) or even longer lags of 40 days (lag 1-40). The choice of lag length has largely depended on the study objectives, the quality of the data available, and any proposed underlying biological mechanisms.

The lagged effect can equally be thought of in reverse, where the event of interest at a time point has been the result of exposure during preceding days, hence we model the exposure on the preceding days (see Figure 3.2 as an example) to represent the lagged effect.²⁶⁹ Modelling exposure on adjacent days leading up to the event has a number of statistical characteristics that need to be accounted for. Several methods have been proposed including constrained distributed lag and the lag stratified models. Firstly a clear understanding of the unconstrained distributed lag model is required.

2.9.2 Unconstrained distributed Lag

A distributed lag model assumes a change in air pollution on day 0 has an effect on the event of interest (e.g. mortality) also on the subsequent days 1, 2, 3, etc. Again using the Poisson regression model an unconstrained distributed lag model can be written as:

$$\log(E(Y_t)) = \text{confounders} + \beta_0 Z_t + \beta_1 Z_{t-1} + \dots + \beta_q Z_{t-q}$$

where, Z_{t-q} is the exposure level on q days prior to the event day t with a lag length q , with the effect experienced on day t is the total effect of all the days prior i.e. q to t . The total effect of a uniform increase of one unit across day’s q to t can be written as:

$$\beta_0 + \beta_1 + \dots + \beta_q = \sum_{q=0}^q \beta_q = \beta^*$$

The equation can be revised to be in the form:

$$\log(E(Y_t)) = \text{confounders} + \beta^* (w_0 Z_t + w_1 Z_{t-1} + \dots + w_q Z_{t-q})$$

where $w_i Z_{t-i}$ is the weighted average pollution level on the i^{th} ($i=0,1,2,\dots,q$) days prior to death; note w_i are weights that sum to one. The effect of an increase in pollution on any individual day within the lag period will be experienced on same day and q days that follow.

Schwartz (2000) and Braga et al. 2001, both used the Unconstrained Distributed lag (0-5 days) model with increases in PM_{10} , for all-cause mortality¹² and cause specific mortality.¹⁵⁰ Air pollution data contains autocorrelation on adjacent days; therefore effect sizes produced would contain high collinearity. However, the sum of β_q 's (or β^*) is an unbiased estimate of the overall effect.¹⁵⁰ The unconstrained lag distributed model has tended to be applied to a short lag periods of less than 5 days.^{12,80,150,270} Using an overall effect size and a short lag period mean that it is difficult to assess the shape and length of the lag effect.

2.9.3 Lag stratified models

A model that contains only the same day exposure (lag 0) can also be considered a special case of the constrained model, where the β 's are constrained across the lag such that $\beta_1=\beta_2=\dots \beta_{q-1}=\beta_q=0$ if in the unconstrained distributed lag model.¹² The intermediary stage between this and the full unconstrained lag model is a lag stratified model, where the full lag period is stratified into shorter periods of, for example 6 days each. The coefficients within the lag period are constrained to be equal such as for lag 1-6:

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_{1-6}$$

which is equivalent to replacing the individual pollution concentration for each day (unconstrained lag distribution) with the mean pollution concentration for the stratified lag periods. The coefficient represents a common increase in exposure across all six days opposed to an increase on an individual day. Therefore coefficients related to increase in the average represent six times the coefficients in an unconstrained lag i.e. a constant increase on each of the 6 days.²⁷¹ The lag stratified model is useful for assessing longer lag periods and has been employed in three papers investigating black smoke air pollution and temperature on mortality over 30 days.^{13,272,273}

Results from the lag-stratified model are easier to interpret than some of the more complex models that have been described earlier, particularly useful when comparing data from multiple cities. However, the lag stratified model is not flexible enough to adequately represent more complex lag distribution shapes, particularly within a stratum.²⁷¹ More flexible methods would be required to see these short-term changes.

2.9.4 Constrained distributed lag

The Schwartz (2000) paper, alongside the unconstrained lag distribution model, proposed a constrained lag distribution model using a polynomial function.¹² The approach, originally described in Gaussian data by Almon in 1965, constrains the shape of the effect estimates β_q 's for the days of the lag period to follow a polynomial function:²⁷⁴

$$\beta_j = \sum_{k=0}^d \eta_k j^k$$

where j is the lag day $0, \dots, q$ and d is the degree of the polynomial distributed lag model. Using the Poisson regression model from Chapter 2.9.2, the polynomial constraints can be written as:

$$\begin{aligned} \log(E(Y)) = & \text{confounders} + \eta_0 Z_0 \\ & + (\eta_0 + \eta_1 + \eta_2 + \dots + \eta_d) Z_1 \\ & + (\eta_0 + 2\eta_1 + 4\eta_2 + \dots + 2^d \eta_d) Z_2 \\ & + \dots + (\eta_0 + q\eta_1 + q^2\eta_2 + \dots + q^d \eta_d) Z_q \end{aligned}$$

which can be fit in a model where the covariates are the collective weighted sums of the of the exposure variables Z as shown, with the estimates η 's being the parameters of the model.

$$\begin{aligned} \log(E(Y)) = & \text{confounders} + \eta_0 (Z_0 + Z_1 + Z_2 + \dots + Z_q) \\ & + \eta_1 (Z_1 + 2Z_2 + 3Z_3 + \dots + qZ_q) \\ & + \eta_2 (Z_1 + 4Z_2 + 9Z_3 + \dots + q^2 Z_q) \\ & + \dots + \eta_d (Z_0 + 2^d Z_1 + 3^d Z_2 + \dots + q^d Z_q) \end{aligned}$$

The Schwartz paper selected a second degree (quadratic) polynomial to investigate a 0-5 day exposure period prior to mortality.¹² The quadratic polynomial is a popular choice for short lag periods.^{151,275}

Using a quadratic function means the model will fit a quadratic curve to the data even if a more suitable function is required smoothing out important fluctuations. This is unlikely to be a problem when the lag period is short but longer lags may increase the chance of bias. Increasing the degrees of freedom in the polynomial function will result in a more flexible shape should one exist. Over greater lag lengths between 8 and 40 days cubic (3rd degree) polynomial functions^{276,277} and 4th degree polynomial^{152,153,278} have been fitted. There are no restrictions, but the number of degrees of freedom

controls the flexibility with a low number potentially smoothing out underlying true relationships or a greater number influenced by too much noise. Even so, some studies have reported a 4th degree, 6th degree, or an 8th degree polynomial function for both pollution and temperature.⁷⁹

2.9.5 Penalised spline

Unconstrained lag models are subject to collinearity making it difficult to determine the true shape of the lagged effect, and constrained models can lack flexibility to account for smaller localized structures. A penalized spline smoothing has been proposed²⁷⁹ that extends the constrained polynomial function, described in Chapter 2.9.4, to a regression spline such that β_j becomes a piecewise polynomial with a dth degree that was described in Chapter 2.5.3. As it is, the estimated effect sizes β_j lead to a function that is too flexible and so a penalty term is added to the least squares criterion. This method, called penalised spline smoothing, has been applied to mortality data from Milan.²⁷⁹ However, this method has primarily been used when investigating the harvesting hypothesis and will be discussed further in Chapter 2.11.

2.10 Modelling exposure 3: Exposure-response and lagged relationship

2.10.1 Distributed lag non-linear models

The effect of a non-linear dose-response and the lag distribution have generally been investigated separately, assuming in the lagged-response models that exposure was exhibiting a linear relationship with mortality. The assumption of a linear relationship has been widely observed at very short term lags (lag 0 or 1),^{257,258,260,262,263,280,281} however this may not be the case over longer lag periods. Armstrong (2006), proposed the use of ‘cross-basis’ functions to fit a non-linear dose-response relationship across a distributed lag function. The case he outlined was investigating the relationship between temperature and mortality across a lag period of 30 days.²⁷¹

The ‘basis’ function is a designated term for any transformation function of the original variable x that generates a new set of variables. Here x can represent either the daily exposure range or the lag period. These functions apply a non-linear relationship directly calculated from the original variable, in the form of a parametric smoothing curve such as polynomials, natural cubic splines, indicator variables, or a segmented linear spline. The ‘basis’ functions are chosen independently for exposure and lag such that they relate to the potential shape, level of accuracy, and their ability to interpret the results. The relationship exposure-response and lagged-response represented by the sum of a set of ‘basis’ functions are then ‘cross multiplied’ (for the exposure and lag) to create a set of linear terms that can be included in the model. These cross multiples are then the designated ‘cross basis’ functions and are defined as:

$$x_{b,p,i} = \sum_{l=0}^L l_p x_{b,i-l}$$

where, l_p and x_b represent the basis function of the lag distribution and exposure variable at each lag time point $i - l$, respectively.

In the example given by Armstrong (2006) on a London data set, and in a further paper by Gasparri et al. (2009) based on the NMMAPS dataset, natural cubic splines were adopted for both temperature exposure and lag across 28 and 30 days, respectively.^{271,282} One limitation is the difficulty in interpreting the effect estimates produced by more complicated basis functions such as these. Armstrong suggested simpler methods, such as the lag-stratified approach, may be preferred particularly if multi-study sites are to be compared.²⁷¹ Alternatively, interpretation of the more complicated terms can be achieved by plotting the change in risk across the exposure and lag simultaneously in 3D surface plots or in cross-sectional slices e.g. change in exposure when lag is fixed at 0, 5, 15, etc days, or across the lag period when exposure is fixed at -10, -5, 0, 5, 10°C, etc.²⁸²

This technique was initially applied to temperature – mortality studies. These have shown the effect of temperature on mortality to be non-linear in shape with a U, V, or reverse J shape. Air pollution studies on the other hand have tended to assume a linear relationship in the immediate time period based on the short-term lag period (0 or 1 day). It may be useful to investigate if a nonlinear relationship persists when lag periods are extended greater than 5 days.

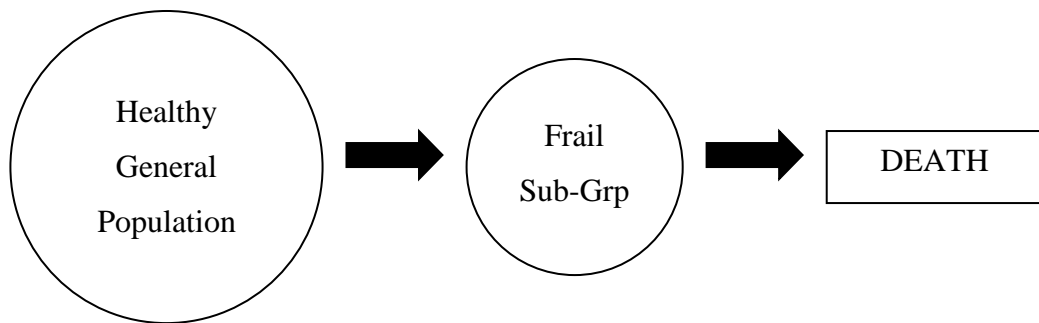
2.11 Harvesting/mortality displacement

Harvesting also known as mortality displacement is an important concept when interpreting delayed exposure effects. The link between short-term changes in air pollution and mortality, or morbidity, has been shown in locations across the world in single and multiple location studies for whole populations. Debate still remains as to how changes in pollution affect individuals in a population; are changes in pollution affecting only subjects who are already frail? In an extreme case these might be subjects who would not be expected to live much longer regardless of the change in pollution level. Schimmel & Murawsky first described in 1976 the hypothesis that only already frail subjects were being affected by exposure to air pollution. They referred to the later named “harvesting” or mortality displacement phenomenon as excess premature deaths.¹⁹⁶ If the “harvesting” hypothesis is true and only frail subjects are affected then

perhaps only a small amount of expected life in days would be lost. Of course if the converse is true and subjects who otherwise would be healthy and alive for a long period of time are dying much earlier due to exposure, then this would have different implications on public policy.

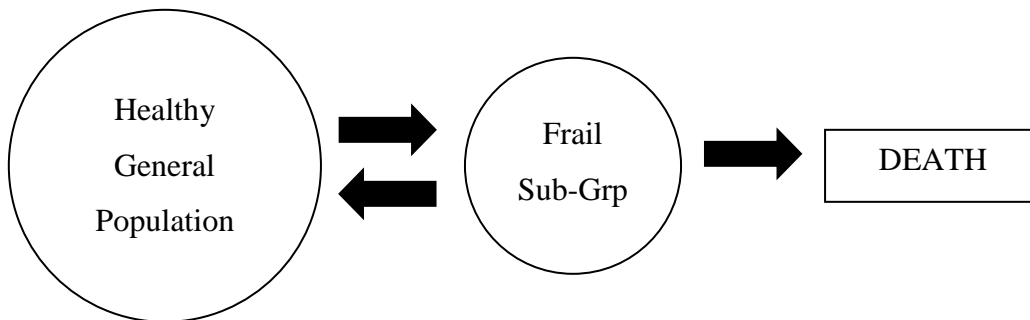
In a very simple model (Figure 2.7) outlined by Zeger et al. (1999) a healthy subject experiences two phases; moving first from the healthy general population into a smaller frail sub-population whose death rate is high, regardless of pollution; then after a period of time mortality will occur.²⁸³

Figure 2.7 - A simple 'harvesting' model showing continuous transition of subjects from health general population to frail sub-group leading up to death



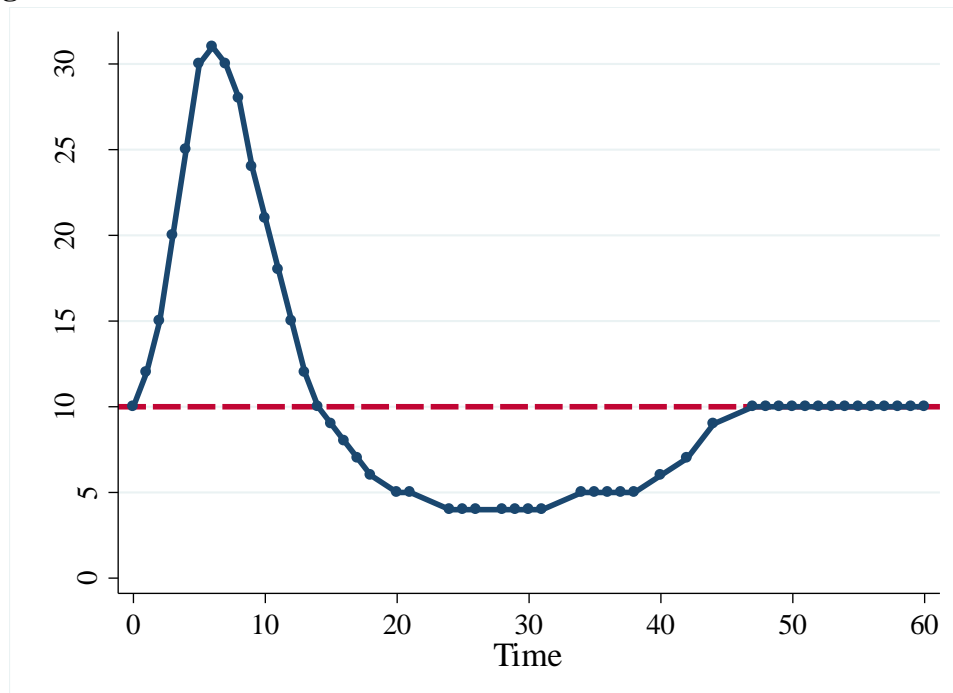
A subject located in the frail sub-group would be expected to have a short time to death compared to the healthy population, regardless of pollution exposure, in which the likelihood of death increases as air pollution levels increase. The size of the frail sub-group is then an indicator of the length of time the subject is present. Subjects are likely to exit the frail group either via death or through recovery back to the health general population (Figure 2.8). Under no external exposures, over time the rate at which the frail group is replenished would be in equilibrium and the size of the frail group would be constant.

Figure 2.8 - A 'harvesting' model showing transition of subjects from health general population to frail sub-group (with the possibility of return) leading to death



A small frail group would indicate that subjects are consistently dying quicker than the frail group can be replenished i.e. due to exposure to air pollution. Whereas a large frail group would indicate that health in the general population is decreasing faster than the death rate or recovery rate i.e. deaths from exposure to air pollution is only affecting those already with a short expected life span.²⁸³ In this case, we consider that death only occurs in a subject once entered into the frail group, and air pollution is causing deaths in subjects only a few days earlier than would be expected. As illustrated in Figure 2.9, under change in air pollution the rate of death would initially increase quickly to a peak. A subsequent drop in the membership of the frail group would cause mortality rates to drop below the expected mortality rate for a period of time while the frail groups numbers are replenished at which point the mortality rate would begin to return gradually to an expected rate.²⁸³ This has been termed as the 'rebound' effect.⁸⁴

Figure 2.9 - The likely change in mortality rate from the Expected Mortality Rate (Red dashed Line) when in the presence of short-term harvesting conditions leading to the rebound effect.



This ‘rebound’ effect is unlikely to occur over a very short or a very long period of time, hence most techniques proposed to model harvesting have investigated the rebound effect within a few weeks. Several early methods were proposed such as the frequency domain log linear model,²⁸³ time domain log linear model,^{284,285} the STL algorithm (Seasonal decomposition Time-series LOESS),²⁸⁶ and penalized spline smoothing²⁷⁹ but the most commonly used method has been the distributed lag models.

2.11.1 Distributed lag methods in harvesting

The introduction of distributed lag models (see Chapter 2.9.4) lead to a more straightforward method of assessing mortality displacement, as it will provide an estimate of the cumulative effect across the lag period.¹² The use of a distributed lag model to assess mortality displacement has been used in papers in relation to particulate matter^{152,278,279} and Ozone levels.²⁸⁷ As discussed, under short-term harvesting a rebound effect should present a period of lower than expected death rate exists (see

Figure 2.9). The cumulative effect estimate should tend towards zero compared to the base estimate. Zanobetti et al. separated the lagged effect of TSP into three sections; the initial increase, the drop below zero, and the return to a positive effect on daily mortality. They found a drop in the estimated effect at section 2 that remained greater than zero before increasing again. Hence the cumulative estimated effect increased rather than decreased towards zero. Similar results occurred, even at longer lag periods, when unconstrained, cubic and 4th degree polynomial distributed lag models were applied in all-cause,²⁷⁸ and respiratory and cardiovascular mortality.¹⁵² Indicating that mortality displacement potentially occurs for significantly longer than the short-term harvesting hypothesis would suggest. The unconstrained and penalised distributed lag models were applied for ozone levels 21 day lag in two studies.^{149,287} In both papers the cumulative effects across the 21 days tended to be at least as large as the lag 0 effects. This may have been due to the length of the lag and a longer one may show a tendency towards zero and harvesting.

Roberts (2004) performed simulations under varying harvesting conditions to determine the ability of the distributed lag model to measure the harvesting effect.²⁸⁸ Mortality time-series data were generated under varying conditions of the harvesting model with simulations split into three groups regarding the frail group. 1) An enter only model where pollution exit effects were set to zero, 2) An exit only model where pollution effects regarding entry to the frail group are set to zero, and 3) an enter & exit model where both conditions are varied. The total effect estimates from a distributed lag model indicated that if the mean time in the frail group is shorter than the length of lag used in the model, then the effect estimates adequately show the occurrence of harvesting. This may indicate that the usual lag of 40 days used in assessing mortality displacement may not be long enough to account for the mean time in the frail group and determine if harvesting exists.

2.12 Missing data

2.12.1 Introduction

Data collection of non-simulated real time data, no matter how meticulously performed, is susceptible to the presence of missing data. This is common in epidemiological studies where the potential for missing data is high and it can impact on the quality and validity of the results. Even when measurements are meticulously recorded there is potential for sporadic missing data within a dataset. Many default analysis techniques use a 'complete cases' approach, excluding all participants with even one missing observation. This may substantially reduce the sample size, in turn causing a loss of power; but it can also increase the chance of bias if systematic differences are present between observed values and missing values.^{289,290} When possible, it is important to consider employing an adequate method of dealing with missing data, especially when ignoring missing data would result a substantial loss in the sample size.

2.12.2 Exposure data characteristics & missing data implications

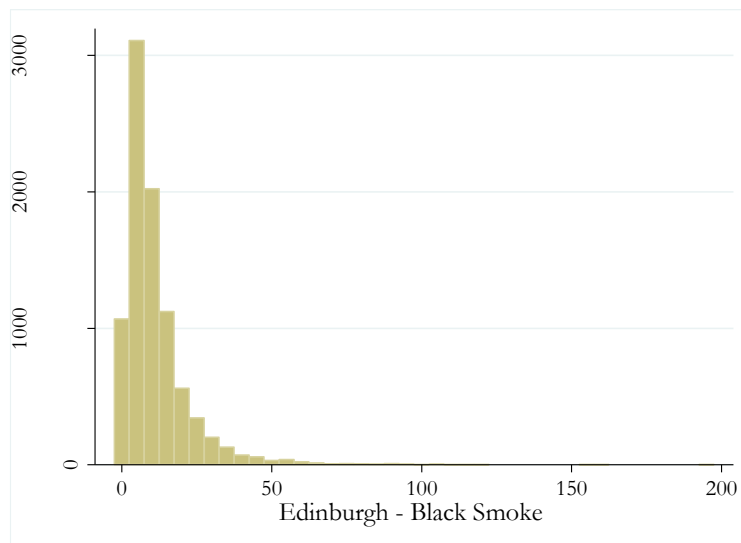
Epidemiological studies investigating air pollution and meteorological data rely on measurements made by government run, fixed, monitoring sites across cities.^{177,291} The alternative, personal monitors, are expensive requiring many years to produce a large enough sample. Despite significant increases in the sample size for fixed monitor studies there are still limitations. One being that the presence of missing data, may occur due to several reasons including human error, changes in site location, mechanical or scheduled breakdown of automatic equipment. Some sites may be in hard to reach places and so if a breakdown occurs it may take time for a mechanic to be sourced and reach the site, meaning multiple days in a row may be missing.

In study designs such as the time-series and case-crossover designs, one missing exposure day may correspond to the loss of a number of participants in a complete cases analysis. Missing data can cause bias if systematic differences exist between the observed data and the missing data. For example, if missing pollution data is more

likely when temperature is higher and temperature influences the daily level of pollution then missing data will potentially underestimate the influence of pollution and temperature on mortality. Accounting for missing data can require a large amount of additional modelling complexity and so it may not be worth recovering if small amounts of information are missing.

The data structure and characteristics are important when attempting to account for missing data. Unlike temperature which follows a parametric normal distribution, pollution exposure data tends to be in a non-negative, skewed continuous form and is often considered to be on a log-normal scale (e.g. Figure 2.10). This may mean standard parametric assumptions of normality are not appropriate and a log transformation or equivalent method will be required to suitably fit missing data replacement techniques.

Figure 2.10 - Histogram of black smoke air pollution measurements from Edinburgh 1979-2009



2.12.3 Missing data processes & definitions

Missing data, located either in the predictor or outcome of interest, is defined as data in which a value could have been recorded but for some reason was not. The question is then, given missing data will the analysis give valid effect estimates and standard errors compared to results from a complete dataset? To answer this, missing data is classified into three types; Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR).²⁹²⁻²⁹⁴

To help define these concepts, the complete set of explanatory/outcome data is X ; this is comprised of either observed X_O or unobserved X_M , i.e. missing data. The probability that a value or set of values are missing given the observed and missing observations can be written as: -

$$Pr(M|x_O, x_M)$$

where M is a missing data indicator set to be $M = 1$ if X is observed and $M = 0$ if X is missing.

Missing completely at random (MCAR) is when the probability of an occurrence of a missing value is not dependent on the value of any observed or missing value. For example, if a daily estimate of pollution is missing due to the measurement device breaking down, assuming the breakdown is not related to any other factor but through random chance alone. The notation therefore becomes:-

$$Pr(M|x_O, x_M) = Pr(M)$$

If missing data is MCAR then a ‘complete cases’ analysis should give unbiased results, although it will be subject to reduced power given the loss of data.

Suppose that temperature is being measured. If the pollution monitor is more likely to breakdown - causing missing data - when temperature becomes colder this would then be deemed *Missing at Random (MAR)*. Here the probability of a missing value is dependent on an observed value.

$$Pr(M|x_O, x_M) = Pr(M|x_O)$$

If the probability that the value is missing is related to an unobserved value including the value of the measurement itself then it is deemed *Missing not at Random (MNAR)* i.e.

$$Pr(M|x_O, x_M) \neq Pr(M|x_O)$$

Missing not at random in our pollution monitor example may occur in the pollution measurements if the measuring device broke down due to a very windy day, where wind speed was not measured, or due to a day with an extreme amount of pollution causing the device to overload and breakdown.

The chances of the missing value being MAR can be increased by simply collecting more information in the form of explanatory variables on the subject.²⁹⁵ For example, the MNAR example would become MAR if daily wind speed were measured. To account for missing values within the data, an underlying model which adequately represents both the observed data and the missing data is required in order to attain MAR. However, it can be difficult in real measured data to determine which of the three forms of missing data are present, particularly with MAR as unobserved values that predict missing values may still be present.²⁹⁶

2.12.4 Missing data characteristics and patterns in time-dependent data

A clear understanding of the data characteristics and the potential corresponding missing data patterns is important. In order to firstly, judge if missing data is likely to be MCAR, MAR, or MNAR given the available covariates and secondly, that those covariates adequately predict the missing value i.e. they produce suitable replacement observations. In traditional study designs each observation, whether present or missing, is related to a subject characteristic, for example demographic or clinical data at a particular time point. This means that to improve the likelihood that the data is MAR further subject specific data could be collected.

Conversely, observations in a time-series study design represent repeated values over time on variables that are thought to be related to the subject outcome. Therefore variables that explain the missing data pattern or the missing data characteristics need to be representative of the time dependent factors. This means accounting for time descriptive variables that represent systematic patterns, for example, between days of the week, calendar months, season, year, calendar day, or study day. These time-descriptive variables mostly provide information regarding the long-term time trends in the data. The variable containing the missing observations can itself provide information regarding short-term trends and thus provide information to help predict missing data. Time-series data containing autocorrelation means that variables representative of the data at time points prior to or following the day of interest (time window) are important explanatory variables. The width of the ‘time window’ around the day of interest may be dependent on the strength of autocorrelation and the length of short-term patterns in the data.

Additionally, supplementary time-series variables that correlate with the primary variable may provide important information. For example, it may be the case that pollution may increase in the summer months when temperatures are high or when wind speed is low. Daily temperature or wind speed data for the same day or across a similar ‘time window’ would be important considerations when investigating missing data in a pollution variable. This relies on the additional time-series variables being collected concurrently ideally with as little missing data as possible.

Once appropriate missing data assumptions (i.e. MCAR or MAR) are judged to be present and data the appropriate characteristics identified, so long as the appropriate measurements are obtained, a method of analysing data with missing observations is applicable.

2.12.5 Modelling with missing data present - Complete Cases

The standard approach to dealing with missing data and the default process in most statistical packages is to ignore the missing data completely and run a ‘complete cases’ analysis. Here, the analysis only includes those subjects with complete data i.e. no missing values in any variable included in the model. For example, in a regression model comparing blood pressure (dependent variable) with cholesterol adjusted for age, any subject with a missing value in any parameter (blood pressure, cholesterol, or age) would be excluded from the analysis. This means if age is missing then we can lose important information on blood pressure and its relationship with cholesterol.

If a ‘complete cases’ analysis is employed and the missing values present in the data cannot be considered MCAR, or MAR given the covariates in the model and a maximum likelihood estimation, then analysis estimates (e.g. means, ratios, regression coefficients) and corresponding standard errors are likely to be biased. Resulting in incorrect conclusions, p-values and confidence intervals.²⁹⁷

A ‘complete cases’ analysis can also be considered inefficient, even if MAR is present in maximum likelihood estimation. To reduce bias in estimates, the model often includes covariates thought to account for potential confounding. If missing data is present in any covariate, then the sample size used in the ‘complete cases’ analysis will be reduced; significantly, if many covariates are included. The reduced sample size causes the standard errors to become larger, affecting reducing the power to correctly determine a significant result.²⁹⁷ It is therefore advantageous to somehow try to account for missing data in the explanatory variable of interest and any covariates, particularly if a large amount of data is missing when extra covariates are included.

2.12.6 Modelling with missing data present - Imputation

The central idea of imputation is that the missing data are filled in with a value considered to be a credible replacement of the true unobserved value. These imputed values, together with the observed data, then allow the analysis to continue without a significant reduction in the sample size. The following describes methods for generating a valid imputation starting with the simple, single imputation methods and building to more complex multiple imputation techniques. Note, it is considered of little value to account for missing data in the dependent variable, as it is likely to add little additional information particularly if the missing dependent variable is considered MAR given the covariates in the model.²⁹⁸ Hence the following will focus on techniques applied to the independent variables.

2.12.7 Modelling with missing data present - Mean imputation

The simplest imputation method is to replace any missing value with the arithmetic mean of the observed data for the variable. Mean replacement is not ideal as the distribution of the data becomes distorted. The addition of centrally located imputed values results in the mean of the variable remaining unchanged so long as MAR is present. However, with the increased sample size the standard errors will be reduced because the variance of the new dataset will be less than the true dataset.²⁹⁹ Smaller standard errors will result in any between group comparisons being exaggerated and a statistically significant result can be found where none exists, hence a greater chance of a false positive result i.e. type 1 error.

2.12.8 Modelling with missing data present - regression model

An 'imputation model' in the form of a standard regression model is fitted to the observed data with the variable containing the missing data set as the dependent variable; this model is then used to predict the value of the missing observations. Using appropriate regression models, predictions can be made for continuous, binary, and categorical variables. The predictions from the imputation model are likely an improvement on the mean imputation as the predicted values will probably be closer to the true values; however they may also increase the chance of incorrect conclusions about the data. The regression model will only add a small amount variation hence as with mean imputation; the potential for type 1 error is still greatly increased.

2.12.9 Modelling with missing data present - regression model with random variation

To increase the variation we can modify the regression model described in Chapter 2.12.8. Once predictions are generated using the regression model in the observed data, a random error factor is added to the mean prediction based on a randomly allocated value from the normal distribution with a mean zero and variance equal to the residual variance of the original regression model. This method, called stochastic regression imputation, is considered an improvement from the standard regression model due to the added random variation. However, the imputations are still overly precise as they are based on the observed data and do not include the potential for additional uncertainty. Here the effect estimates – and variance - will not change, but the sample size will increase meaning the standard error will be smaller (though to a lesser degree than the mean and regression methods) resulting the incorrect p-values and confidence intervals.

2.12.10 Modelling with missing data present - Multiple imputation

In 1987, Rubin proposed a method to impute missing data that generates valid results with a reduced likelihood of producing biased effects and artificially replicating the true standard errors. Which will improve the validity of any conclusions based on hypothesis testing.³⁰⁰

Rubin proposed a three step process; Imputation, Analysis, and Combination:²⁹⁰.

1 – The imputation process is carried out not just once but multiple times. Using an *'imputation model'* multiple datasets are created to contain credible imputed values with random variation that replace the missing values. The number of imputed datasets will be discussed further in Chapter 2.12.15, though is typically between 3 and 10.

2 – Each complete dataset containing imputed values is analysed individually in a *'substantive model'* using the appropriate modelling technique in a standard 'complete cases' analysis techniques to provide an estimate of the parameter of interest.

3 – The estimates of the individual imputed datasets are combined using rules proposed by Rubin, where the combined variance of the effect estimate is calculated in a way which accounts for the uncertainty in the imputations.

2.12.10.1 Rubin's Rules

To combine the effect estimates Q_1, \dots, Q_m and variance estimates V_1, \dots, V_m produced by the m imputed datasets we follow a set of rules laid down by Donald Rubin in 1987.³⁰⁰

Rule 1: The combined estimate Q^* , and the mean within-imputation variance V^* are calculated by:

$$Q^* = \frac{1}{m} \sum_{i=1}^m Q_i \quad V^* = \frac{1}{m} \sum_{i=1}^m V_i$$

Rule 2: Calculating the between-imputation variance B means that the total variance T is defined as:

$$B = \frac{1}{m-1} \sum_{i=1}^m (Q_i - Q^*)^2$$

$$T = \text{Var}(Q^*) = V^* + \left(1 + \frac{1}{m}\right) * B = \frac{1}{m} \sum_{i=1}^m V_i + \left(1 + \frac{1}{m}\right) \frac{1}{m-1} \sum_{i=1}^m (Q_i - Q^*)^2$$

Confidence intervals and significance tests can then be calculated in the standard way using:

$$Q^* \pm (t_k * \sqrt{T}) \text{ and } \frac{(Q^* - Q_{null})}{\sqrt{T}} \sim t_k$$

where k equals:

$$k = (m-1) \left[1 + \left(\frac{V}{B}\right) \left(1 + \frac{1}{m}\right) \right]^2$$

Rubin's rules can be used to estimate the amount of information lost due the missing data. An estimate of the uncertainty in the model coefficients will allow for a comparison between 'complete cases' and multiple imputation analysis.³⁰¹ The Fraction of Missing Information (FMI) is calculated by dividing the amount of missing information by the amount of complete information as derived by the relative increase in the variance of an estimate due to missing data.³⁰⁰ The FMI is defined as:

$$FMI = \frac{\left(r + \frac{2}{k+3}\right)}{r+1}$$

where r equals:

$$r = \left(1 + \frac{1}{m}\right) * \frac{B}{V^*}$$

2.12.11 Multiple imputation methods – Univariate ‘imputation model’

The two techniques employed to impute missing values for a univariate variable containing missing data are described below:

1 – A *simple regression model* as described in Chapter 2.12.9 is fitted and used to simulate coefficients and variances at random from their Bayesian posterior distribution i.e. the distribution of the regression coefficients when the data is observed. The first set of imputed values are then randomly generated from a normal distribution with a mean set to be the mean of the predictions given the simulated regression coefficients and standard deviations. To increase the number imputations the process is repeated by generating a second set of randomly drawn regression parameters and imputing the missing data.³⁰⁰

2 – *Predictive mean matching (PMM)* begins by applying the regression method just described to predict a value for all the missing values and the observed values. For each missing value, a k sized set of observed values are identified whose predicted value was closest to the predicted value of the missing value. The missing value is then replaced with a randomly chosen observed value from the k set of observed values.³⁰²⁻³⁰⁴ Little (1986), originally used the nearest neighbour.²⁸⁹ If the k set of nearest neighbours is small certain imputed values may be repeatedly picked and the imputations would fail to reflect any uncertainty in the missing data, which in an extreme case would become the same as a single imputation. Though sets of 3-5 nearest neighbours have been proposed,³⁰² the current recommendation proposes to randomly choose from a set of at least 10 observed values. The PMM method is useful when replicating the observed data in the imputed values is important. This may occur when the observed data does not follow standard regression modelling assumptions such as in a highly skewed distribution, a fixed range, or a non-linear relationship none of which can be suitably replicated through a transformation. Recent improvements in computation and statistical packages have meant PMM is no longer considered to be too computationally intensive, even in large datasets with multiple variables containing missing data.^{305,306}

2.12.12 Multiple imputation methods - Multivariate ‘imputation model’

Studies tend to contain missing data in more than one variable. The following methods have been developed to account for missing values in multiple variables simultaneously.

2.12.13 Multiple imputations by chained equations (MICE)

Multiple Imputation by Chained Equations (MICE),³⁰⁷ is also known as Sequential Regression multiple imputation³⁰⁸ or Fully Conditional Specification.³⁰⁹ In MICE a separate imputation model is described for each variable containing missing values, with the multivariate missing data imputed by a series of univariate conditional models.

Initially, the missing values are filled with randomly chosen values from within the observed data for each variable. The randomly filled in data is then used in a univariate regression imputation, where the first variable with its missing values set back to missing, is regressed on a set predictors variables with the missing values filled in. A new set of imputed observations are predicted as described previously (Chapter 2.12.11). The process repeats for the second variable containing missing values regressed on to the same predictors this time including the first variable with its newly imputed values. Once all variables containing missing values have been imputed the cycle starts again. The cycle is repeated with newly imputed observations for a second, third, fourth, etc cycle until the imputations converge becoming stable and considered unrelated to the starting values; this is known as a burn in period. Once the burn in is complete the imputed values from the next cycle is saved as the first set of imputations. The whole process is then repeated to obtain the second, third, fourth, etc set of imputations.³¹⁰

One advantage of MICE is that it allows for differing data structures within the univariate modelling stage. This means that incomplete variables which are not normally distributed such as count, binary, or ordinal data, can be modelled using appropriate alternatives to the linear regression such as Poisson, negative binomial, logistic, and ordinal logistic regression models.

2.12.14 Multiple imputation - Joint modelling

One alternative to the MICE procedure is Joint Modelling (JM).³⁰⁹ In JM observations are grouped depending on their missing data patterns. Imputations are then generated within the groups according to a joint model of those variables that are common to all observations. This has been applied to continuous variables and categorical variables in the form of multivariate normal model and log linear respectively.³⁰¹ Joint modelling begins by identifying a parametric multivariate density distribution based on the data, given the model estimates. An appropriate prior distribution is defined for the model parameters leading to each missing data pattern having an appropriate sub-model derived. Imputations are then drawn from the sub-model for each missing data pattern.²⁹⁹

The main disadvantage of this method is that unlike MICE, it is not able to account for more complex data structures within the JM as all variables to be imputed need to be on the same continuous or log-linear scale. In reality varying data structures are likely to be present in most studies that record lots of data in on each subject.²⁹⁹ In most cases the solution is to perform a transformation on the variable that requires imputed values; however this means that the imputed observations will often need to be rounded. Rounding imputed variables to the nearest integer value will incur bias in the model estimates.³¹¹ Van Buuren in 2007 concluded that the MICE procedure is a more flexible and convenient method of dealing with missing data in complex data structures,³⁰⁹ and has been shown to produce unbiased estimates^{308,312}

2.12.15 Multiple imputation considerations

Several considerations need to be taken into account when applying multiple imputation to a dataset.

2.12.15.1 Number of imputations

The number of imputations is an important consideration. Multiple imputation can be computationally intensive and time consuming; increasing the number of imputations may take considerable time without adding useful information to the results. Schafer (1997) indicated a relatively small number of imputations would provide an efficient estimate of the standard error even if the FMI was large.³⁰¹ Schafer showed using Rubin's second rule that the ratio of relative efficiency of estimates compared to an infinite number imputations is equivalent to:

$$RE = 1 + \frac{FMI}{m}$$

If the FMI was 0.25, then 5 imputations would attain a relative efficiency loss of 5% compared to an infinite number of imputations. Schafer suggested that in the majority of cases no more than 10 imputations would be needed,³¹³ though even this may not be the case if the FMI is very large.

Variation in final results due to choosing a finite number of imputations is a concern, especially if the combined confidence intervals and p-values indicated a borderline result. If variation in the number of imputations changes the final conclusion any results would be undermined. Variation in the results or 'Monte Carlo error' can be estimated by repeating the full multiple imputation process several times with different starting seeds, and comparing the variance of the estimates as the between-imputation variance B divided by the number of imputations m . Alternatively, the 'Monte Carlo' error can

be calculated using a Jackknife procedure, in which, one observation is removed from the dataset, the multiple imputation procedure is implemented and the effect estimate and variance of combined results after multiple imputation are stored. Once this is repeated for all observations within the dataset the stored results are compared.³¹⁴

2.12.15.2 Explanatory variables in the ‘imputation model’

As discussed previously (Chapter 2.12.3), to improve the likelihood of the MAR assumption being true, it is important to collect as many potential explanatory variables as possible and including them in the ‘imputation model’. This does not mean variables should be included, as too much information may cause overfitting or cause the ‘imputation model’ to become unstable in small data with sample size or a large FMI. Careful consideration should be taken when planning the ‘imputation model’,²⁹⁶ explanatory variables should include:

- All variables included in the ‘substantive model’
- Variables not in the ‘substantive model’ but do provide information regarding the probability of a missing value and the likely value currently missing.
- Variables not in the ‘substantive model’ but provide information regarding the likely value currently missing.
- There is no need to include variables not in the substantive model and only provide information regarding the probability of a missing value.
- The variables included should reflect any interactions, non-linear relationship, or multilevel structure.
- Where possible, passive variables should be included. Passive variables are secondary variables that are generated from the primary variables to be imputed.
- The outcome variable of the ‘substantive model’.

2.12.15.3 ‘Substantive model’ outcome as a ‘imputation model’ predictor

The imputation techniques described here concerns the imputation of missing predictor data based on the available observed data. The outcome variable from the ‘substantive model’ contains important information that should be included within the ‘imputation model’, when imputing any variable that is also to be included within the main analysis.²⁹⁸ If the percentage of missing data in the imputed variable is large then not including the outcome variable as a covariate in the imputation model will cause the association between the imputed predictor variable and outcome in the main analysis model to be diluted, as the imputed values will not carry the same level of association as the observed values.³¹⁵

2.12.15.4 Limitations

Multiple imputation is able to handle large amounts of missing data. However, if any part of the imputation procedure has been defined incorrectly producing bias, then the result will be exaggerated in data with large amounts of missing values. This may happen if the ‘imputation model’ has been miss-specified e.g. ignoring the outcome variable, or other strong predictors of missing values, or the missing data is not MAR.

The MAR assumption is an integral part of the MI procedure. If the data is not considered MAR then the imputations, and subsequently the results, will likely be biased. In order to be confident that the data is MAR it is important that as many potential covariates are recorded and included in the imputation model as is possible. It should be noted that the MAR assumption is not a testable assumption.

2.12.16 Multiple imputation – with multilevel data

Multilevel data is typically characterised as being when single data observations, usually based around individuals, cannot be considered independent of each other and instead are inherently clustered together by a common factor. The typically used example is school exam results, where the individual level data is each child's exam result and the cluster is the class, or school, or local authority.

Here in the time-series design to be used in this thesis, we are concerned with comparing pollution levels on different days and so the individual level data is the pollution measured on each distinct day. The estimated pollution levels of the city are measured by multiple monitors across the city, (see Chapters 1.3 & 3.4) with the data restricted to one available monitor for any individual day. Across the 30 year study period multiple monitors are used and so the daily pollution estimates are clustered within monitors.

A multiple imputation model is derived from the conditional distribution of the missing data given the observed data. The conditional distribution of the missing data will depend on any correlated variation in the data, which in the multilevel context means each observation will be correlated with the other observations from within the same cluster. The aim is to account for the multilevel structure and make sure the imputed data contains the same variance and covariance properties as the actual data.²⁹⁶ Currently, methods of dealing with multilevel data in terms of imputing the data are limited; two potential methods are available both with their advantages and disadvantages.

The only option known at the time that fully incorporated a multilevel structure into the imputation model was proposed by Carpenter et al. (2011) and implemented in their freeware software REALCOM.³¹⁶ In order with the aim to reduce error caused by failing to take into account both between and within cluster variation in the imputations. Carpenter et al. approximated the underlying predictive distribution by using a Joint Modelling approach and fitting a multivariate latent normal model with random effects.

A Monte Carlo Markov Chain simulation is subsequently used to fit the imputation models and generate multiple imputations.^{296,317}

The main limitation (with respect to the air pollution data) is that unlike MICE, the model relies on all covariates in the being complete data i.e. no missing other than in the variable to be imputed. The autocorrelation present in time-series pollution data and short-term time dependent changes means the pollution measurements on days leading up to and following the imputed day (lag 0) are likely to be included as covariates. Anything between a week and 30 days prior and post lag 0 may be required. If this is the case, then it is feasible that in a complete cases imputation model a considerable amount of data would be lost as one missing day could relate to 60 days excluded.

The simplest method is to perform the MICE technique separately within each cluster. This means that all data can be employed even if missing data is present in an important predictor of pollution data. This treats the measurement between monitors and subsequently the imputed values as independent. Though if imputations are modelled using values from within the monitor some between monitor correlation should be retained. This means the imputation model was solely dependent on the values within the monitor, if the monitor has a short activity period with limited measurements available then certain characteristics present in the monitor and within all monitors may be missed such as long-term time trends and seasonal changes if the time is less than a year.

2.13 Aims and objectives of the thesis

2.13.1 Overall aim

The overall aim of this thesis was to describe and compare the delayed effect of air pollution on acute mortality risk from pneumonia, chronic obstructive pulmonary disease, and ischaemic heart disease.

2.13.2 Specific objectives

The specific aims of this thesis were:

- 1- To determine for each cause of death an appropriate lag-response structure.
- 2- To evaluate exposure-response linearity for the entire lag period.
- 3- To investigate in detail the influence of temperature on the three specific causes of death.
- 4- To investigate the influence of missing pollution exposure data and the usefulness of appropriate missing data analysis techniques.
- 5- Using hospital admissions data, attempt to determine if subject location during exposure influences the results.
- 6- In the main analysis, determine the influence of exposure measurement characteristics in the form of extreme outliers and missing data.

3 STUDY METHODS AND MATERIALS

3.1 Introduction to methods and materials

This chapter outlines the data and analysis methods applied in this study. Beginning with the data the study population, the study time-period, any subject inclusion or exclusion criteria and the data collection are all outlined. A description of the available pollution and temperature exposure data will be provided in terms of the monitoring types, their characteristics (active period, missing data, etc), and the final monitor inclusion/exclusion procedure. Then methods to generate the parameters to represent the effect of the exposure variables in the analysis model, including any manipulation to the exposure data, analysis accounting for missing data, and sub-group analysis are discussed. The outcome data ‘daily death count’ and its characteristics are defined initially for all-cause deaths but primarily for cause-specific definitions relating to the study objectives. An explanation of the study design, modelling methods, and finally the statistical analysis procedure are reported.

3.2 Study population and study period

The initial cohort contained all persons located in Scotland between January 1980 and Dec 2011. A person entered the study if they died within the study period from a non-accidental or non-self-inflicted death. The study subjects were assigned an underlying cause of death classified either by ICD-9 (pre 2001) or ICD-10 (post January 2001).

3.3 Participant data recruitment, administration, and acquisition.

This was a records-based study of historical events, with no active recruitment involved. Routinely collected sources of data were used for mortality and hospital admissions information related to each subject. Scottish mortality and hospital admissions data were provided by the Information Services Division (ISD) Scotland, collected as part of their statutory responsibilities.³¹⁸ The primary dataset request was made to ISD Scotland. Mortality and corresponding hospital admissions data for Scotland, for the cities Edinburgh, Glasgow and Aberdeen, collected between 1980 and 2001 were already possessed by The Centre for Occupational and Environmental Health, and have previously been used to investigate temperature mortality.^{13,272,273} A request was made to ISD to update the mortality & hospital admissions data for the years 2001 to 2011. The data file included demographics such as gender, age of death, and sector level postcode, as well as information on date and cause of death.

When a death occurs, a doctor completes a death certificate containing information relating to the subject, as well as the date of death, and cause of death. The doctor was asked to fill out the cause of death fields such that, in the opinion of the doctor, the first cause of death field represents the final contributory factor leading to death, often defined as the ‘immediate’ cause of death. The following fields relate to the contributory causes in the reverse order that the doctor felt lead up to death, with the final field relating to the cause that began the chain of events, defined as the ‘underlying’ cause of death. A separate set of fields provide the secondary causes that were present but not felt to be part of the central chain of events. Each death was registered with the National Records of Scotland,³¹⁹ where the order of causes may be changed such that the most appropriate underlying cause of death was chosen as defined by the World Health Organisation (WHO) and the International Statistical Classification of Diseases and Related Health Problems,³²⁰ ICD-9 (pre 2001) and ICD-10 (post 2001).¹⁶⁹ Once processed the data were passed to ISD Scotland. Once cause of death has been processed by GRO Scotland the chain was removed. The underlying cause of death was set in the data file as the primary cause of death and all other [secondary] causes were given no clear order.

3.3.1 Participant consent

It was not possible for this study to have a detrimental effect on any of the study participants as subjects can only enter the study once they have died. The data requested did not include any directly identifiable information, meaning it was not possible to contact the subject or a family member to gain consent; in any case given a dataset of this size it would have been highly impractical.

3.3.2 Ethical & security considerations

The deceased study participants cannot be negatively influenced however there is still an ethical duty to use anonymised data when using their records. To match exposure data with other participant data we required the date of death and sector-level postcode. The sector-level postcode is between four and six characters long including the space (e.g. M13 9) of a 7 to 8 character long UK postcode (e.g. M13 9DD). Sector postcode in 2011 corresponds to 11,197 unique codes or an average 2,358 homes per sector postcode in the UK.^{321,322} This differs depending on population density, although postcode sector alone would not allow identification of a subject. However, the additional variables of cause of death, gender and age when combined may allow identification to be possible in some cases. The presence of semi-identifiable information, and the inability to gain participant consent, means data protection was an important issue. The appropriate data protection procedures were implemented,³²³ including acquiring the ethical approval along with implementing the relevant information governance procedures regarding storage and day to day use.³²⁴

The study was reviewed by the NHS Research Ethics Committee and a favourable ethical opinion was received in November 2011 (REC reference 11/NW/0768). The study was also submitted for endorsement by the University of Manchester senate committee on the ethics of research on human beings. The Information Services Division (ISD) Scotland, who provided the primary dataset, requested a Privacy Advisory Committee (PAC) approval be obtained. This was equivalent to the NHS Confidentiality Advisory Group (formally known as National Information Governance

Board for Health and Social Care – replaced May 2013) ethics approval. The PAC approval was granted in May 2012.

The semi-identifiable nature means that adequate data protection procedures were put into place. All directly identifiable data (Names, Full Address, NHS Number, etc) were removed prior to extraction and the data were provided in the post via recorded delivery on a CD. The CD was encrypted using WinZip V9.0 or higher with an encryption level of 256bit as instructed by ISD Scotland. The data were stored using the formats suitable for the statistical packages SPSS or Stata on the University network storage infrastructure housed at the University's data centres. The network storage infrastructure and data centres were implemented and maintained by the University of Manchester IT Services Division and were internally and externally audited at regular intervals. The data were accessed via a mapped network drive on a password-protected Dell desktop computer that was located in a secure swipe card accessed room. The computer has Windows 7 with the firewall enabled and configured to prevent remote access and programmed to automatically update antivirus signatures, and download and update any Microsoft Operating system and application security patches on a weekly basis. The computer does not cache copies of the data with all versions to be stored on the network storage infrastructure. All data manipulation and analyses of sensitive material were performed by the primary investigator only, with publication and dissemination at the aggregated level only.

3.3.3 Exclusion criteria

The initial study population comprised of all residents within Scotland during the study period, those located outside major urban areas of Scotland were excluded due to limited pollution monitoring in the area and an expected low level of pollution exposure. Home postcode populations (at the 'sector' level) were further excluded if located within an urban area containing a small exposed population, defined to be less than 50,000 people, as estimated by the National Records of Scotland in the mid-2011 and mid 2012 population estimates.³²⁵ Urban areas were also removed if exposure data were unavailable due to a lack of monitoring during the study period. These areas were identified using the 'Data Availability' tool on the Department for Environment Food & Rural Affairs.³²⁶ The study population was then restricted to those exposed in the urban areas: Aberdeen, Edinburgh, Glasgow, and Inverness. Though small, Inverness was kept in order to maximise the PM10 and NO2 data considered to be sparse in the remaining cities.

3.3.4 Expected sample size

The number of deaths in this period was dependent on the Scottish death rate during the study period. Death rates have shown a gradual downward trend since 1980 with the total number of deaths recorded by GRO Scotland to be 53,856 and 53,967 for 2009 and 2010 respectively.³²⁷ A potential dataset could then be drawn from approximately 1.9 million deaths during the study period. This was reduced when inclusion and exclusion criteria were employed, and the focus changed to densely populated urban areas such as Edinburgh and Glasgow. Death rate information for council areas in Scotland was only available for 1991 onwards,³²⁷ even so it was estimated that a potential all-cause of death dataset would be approximately 800,000 subjects.

3.3.5 Power calculation

Standard power calculations are difficult to perform due to characteristics of the study design, such as no 'non-exposure' comparison group. To gain some insight into the likely power of the study crude power calculations were performed based on data in a case-crossover design for subjects with 'pneumonia' as their primary cause of death (N=4575) within Edinburgh (1981-2001) and corresponding available controls days (N=15233). A 30 day average of daily black smoke exposure was categorised into: <10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, and >70 μgm^{-3} . Study power to find a significant effect ($p<0.05$) for the higher categories versus the surrogate 'non-exposure' base category <10 μgm^{-3} was evaluated using a matched case-control sample size calculation where the life-time exposure was assumed to be constant between the case and control groups. The results were expressed as the minimum value of the true rate ratios that could be detected with 80% power given the observed sample sizes. The minimum rate ratios associated with an individual day within the 30 day period were calculated to be 1.003 (10-20), 1.005 (20-30), 1.009 (30-40), 1.021 (40-50), 1.019 (50-60) and 1.017 (60-70) 1.035 (>70 μgm^{-3}).

These calculations indicated a lack of power to observe the potentially small effects experienced in pollution studies, especially for the largest pollution concentrations, where number of exposed days was small. This power calculation may be under representative of the actual power given the data used, as it cannot account for factors present in this study. Exposure variables, both pollution and meteorological, were modelled as continuous variables that account for delayed effects and exposure-response relationships. These factors were not present in the power calculation, meaning the true power was likely to be greater than we can estimate.

3.4 Independent exposure data

Several pollution exposure measurements are commonly used due to their toxicological makeup, proposed influence on biological outcomes, and availability. These include: black smoke (BS), sulphur dioxide (SO₂), particulate matter with aerodynamic diameter less than 10µm (PM₁₀), fine particulate matter with an aerodynamic diameter less than 2.5µm (PM_{2.5}), ozone (O₃), carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO₂). Though fixed-site monitor exposure data were freely available via an internet request, complete data from one measurement source covering the entire 30 year study period were unavailable. Both pollution and meteorological exposure data such as temperature, relative humidity and wind speed were available for Scotland in some but not all areas. The following will describe data extraction, any procedure regarding monitor choices, and manipulation of the exposure data.

3.4.1 Accessing exposure data

Unlike health data, access to pollution and meteorological exposure data requires no ethics approval. Pollution data can be freely acquired from publicly accessed databases and was obtained through an internet request from the *Department for Environmental Food and Rural Affairs* (DEFRA) or alternatively the *Scottish Air Quality and Statistics database* at the following websites:

<http://www.defra.gov.uk/>

<http://www.scottishairquality.co.uk>

Daily, weekly, or monthly pollutant data were available from 120 plus automatic and 150 plus non-automatic monitoring stations located across the UK. These 300 pollution monitors were grouped into networks that monitor certain information, or for particular reasons, and may be run only for specific periods of time. All available data were obtained for all monitors active during the period Dec 1979 to Dec 2011 located in densely populated regions defined by the inclusion and exclusion criteria (Chapter 3.3.3).

Meteorological data were obtained from the British Atmospheric Data Centre (BADC), which provides data for free if the purpose is academic. The BADC website is.

<http://badc.nerc.ac.uk/home/index.html>

The BADC contains up-to-date met office data from more than 100 UK sites in addition to data from other sources. For the purpose of this study, the database used was the UK Met Office MIDAS Land Surface Observations Database. This contains both hourly and daily data from UK stations in 110 UK counties since 1853. Data for the period Dec 1979 to December 2011 was provided on request.

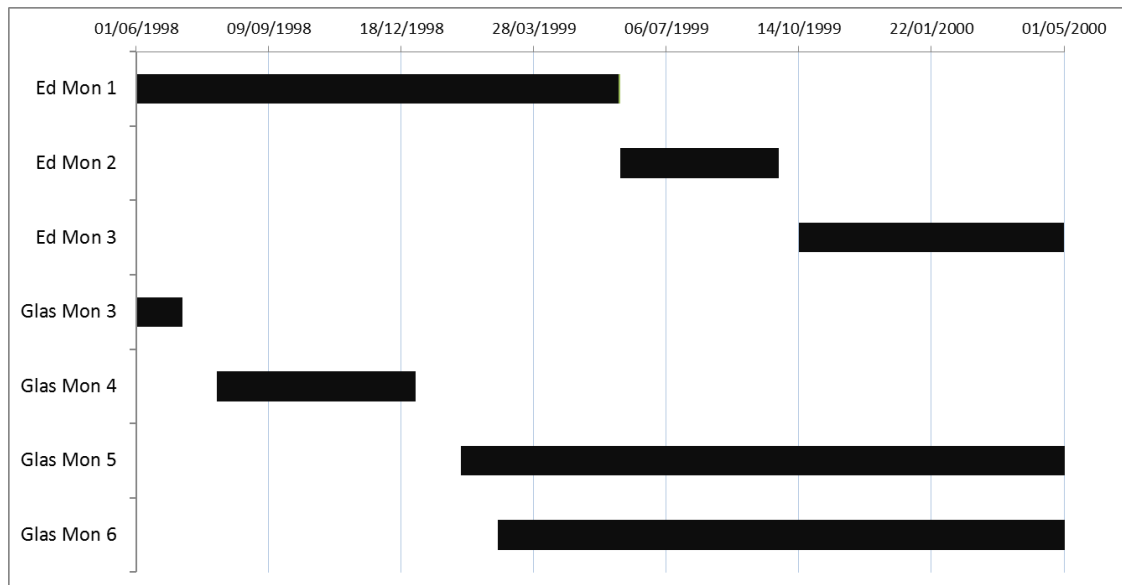
3.4.2 Defining the exposure data

To focus the analysis the pollutants black smoke, sulphur dioxide, nitrogen dioxide, and the particulates were identified and analysed, with black smoke as the primary pollutant of interest. Black smoke was chosen as it represents a good measure of overall pollution relating to combustion and health outcomes, encompasses a intermediate particulate size thought to be approximately PM₄, and measurements were taken for the majority of the 30 year study period.⁴⁹⁻⁵¹ The secondary pollutants have all been shown to be linked to the outcomes of interest, and provide an interesting comparison to the primary pollutant.

3.4.3 Monitor identification criteria

Ideally multiple concurrently active monitors located across the city for the full study period would be employed. However, with such a long study period many monitors have had periods of dormancy, some concurrently, and some consecutively, and some after a delay. To illustrate the issue a fabricated example of several monitor activity periods is given in Figure 3.1 for Edinburgh and Glasgow. Exposure recordings from different monitors contain variation, even disregarding sites location and differences in measurement techniques. Differences inherent in the monitors themselves will cause extra random and systematic differences in the measurements. Exposure data measured from several monitoring stations means that heterogeneity between monitors may be present. A difference between the true exposure and the measured exposure which is consistent within the monitor would have little importance in this analysis as it would likely only influence the intercept. However, if variation in the exposure measurement differs between monitors then effect estimates related to the change in true pollution may be biased if they were treated as homogeneous in the analysis model. To remove as much bias as possible we need to minimise the amount of between monitors variation.

Figure 3.1 – Fabricated example of monitor ‘active periods’ for seven monitors in Edinburgh (Ed) and Glasgow (Glas)



Due to the study length no monitor runs for the entire period, and so a procedure was defined for identifying the optimum choice of monitors. The following was used to identify pollution monitors and define the pollution estimates for each city:

- All monitors within the Electoral boundary as outlined by the Local Government Boundary Commission for Scotland were included in the monitor selection process.
- Types of monitor site were ranked in the following priority order: background, roadside, and finally industrial/hotspots, with industrial/hotspots to be avoided unless absolutely necessary.
- To reduce between monitor biases, monitors covering the longest consecutive time period possible were given priority, but were truncated to the end of the nearest complete month.
- If more than one monitor was suitable for the individual exposure assessment and concurrently active, then measurements were assigned to the nearest subject based distance between home postcode 'sector' and monitor postcode.

The following criteria were used to determine the choice of meteorological monitors and define the construction of temperature estimates for each city.

- All meteorological monitors consistently recording measurements for all hours of the day were considered.
- Priority will go to monitor(s) located at the city centre or within the electoral boundary as outlined by the Local Government Boundary Commission for Scotland.
- If multiple monitors existed, to reduce between monitor biases, monitors that covered the longest consecutive time period were chosen.
- Where meteorological monitors within the city were not available, the closest suitable monitor was used.

3.4.4 Exposure parameters - Construction

The exposure data for temperature and air pollutants (BS, SO₂, NO₂, PM₁₀, & PM_{2.5}) were stored as hourly measurements (temperature) and daily 24 hour averages (pollution). Any day with more than 50% of hours with measured temperature data were included. To maximise the exposure-subject relationship the daily average temperature during the likely active hours during the day (7am to 11pm) was calculated. The daily temperature and pollution measurements were referred to as the same day exposure parameter (i.e. lag 0), the exposure for the day prior was set as lag 1, and so on with the maximum analysed here as the 30 day prior exposure window prior (i.e. lag 30). Any average exposure across a lag period e.g. average over lag 0 and lag 1 was denoted as a lag 0-1 parameter. Lag 0 was dropped from all analysis due to uncertainty over the period the 24 hour pollution average covers (i.e. midnight to midnight or 9am to 9am) and to reduce the likelihood of measurements being taken after the time of death. Figure 3.2 represents an example of the 30 day exposure period (X) leading up to a subject's case or control day (O).

Figure 3.2 - Example of the 30 day exposure period (0=day of death, X=exposure period)

August						
SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23 X	24 X	25 X	26 X
27 X	28 X	29 X	30 X	31 X		

September						
SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
					1 X	2 X
3 X	4 X	5 X	6 X	7 X	8 X	9 X
10 X	11 X	12 X	13 X	14 X	15 X	16 X
17 X	18 X	19 X	20 X	21 X	22 O	23
24	25	26	27	28	29	30

If exposure variables representing each individual day were included in the statistical model, this will cause collinearity due to autocorrelation between adjacent exposure days (Chapter 2.9.2). Therefore the exposure window was modelled in three ways: a distributed lag non-linear model, a lag stratified model, and a distributed lag model.

To flexibly model any non-linearity in both the exposure-response relationship and the 30 day lagged-response relationship, new exposure parameters representing ‘distributed lag non-linear’ were generated.^{271,282} A set of parameters representing ‘natural cubic splines’ were generated independently for the exposure range and the lag period. The knot points associated with both exposure and lag were positioned at equally spaced percentiles of their corresponding marginal distribution as recommended by Harrell in 2001 (see Table 3.1), with the first knot point located no less than the fifth smallest value and the last knot located no greater than the fifth largest value of the data.³²⁸ Because a two-stage individual level meta-analysis was to be performed (see details in Chapter 3.9) within each monitor before being combined. The knot positions were based on the monitor with the smallest absolute range and applied to all monitors included.

Table 3.1 - Harrell suggested percentiles for knot positions within a data range.³²⁸

Knots	Harrell's Percentile Position						
	10	50	90				
3	10	50	90				
4	5	35	65	95			
5	5	27.5	50	72.5	95		
6	5	23	41	59	77	95	
7	2.5	18.33	34.17	50	65.83	81.67	97.5

If there are too many knot points, the ‘natural cubic splines’ can be influenced by random noise, if too few potentially important non-linear relationships will be smoothed out. The final ‘distributed lag non-linear’ parameters allow for non-linear changes across exposure range and lag period by combining two sets of parameters using interaction terms. The number of combinations then increases by $e * l$, where e and l represent the number of different knots across the exposure range and lag period respectively. It was decided that more than 5 knots were unnecessary; therefore natural

cubic spline parameters for both factors were generated with 3, 4, and 5 knots points, creating 9 combinations for comparison.

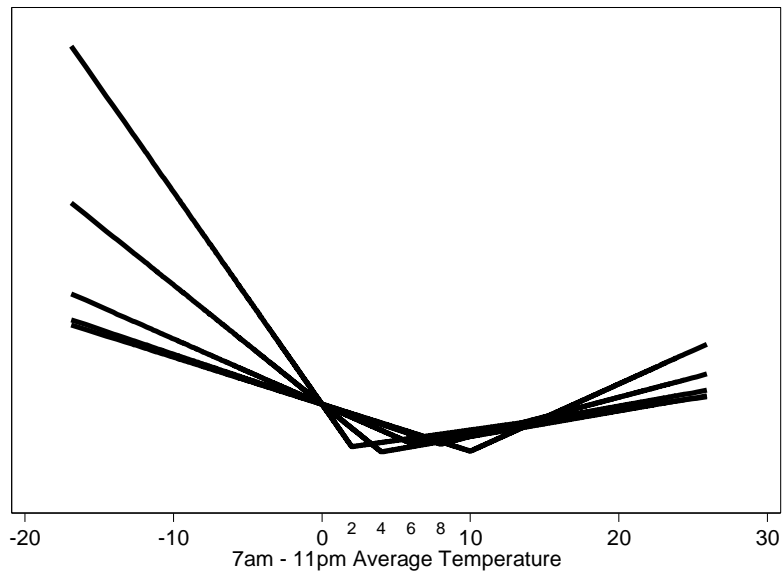
The advantage of the ‘distributed lag non-linear’ model is that it can allow large amount of flexibility in both the exposure range and the lag period, however it may be influenced by data-specific random variation. Additionally, the parameter effect estimates generated from the ‘distributed lag non-linear models’ can be difficult to interpret at face value.²⁷¹ To aid interpretation the relative risk compared to lag 1 and the mean exposure was predicted and plotted across both the exposure range and the 30 day lag period in a 3D surface plot. Further direct interpretation was attempted through the ‘distributed lag’ and ‘lag stratified’ models. Firstly, the ‘distributed lag non-linear’ model, indicated if non-linearity across the exposure was present across the full lag period and secondly, if non-linearity was present it indicated the appropriate number of knot points. Any non-linear exposure-response relationship was modelled using a double ‘linear threshold’ model (Chapter 2.8).

To determine the number and position of the knot point the following procedure was implemented:

- To determine the exact position, the exposure values for lag 1 (day before) were used.
- If one knot was required. The exposure range was split initially at the 5th percentile and two continuous linear terms connected the exposure range above and below the 5th percentile value. The two terms were fitted in the model and the goodness of fit statistic - Akaike Information Criteria (AIC) - was recorded.
- The process was repeated with the knot point moved at a time in small increments (2°C temperature, 2% per pollutant) along the exposure range until the 95th percentile. Each knot point was modelled separately and each time the AIC was recorded.
- Finally, the most optimum model i.e. the most optimum knot point, was identified as the one with the smallest AIC value.

Figure 3.3 gives a graphical example of several separate threshold models for temperature, where the threshold knot point was moved along the temperature exposure range for each model.

Figure 3.3 - Graphical representation of several example threshold models with the knot point moving along the temperature exposure range.



If two knot points were more appropriate then the double linear threshold model was extended to a triple linear threshold with two knot points. Here:

- The exposure range was again split with a knot at the 5th percentile point.
- Whilst maintaining the position of the first knot, a second knot was created at the 5th percent plus 2oC (2% if a pollutant), creating three linear parameters connecting across the exposure range.
- The three continuous linear parameters constrained to connect at each knot point were modelled, and the AIC value recorded.
- The procedure was repeated firstly moving the second knot point along in the small increments towards the 95th percentile, each time recording the AIC value.
- Once completed, the 1st knot point was advanced by a small increment, and the second knot point procedure repeated.
- Once the 1st knot reaches the 95th percentile point, the most optimum model combination with the smallest AIC value was chosen.

The optimum exposure-response relationship identified for lag 1 was then replicated for the entire 30 day lag period. The lagged-response effect was then modelled using a 'lag stratified' model and a 'distributed lag' model. The 'lag stratified' model overcomes

autocorrelation by stratifying the lag period into groups and modelling the average exposure. Initially an average exposure for the entire 30 days was modelled, with a subsequent model containing 5 lag periods of 6 days each, set to be 1-6, 7-12, 13-18, 19-24, and 25-30 days. The average exposure over prior 6 (or 30) days was included as a parameter in the model. Effect estimates associated with lag periods will represent an average unit increase across the x days, or in other words a unit increase occurring on each day within the lag period. To simplify comparisons between lag periods of different lengths, the effect estimates were adjusted to represent the effect of a unit increase on one single day within the lag period.

The 'lag stratified' model applies a strict flat relationship across each lag period, this means any within lag period (1-30 or 1-6, 7-12, etc days) relationships were smoothed out. The 'distributed lag model' provides a useful half way measure, allowing for potential short-term (day to day) changes and yet less subject to data specific random variation. It produces a smooth lag-response relationship through a polynomial parametric function. Use of a polynomial function can grant greater flexibility, but it can also impose a structure without flexibility perhaps implying a relationship may not be present. For example, a 1st degree polynomial will force a linear lag-response relationship regardless of its true nature. In order to allow the data to dictate a higher order polynomial function can be applied. In this case a cubic function was chosen. The estimates of the coefficients for the cubic function can reveal both complex (cubic) and simplified (e.g. quadratic, linear, flat) relationships across lag-response structure.

Measures from multiple pollution/temperature monitors within multiple cities were used. The transformed exposure variables representing the above exposure and lagged structures were generated for each pollution monitor separately. As each monitor differs in the exposure range it has measured, to be consistent all transformed exposure variables regardless of the monitor were generated assuming the range identified by the monitor with the smallest exposure range. Once all new variables were generated, the day-to-day exposure estimates for each monitor were matched to the case and control days of the nearest subject based on the shortest direct distance between the centre of the subject's home 'sector' postcode and the monitor postcode.

3.5 Study outcome or dependent variable

The primary outcome was based on the subject's day of death. A relative risk comparison was made between exposures related to day of death and exposures related to a day, or days, when the death did not occur. The primary analyses were related to the specific causes of death pneumonia, chronic obstructive pulmonary disease (COPD) and ischaemic heart disease (IHD). Study subjects were defined as having one of the three causes of death if they had the International Disease Classifications (ICD) code for pneumonia (ICD 9 – 480-486, ICD 10 – J12-J18), chronic obstructive pulmonary disease (ICD 9 – 490-492, 494, and 496 ICD 10 – J40-J44, J47) or ischaemic heart disease (ICD 9 – 410-414, ICD 10 – I20-I25). Two sets of cause of death indicator variables (e.g. pneumonia present yes/no) were created. The first set was based on the primary cause of death field only, and the second based on any cause of death field (i.e. the primary cause of death field and all secondary cause of death fields). Secondary objectives were achieved by restricted the data to the relevant sub-groups and evaluating the effect estimates produced.

3.6 Study design

The time-stratified case-crossover design was chosen as, within the design, it elegantly and automatically accounts for a large amount of confounding through time-dependent and subject-specific factors (see Chapter 2.6). Additionally, the design allows for the stratification and comparison subject split by secondary covariate information such as demographic and hospital admission data for each subject. The time-stratified case-crossover design compare exposure between days associated with the subject's day of death (the case day) and those with a similar time period when death did not occur (the control day). Control days were selected to be the same day of the week for each week within the same month. See Figure 2.6 as an example calendar for September. If a study participant died on the Friday 22nd September (X) then their control days were all remaining Fridays within September (O), in this case the 1st, 8th, 15th and 29th.

3.6.1 Data manipulation

The three raw datasets representing Mortality, Hospital Admissions and Exposure data were manipulated and combined in order to represent the time-stratified case-crossover design. The procedure had the following stages:

- 1) All subjects failing to match the inclusion and exclusion criteria were removed from the mortality data. Leaving a dataset that represents subjects with a home postcode in the urban areas Aberdeen, Edinburgh, Glasgow, and Inverness and a death between Jan 1980 and Dec 2011. Using the cause of death fields and ICD codes, new variables were created to identify subjects who died from pneumonia, COPD, and IHD based on primary cause of death field only (also known as the underlying cause of death) and based on any cause of death field (primary and any secondary cause of death field). Each data row represents a single subject identified from their ISD-provided unique identifier.
- 2) The mortality dataset was converted to represent the case-crossover design, such that each individual has a line of data for the case day (date of death) and up to five control days, set to be the same day of the week for each week of the same month. This was done by firstly creating a separate template data file:
 - a. The template begins with a single date variable for every day within the study period (variable name=day0), a unique id (id=1,2,3...N where N was total no. calendar days between 1st Jan 1980 and 31st Dec 2011), and a separate date of death (dod) variable equal to day0.
 - b. Additional variables were created to represent control dates at the following seven day intervals day0-28, day0-21, day0-14, day0-7, day0+7, day0+14, day0+21, & day0+28.
 - c. Reshape the dataset into a 'long format' such that each 'dod' date now has 9 rows each (1 case date and 8 control dates). Drop those control dates located in a different month to the case date. Create a binary outcome variable identifying the case (1) and the control (0) rows. Leaving each 'dod' with 4/5 rows, 1 case and 3/4 control days depending on location within the month.

- d. The mortality data file was then matched to the case-crossover template such that the subject specific data were repeated for each case and controls days. Note, in a mortality dataset with multiple deaths per day this may need to be repeated in separate datasets relating to a single individual per day, and then re-combined (appended) later. A fabricated example of the data file is given in below in Table 3.2.

Table 3.2 - Generated example of the data structure

id	date	outcome	dod	diffdays	gender	age	city	Subj Covars ---->
1	02-Jan-80	1	02-Jan-80	0	M	65	ABR	----->
1	09-Jan-80	0	02-Jan-80	7	M	65	ABR	----->
1	16-Jan-80	0	02-Jan-80	14	M	65	ABR	----->
1	23-Jan-80	0	02-Jan-80	21	M	65	ABR	----->
1	30-Jan-80	0	02-Jan-80	28	M	65	ABR	----->
2	04-Jan-80	0	11-Jan-80	-7	M	82	GLAS	----->
2	11-Jan-80	1	11-Jan-80	0	M	82	GLAS	----->
2	18-Jan-80	0	11-Jan-80	7	M	82	GLAS	----->
2	25-Jan-80	0	11-Jan-80	14	M	82	GLAS	----->
3	02-Jan-80	0	23-Jan-80	-21	F	84	EDIN	----->
3	09-Jan-80	0	23-Jan-80	-14	F	84	EDIN	----->
3	16-Jan-80	0	23-Jan-80	-7	F	84	EDIN	----->
3	23-Jan-80	1	23-Jan-80	0	F	84	EDIN	----->
3	30-Jan-80	0	23-Jan-80	7	F	84	EDIN	----->
4	05-Jan-80	0	26-Jan-80	-21	M	48	ABR	----->
4	12-Jan-80	0	26-Jan-80	-14	M	48	ABR	----->
4	19-Jan-80	0	26-Jan-80	-7	M	48	ABR	----->
4	26-Jan-80	1	26-Jan-80	0	M	48	ABR	----->

- 3) Finally, independent variables representing the pollution and meteorological exposure, or additional covariates created from the hospital admissions data can then be matched to the case-crossover outcome file based on the unique identifier defined by ISD Scotland, or the date and location variables.

3.7 Analysis with missing exposure data

Ideally the exposure dataset would contain measurements for all days within the study period in this case 01/01/1980 – 31/12/2011. In a standard ‘complete cases’ analysis, any missing exposure data on either the case day or all of the control days would result in the subject being removed from the analysis entirely. To avoid this and maximise the available data a ‘multiple imputation’ procedure replaced missing values with several imputed observations. As explained in full in 2.12, multiple imputation is a validated method in statistics under given assumptions. However, patterns of missing exposure data may not fit and their impact in the case-crossover analysis has not been showed. Hence, a simulation study determined the influence of missing exposure data on estimates of the true exposure-mortality relationship when analysed using the standard complete cases method and when using missing data techniques such as multiple imputation. The missing data technique was then applied to the main analysis and compared to the standard results from a complete cases analysis.

3.7.1 Imputation model – The data structure and primary variable

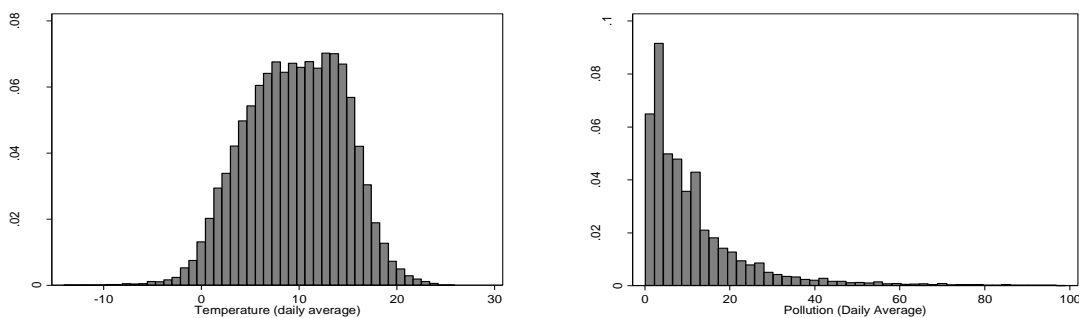
To impute missing exposure values the dataset was structured in a time-series format where each line of data represents the exposure measurement (and associated factors) for an individual day within a monitor for that period. Table 3.3 gives a fabricated example of the data structure. Note, in the example Edinburgh monitor 1 was active until 31st May 1999 after which it was replaced by Edinburgh monitor 2. Glasgow monitor 5 active period overlaps with Edinburgh monitor 1, and so dates were not considered unique in this dataset (see example in Figure 3.1). The primary information to be imputed relates to the same day exposure observations (‘Exp lag 0’ in Table 3.3) for any day within the activity period of each monitor. This was denoted as a present observation or a blank missing value. Illustrated in Figure 3.4 are examples of the likely temperature and pollution distributions. Whereas temperature can be considered as following the Gaussian distribution, pollution measurements tend to be highly skewed by some extreme values. To improve the distributional properties all pollution exposure variables (x) were log transformed, with zero values accounted for by adding a constant (c), where c is the smallest observed value divided by two. The dataset contains a

number of additional variables that will help describe the characteristics and predict the observed and missing values.

Table 3.3 - Example of the exposure dataset in preparation for multiple imputation analysis

Date	Exp lag 0	BS Monitor	Area	Additional Factors-->
:	:	:	:	--->
:	:	:	:	--->
25/05/1999	48	ED Mon 1	Edinburgh	--->
26/05/1999	35	ED Mon 1	Edinburgh	--->
27/05/1999	36	ED Mon 1	Edinburgh	--->
28/05/1999	.	ED Mon 1	Edinburgh	--->
29/05/1999	.	ED Mon 1	Edinburgh	--->
30/05/1999	65	ED Mon 1	Edinburgh	--->
31/05/1999	119	ED Mon 1	Edinburgh	--->
01/06/1999	98	ED Mon 2	Edinburgh	--->
02/06/1999	66	ED Mon 2	Edinburgh	--->
03/06/1999	.	ED Mon 2	Edinburgh	--->
04/06/1999	22	ED Mon 2	Edinburgh	--->
:	:	:	:	--->
:	:	:	:	--->
:	:	:	:	--->
:	:	:	:	--->
:	:	:	:	--->
25/05/1999	88	Glas Mon 5	Glasgow	--->
26/05/1999	76	Glas Mon 5	Glasgow	--->
27/05/1999	115	Glas Mon 5	Glasgow	--->
28/05/1999	.	Glas Mon 5	Glasgow	--->
29/05/1999	126	Glas Mon 5	Glasgow	--->
:	:	:	:	--->
:	:	:	:	--->
31/11/2011	86	Glas Mon 10	Glasgow	--->

Figure 3.4 - Example representations of exposure data distribution



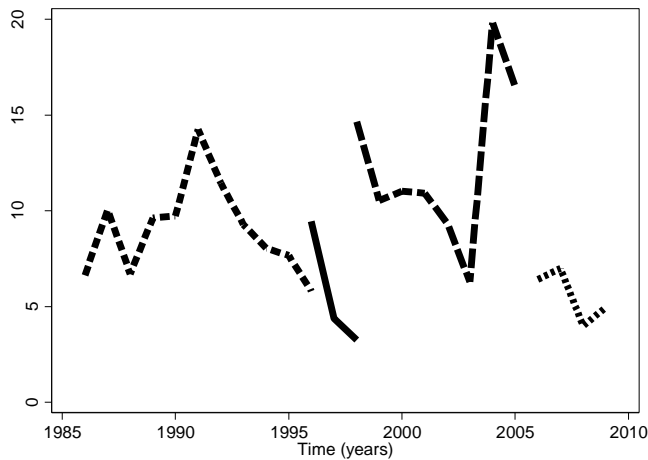
3.7.2 Imputation model - Method

The Multiple Imputation by Chained Equations (MICE) procedure was chosen to account for missing exposure data. The procedure has been explained in detail elsewhere (Chapter 2.12.13) but summarised here. Missing values were replaced with imputed values found by fitting univariate model to the observed data using variables thought to predict missing values or the underlying properties of the true values observed. New model parameters were generated from simulated draws of their posterior predictive distribution, and used to generate newly imputed values. When there were several variables with missing data, the procedure repeats firstly one by one for all variables in the model that contain missing data, and then the entire cycle, until the imputed values converge to a consensus (the ‘burn’ in period) and a set of imputed values were saved.

To ensure the distributional properties of the observed exposure data (Chapter 3.7.1) were replicated in the imputations, a log transformation with an adjustment for zero values (in this case $\text{zero} + 0.5 * \text{minimum observed pollution value}$) was performed and a predicative mean matching (PMM) applied under the MICE procedure. Also discussed in more detail in Chapter 2.12.11, PMM predicts a new value for both the missing and observed values, and then randomly selects a new imputed missing value from a pool of observed values that have the closest matched predicted values in this case set to be one of the nearest 10 observations.

In order to cover the study period and the different locations both within and between cities, the observed pollution measurements were sourced from several different monitors. This clustering within the exposure data means measurements within a monitor were likely to contain similarities not present for measurements from different monitors. An example of the between monitor heterogeneity in exposure measurements is given in Figure 3.5 for four monitors active within distinct time-periods but producing differing average pollution estimates.

Figure 3.5 – Example of the change in the pollution levels for differing monitors over time.



This may be true for both the measurements and the causes of missing observations. Applying the multiple imputation procedure across the entire exposure dataset regardless of monitor will create biased imputation results. In the absence of a suitable multiple imputation procedure at time of the analysis that can allow for multilevel data in a chained equation with predictive mean matching, imputed values were generated within the clusters of active time periods for the monitors. Bias may be increased within small sample size monitors, especially if the percentage of missing data were also large, as monitors with small active periods may have a reduced ability to account for certain time-dependent trends e.g. long-term time trends.

Here the final imputed dataset contained 10 imputed observations for each missing value chosen after a burn in period of 50 iterations and from a predictive mean matching pool of the 10 closest matched observed values.

3.7.3 Imputation model – Primary model predictors

Careful consideration was required when identifying the most appropriate imputation model. To reduce any potential bias and to improve the chances of MAR, variables should be included if they were present in the ‘substantive’ model (the main analysis model), or were predictive of the presence of a missing values and the observed values.²⁹⁶

The exposure data on adjacent days is correlated: not only will the level of exposure likely be followed by a similar level, but a day with a missing value was also likely to be followed by a missing value. This means that including adjacent days as predictors will likely improve the predictive ability of the imputation model. To improve the replication of short-term patterns in the observed exposure (i.e. changes within a few weeks to month) the imputation model included parameters representing the days prior and post the day to be imputed. The predictive influence of exposure may be greater than just for the adjacent day or days, hence it was hypothesised that a window of approximately 30 days either side of the day to be imputed would be required. Therefore the primary set of explanatory variables to be included in the imputation model were the exposure values themselves for the 30 days prior to and 30 days following day zero.

The presence of collinearity in the imputation model between potentially highly correlated adjacent exposure variables was a concern. Collinearity was monitored and if imputations appeared to be affected then measures were proposed to use a similar approach as the lag stratified model, i.e. to include parameters that represent the average of x days leading up to the day to be imputed. Including explanatory variables created from the primary variable to be imputed means that these variables will also contain missing observations that would require imputing.

Temperature is not only a primary confounder, and included as a covariate in the substantive model, but also it is likely to be predictive of both the pollution level and any missing pollution data. Pollution levels are highly correlated with temperature. This may be due to related weather factors such as still, cold weather trapping pollution causing greater concentrations, or due to chemical changes in warmer temperatures such as between nitrogen oxides and ozone, or changes in human behaviour resulting in increased car use, or decreased home heating. Temperature was therefore considered an important explanatory variable and was included in the imputation model. Temperature variables representing the same day and 3 days prior were included. As pollution measurement on adjacent days provided a lot of information regarding short-term patterns, and temperature measurement several days prior were thought to have little predictive influence on pollution value or the chance of a missing value. Hence

temperature variables representing several days prior and any days post day were not required. All exposure variables containing missing data were imputed as part of the chained equations procedure.

3.7.4 Imputation model – Additional covariates

To improve the accuracy of the imputed values and maximise the chance of MAR being present as many explanatory variables as possible should be measured and considered (see Chapter 2.12.15). The nature of the dataset means that this relates to time-dependent factors relating to the confounders discussed in Chapter 2.4.1. Given the length of the study period, long-term time trends in the daily concentration level of pollution were likely to be present. In most pollutants this tends to be a downward trend due to pollution reduction policies and improved technologies. Long-term time trends were accounted for using a linear variable representing date within the study. Differing exposure levels (and missing data levels) were also seen to fluctuate between, years, seasons and over the course of a single week. A number of binary (yes/no) explanatory variables were included to represent each year the monitor was active, each month of the year and each day of the week. Including months rather than a simple winter vs summer variable helped account for seasonal differences and monthly differences, which may be important given the monthly clustering of cases and controls within the time-stratified case-crossover study design. To reduce the number of parameters and the chance of overfitting, for each binary time-dependent parameters one category was removed. For example, the seven binary day of the week variables will be six with Sunday removed, the remaining variables when all set to 0 will identify the Sunday.

Imputed observations must retain the relationship between observed data and the main outcome variable of the substantive model. In an attempt to preserve this relationship the dependent variable of the substantive model should be included as a covariate of the imputation model. A direct representation of the matched binary case vs control(s) and the time-series exposure data structure was not available. Instead, to maintain the relationship with mortality, a discrete count variable representing the number of all-cause deaths and cause-specific deaths occurring per day within each corresponding city was generated and included as a covariate.

3.7.5 Imputation model - Summary

In summary the following outlines the proposed multiple imputation procedure to be performed.

- The primary variable to be imputed was calendar day (day 0) pollution exposure measurement.
- Primary predictors of the day 0 pollution exposure was pollution measurements on the 30 days prior (day 0 - 30) and 30 days post (day 0 + 30) the day 0. Same day (day 0) temperature, average temperature on the three days prior (or average of three days), in order to aid the imputations ability to represent short term variation and long term trends in pollution exposure and their relationship with temperature.
- Important covariates included were; count of deaths on day 0, calendar date, binary (yes/no) year, month (Feb, Mar, etc.), and day of the week (Mon, Tue, Wed, etc.).
- To maintain pollution measurement distributional characteristics a log transformation and predictive mean matching technique was employed.
- MICE for all exposure variables containing missing data within monitor clusters will generate multiple imputed datasets for each monitor separately.

3.8 Simulation study - Missing exposure data

3.8.1 Simulation study - Introduction

The proposed study design means that if a pollution measurement was missing on either a subject's case day, or all of a subject's control days then the subject was excluded from the main analysis. The total number of deaths in 2009 and 2010 were 53,856 and 53,967 respectively,³²⁷ indicating an average number of deaths of 147.6 per day. Even adjusting for the study population, this means one missing exposure observation in a 'complete cases' analysis can result in a number of lost subjects and information. If the missing exposure data can be considered MCAR or MAR (given the correct covariates in substantive model using maximum likelihood estimation), then effect estimates from complete cases analysis should still be valid. However, effect estimates from pollution studies in particular are notoriously small, so even with large sample sizes any loss of information might impact the ability to identify significant effects. The use of Multiple Imputation is an important tool; however if the imputation model was miss-specified it may cause greater bias.

Therefore a simulation study was performed with the specific aims:

- 1- To determine potential 'missing at random' data characteristics from the main datasets.
- 2- To evaluate the influence of MCAR and the identified MAR characteristics on estimating the true exposure-mortality effect in the standard 'complete cases' analysis.
- 3- To evaluate the ability of the 'multiple imputation' procedure to account for missing data when estimating the true effect estimate.

3.8.2 Simulation study – Data

This simulation study mirrored the analysis in the main study focusing on black smoke air pollution, though here the focus was on the influence of missing data and characteristics relating to missing data patterns. Two real world datasets were used.

- 1- The exposure dataset - To remove added complexity of between monitor biases, the study was based on black smoke pollution recordings from a single monitor chosen from the main study. A monitor containing the largest period of time where all black smoke pollution measurements and temperature measurements were recorded i.e. the longest time-period with no missing data.
- 2- The mortality dataset - A mortality dataset containing all-cause deaths for the corresponding time period and location was identified to match the 'exposure' dataset. The mortality dataset was set-up for the time-stratified case-crossover analysis and contained covariates (e.g. temperature) with no missing data.

3.8.3 Simulation study – Missing data characteristics

A series of different scenarios were applied to the 'exposure dataset' that simulated characteristics observed in the main datasets. Initially missing data were created under a missing completely at random (MCAR) assumption, with increasing percentage of total missing data at 5%, 10%, 15%, 20%, and 25% identified from the main dataset. To do this, each day within the 'exposure' dataset was assigned a number (1,2,...n) and using a pseudo random number generator, an integer between 1 and n was drawn without replacement. The exposure measurement on the identified date was substituted with a missing value. The procedure was repeated until a set of unique dates were identified and the required percentage of overall missing achieved.

For the second set of simulations, missing values were set to mimic missing at random (MAR) data characteristics using covariates that replicated characteristics of true pollution data. Due to the time-series structure of the 'exposure' data these MAR characteristics were time-dependent, for example, increases over time, seasonal or day

of the week. To create a dataset where the extent of missing data were influenced by these time-dependent factors, stratified random sampling was applied in the form of proportional allocation. Appropriate sampling fractions was determined for each stratum that when combined matched the total required percentage of missing data. Again within each stratum simple random sampling without replacement was employed until the sampling fraction was achieved and a set of unique dates also set to missing.

As with the observed exposure measurements, autocorrelation may be present in the probability of a missing observation i.e. a day with a missing observation will increase the chance of a missing observation in the following days. This may occur if a monitor breaks down and there is a delay in it getting fixed. If these ‘blocks’ of missing data are present, then their characteristics, of average length and variation were incorporated into the missing scenarios. In the event of blocking, the data were subjected to sequential sampling procedure. As before, a simple random sample with a uniform distribution was performed to identify the first day to be set to missing. To identify the length of the ‘block’ of missing observations a second random number was generated under normal distribution properties with a mean and standard deviation identified from the main dataset. The initial randomly chosen day and the follow days of length also randomly chosen, were then set to be missing. Note: any negative values generated from the normally distributed random number were ignored and a re-sample was performed. This one-tail truncated normal distribution was thought to match closest the true distribution of ‘blocks’ in the main study dataset.

All MAR characteristics were investigated individually and in combination. Due to the potential number of scenarios and the time frames involved, each scenario was repeated three times assuming 5%, 15%, and 25% of the complete ‘exposure’ data missing.

3.8.4 Simulation study – Procedure & performance

Firstly, the ‘true’ effect estimates for both temperature and pollution (black smoke) were calculated using the complete exposure dataset. The aim here was to determine the effect on the exposure-mortality relationship of missing exposure data in a complete cases design, and the effect under the attempt to account for it in a multiple imputation analysis. Therefore, effect from a simple linear model parameter of the lag 0 (same day effect) temperature and pollution to outcome was used i.e. no attempt here was made to account for non-linear exposure-response or any lagged response relationships.

Next, the following procedure was followed.

- 1- Starting with the 5% missing completely at random scenario, missing data were generated in the pollution measurements within ‘the complete exposure dataset’ creating a new temporary ‘missing exposure dataset’. Note that temperature was left complete at all stages.
- 2- The multiple imputation procedure outlined for the main analysis (Chapters 3.7.1 - 3.7.5) of the study was performed, generating 10 temporary datasets each with different ‘imputed’ observations for every missing observation.
- 3- The 10 ‘imputation’ datasets were simultaneously matched to the ‘mortality dataset’ based on the date associated with ‘case’ day and the ‘control’ days.
- 4- A conditional logistic regression model was fitted, firstly as a complete cases analysis and secondly in a multiple imputation analysis. In both cases the pollution parameter was included as the primary predictor of interest whilst adjusting for temperature as a covariate, and clustering around the subject id.
- 5- The effect estimates and corresponding standard errors for pollution and temperature were extracted and saved.
- 6- The datasets were returned to their original status’s (i.e. a complete ‘exposure dataset’)
- 7- Steps 1 through 6 were repeated 1000 times.
- 8- Once all 1000 simulations were completed, the effect estimates were pooled and a compared with the true pollution and temperature estimates.
- 9- Repeat steps 1-8 for each MCAR and identified MAR scenario as outlined in Chapter 3.8.3)

The performance under each scenario was evaluated and compared with the true effect estimate, based on methods described in 2006 by Burton et al.³²⁹ In brief, results associated with each missing data scenario were summarised in the form of an expected effect estimate of the raw conditional logistic regression coefficients averaged over all simulations $E(\hat{\beta})$. One thousand simulations were chosen, largely due to computing power and time constraints. A power calculation was performed to determine the number of simulations required in order to have the power to detect a difference given the true parameter estimates, were estimated³²⁹ and reported in Table 3.4. Based on the accuracy of the estimate from the true effect (0.0003302) and true variance (0.0007093^{^2}) calculated from the complete dataset. The result indicated, for a two tailed 5% p-value, a 60% power to detect a 15% difference or an 80% power to detect a 20% difference was present, given the constraint of 1000 simulations.

Table 3.4 – Sample size calculation outlining the No. simulations required given the percentage bias in the missing data simulation study

Percent Diff*	Absolute Difference	No. Simulations (Assume 5% Significance)			
		50% Power	60% Power	70%Power	80% Power
1	0.0000033	177256	226043	284798	362171
3	0.0000099	19695	25116	31644	40241
5	0.0000165	7090	9042	11392	14487
10	0.0000330	1773	2260	2848	3622
15	0.0000495	788	1005	1266	1610
20	0.0000660	443	565	712	905
25	0.0000826	284	362	456	579
50	0.0001651	71	90	114	145

*Percentage difference from true estimate (S.E.) of 0.0003302 (0.0007093)

The most useful method will produce unbiased estimates with little variability. To determine if that suitably unbiased estimates were produced, results were evaluated in the form of bias and accuracy. Due to increased uncertainty in small true effect estimates, the difference (bias) between the combined simulated effect estimate and the true effect estimate (i.e. $E(\hat{\beta}) - 0.0003302$) was reported here as percentage bias and standardised percentage bias (i.e. $100*(E(\hat{\beta}) - 0.0003302)/ S.E(\hat{\beta})$). This was a useful way of comparing both the differing missing data scenarios and the analysis methods. Because the standard errors associated with the simulated estimates are based on the

number of simulations (i.e. greater number of simulations reduces the standard error) then testing the difference (or producing confidence intervals) can be manipulated. This results in the potential for a statistically significant bias that was in practice not considered unacceptable.³²⁹ Therefore given the small effect estimate and the power calculation in Table 3.4, the a priori amount of bias considered acceptable was approximately 15%. To compare the accuracy of the two analysis methods, both the bias and variability were simultaneously measured by calculating the Mean Square Error (MSE). When the MSE tends towards zero the more accurate were the estimates of the true parameter.

Even though the focus of the simulation study was on missing data in the pollution exposure and not the temperature covariate, results were important for both. This is because, with regards to the complete cases analysis, it may be of interest to understand how missing data in the main predictor (pollution) can influence bias in the estimates of the covariate (temperature). As missing temperature observations were only present if there was a corresponding missing pollution observation, by imputing the missing pollution observation the observed temperature measurement will return in the substantive model. Hence assuming the relationship between temperature and pollution was maintained, any simulated effect estimates for temperature should return to the true temperature effect. Should this be confirmed, it will help confirm the temperature pollution relationship was maintained throughout.

3.9 Statistical analysis of the main study

All analysis was performed using Stata versions 11 to 13.³³⁰

Standard descriptive statistics in the form of cross-tabulation and graphical techniques were reported to clearly illustrate the properties of the mortality and exposure datasets. A time-stratified case-crossover study design compared the level of exposure leading up to the subject's death with the level of exposure leading up to day(s) when the subject did not die. The analysis model, a conditional logistic regression, calculates the maximum likelihood estimates for a binary dependent variable (case day = 1, control day(s) = 0), conditional on the matching participant groups.^{331,332} The model investigated the relative risk associated with the primary factor - pollution exposure - on death whilst adjusting for the main confounder temperature. The effect estimates were reported in terms of a percentage change in relative risk associated with $10\mu\text{gm}^{-3}$ increase in pollution or a 1°C increase in temperature for any individual day within the lag period. This means that effect estimates associated with a 30 day lag can be easily compared with a 6 or 1 day lag period, as reporting the average exposure means the effect estimates will represent an increase on each day within the lag, i.e. the effect if the increase occurred on each of the 6 days or each of the 30 days. It is recognised that the range in exposure measured may differ across pollutants. If the observed exposure range is distinct from $10\mu\text{gm}^{-3}$ an example of the effect size associated with the pollutant relevant inter-quartile range will also be reported.

The cause-specific analyses investigating pneumonia, COPD and ischaemic heart disease focused on subjects with the cause of death located in at least one of the primary or any of the secondary cause of death fields i.e. 'any' field with the cause of death (AFCOD). The data was stratified by an indicator variable representing each cause of death as defined by the AFCOD. To compare the exposure effect between each combination of the three causes of death, a first principles z-test was performed using the log relative risk (logRR) of two models, where:

$$z = \frac{\log RR_1 - \log RR_2}{\sqrt{(SE \log RR_1)^2 + (SE \log RR_2)^2}}$$

The relative risks (RR) associated with the three causes of death were then compared such that each two-way combination was fitted separately e.g. COPD vs pneumonia, IHD vs pneumonia, IHD vs COPD.

To simplify the model building stage, the confounder – temperature – was modelled initially without air pollution. As temperature is more likely to influence pollution rather than the reverse, the subsequent inclusion of pollution in the temperature model was thought to have minimal influence on the temperature effect. Using the exposure modelling techniques described earlier (Chapter 3.4), a bespoke i.e. uniquely fitted temperature model was generated for each cause of death. Before modelling pollution, these bespoke temperature models attempted to account for as much variation due to the non-linear temperature exposure-response and lag response relationships. The optimum model for temperature was determined by minimising the goodness of fit statistic AIC, and in the case of a tie, the Bayesian Information Criteria (BIC).³³³ The process was then repeated for the pollution data whilst adjusting for the previously defined optimum temperature model. Within each of the three modelling techniques (distributed lag non-linear, lag stratified, and distributed lag) the corresponding temperature parameters were fitted i.e. if pollution was modelled under the lag stratified model then the temperature was also modelled using the lag stratified model.

To account for heterogeneity between monitors (see Chapter 3.4.3) a single pooled exposure effect was calculated in a two stage individual participant data meta-analysis.^{334,335} Here, the same conditional logistic regression model was fitted independently within each monitor (stage 1), and then the individual monitor effect estimates were aggregated using an inverse variance weighted average in the form of a random effects meta-analysis (stage 2). To improve model stability when the number of studies became small, and because both the within and between monitor effect estimates could not be considered normally distributed, a DerSimonian and Laird random effect meta-analysis was applied.³³⁶

3.9.1 Analysis of the hospital admission moderator

In an attempt to improve the relationship between atmospheric exposure and death, hospital admissions information identified if the subject was in the community during the exposure period leading to death. Hospital episode statistics report a subject's date of entry and exit to hospital. These allowed deaths to be classified by subject location during the 30 day exposure period into one of three groups i.e. 'zero days' spent in hospital, '1-29 days' spent in hospital, and 'all 30 days' spent in hospital. Zero days spent implies a community-based death, whereas all 30 days implies a hospital-based death. To compare the effect estimates observed in the three hospital admission groups against the first principles z-test was performed for each two-way combination; 1-29 days vs zero, all 30 days vs zero, and all 30 days vs 1-29 days.

3.9.2 Analysis of outliers and missing data

The majority of the analysis was performed under the 'complete cases' basis. In addition a set of analyses under two scenarios were applied and compared with the complete cases results. The missing exposure data were assumed to be 'missing at random' given the explanatory variables available and multiple imputation techniques already discussed were applied and the results compared with the main analysis. A final analysis looked at the influence of extreme outliers on effect estimates. Here, the results for adjustment for outliers were reported for 'complete cases only'. In this case extreme exposure outliers were removed that were defined as those values greater than the 99th percentile. A multiple imputation analysis would not be appropriate in this analysis of the outliers, because, as the missing values were due to the value itself, the missing values would now be considered 'missing not at random'.

3.9.3 Analysis procedure of the main data

Given the number of different dependent variables (3 causes of death), independent exposure variables (5 pollutants 1 temperature) and analysis scenarios (3 hospital admission status plus missing data and outlier investigation), modelling all combinations would not have been feasible both during the analysis stage and the write up stage. Prior to the analysis, a procedure was agreed that determined the order and analyses of the main dataset to be performed. Unless necessary for comparison purposes each analysis was performed with outcome defined by any cause of death field (see Chapter 3.5). To account for as much confounding as possible, temperature was modelled in a complete dataset and then the influence of hospital admission status during temperature exposure was investigated. Once temperature had been modelled each pollutant was added in a single pollutant model. Again the analysis was performed initially in the complete dataset followed by an investigation of the hospital admission status under a complete cases scenario. The missing data analysis and outlier investigation was then performed in the complete dataset and then in those with zero days in hospital during exposure only. Of the complete data and the zero days in hospital, the results associated then with zero days in hospital were identified as the main result as they represented the most likely direct exposure-response effect.

4 RESULTS – INTRODUCTION

The following chapters will describe in detail the results of the study under a traditional thesis format. The results will begin here with an introduction to the data outlining any relevant final choices when creating the datasets, particularly with respect to the exposure data. The results of the simulation study, including a description of the data used and the missing data characteristics applied, are presented in Chapter 6. Chapter 7 contains the investigation of the effect of the main confounder – temperature – and the bespoke models chosen for each cause of death. Chapters 5 and 8 will then report results associated with the pollution analysis, beginning in Chapter 5 with the outcome of an analysis investigating the influence of hospital admission during exposure on pneumonia mortality. This analysis was expanded in Chapter 8 to the three causes of death, pneumonia, chronic obstructive pulmonary disease (COPD), and ischaemic heart disease (IHD) and to evaluating the influence of outliers, missing data, and variability in the exposure data. The results chapter will also compare the results of the three causes of death to determine if any significant differences in the exposure effect are present. During the course of this study a paper arising from the work undertaken was published relating to the investigation into the use of hospital admission data to improve the relationship between black smoke air pollution and pneumonia mortality in a smaller cohort (1980 to 1996 in Edinburgh). Some of the results have been included here in Chapter 5, however the full paper can be found in Appendix A.

4.1 Data background - Summary of mortality data

The total number of deaths observed between January 1980 and December 2011 in Aberdeen (15.7%), Edinburgh (28.8%), Glasgow (48.3%) and Inverness (7.2%) was 919,301. The variance was typically greater (by 5-10) than the mean of the daily number of deaths for just under half of the exposure monitors indicating slight monitor dependent over-dispersion in the outcome. The study cohort contained 52.2% female and 78.7% aged over 65. Table 4.1 describes the study population split by the three causes of death to be investigated pneumonia, COPD, and IHD.

Table 4.1 – Number of deaths from Pneumonia, Chronic Obstructive Pulmonary Disease (COPD), and Ishaemic Heart Disease (IHD), split by location, age, gender, and prior number of exposure days in hospital

No. (%) Deaths	Pneumonia		COPD		IHD	
	Primary	Any Cause	Primary	Any Cause	Primary	Any Cause
Location						
Aberdeen	8375(15.2)	34057(18.5)	6112(13.9)	12470(14.2)	34659(15.9)	44689(16)
Edinburgh	15070(27.3)	51801(28.1)	12240(27.8)	24980(28.5)	61758(28.3)	79651(28.5)
Glasgow	28254(51.2)	85508(46.4)	23085(52.5)	45058(51.4)	106428(48.8)	135400(48.5)
Inverness	3498(6.3)	13031(7.1)	2540(5.8)	5176(5.9)	15384(7)	19361(6.9)
No. Exposure Days in Hospital^a						
Zero Days	35948(65.1)	106500(57.8)	21080(47.9)	45612(52)	157363(72.1)	186696(66.9)
1-29 Days	13621(24.7)	55210(29.9)	18534(42.1)	34108(38.9)	51217(23.5)	75732(27.1)
All 30 Days	5628(10.2)	22687(12.3)	4363(9.9)	7964(9.1)	9649(4.4)	16673(6)
Age at death						
<40	567(1)	2105(1.1)	106(0.2)	196(0.2)	1105(0.5)	1294(0.5)
40-	3689(6.7)	18162(9.8)	6413(14.6)	12567(14.3)	42181(19.3)	49057(17.6)
65-	15065(27.3)	63766(34.6)	22811(51.9)	46420(52.9)	98752(45.3)	124944(44.8)
80-	35876(65)	100364(54.4)	14647(33.3)	28501(32.5)	76191(34.9)	103806(37.2)
Gender						
Male	24409(44.2)	85390(46.3)	21887(49.8)	44745(51)	110935(50.8)	140968(50.5)
Female	30788(55.8)	99004(53.7)	22090(50.2)	42938(49)	107293(49.2)	138132(49.5)
Total No. Deaths						
Case only	55,197	184,397	43,977	87,684	218,229	279,101
Case & Control	242,861	811,051	193,273	385,645	960,077	1,227,755

^a 'Primary' refers to the primary cause of death field whereas Any Cause relates to any (primary or secondary) cause of death field on the death certificate. ^a No. days in hospital – Number days spent in hospital during the month leading to death.

The data were reported as those identified in the death certificate as having the cause of death located in the 'primary' cause of death field only, and in 'any' cause of death field (i.e. primary or any secondary field). Along with splitting by age & gender, the data were split by the number of days spent in hospital during the month leading to death. In each category the total number of 'case' days (i.e. the total number of deaths) and the 'case and control' days are reported, with 919,301 total deaths the number of 'case and control' days was 4,043,932. To confirm the observed number of deaths for each cause of death was correctly obtained (Table 4.2), an expected number of deaths were calculated using the cause specific ISD Scotland death rates for 1991-2011, adjusting for the study population and study time period. The gradual downward trend in deaths meant the observed number was greater due to the expected number being based on the available later years of 1991 to 2011. The equivalent table for the entire population as reported by ISD Scotland can be found in Appendix C under Table C1.

Table 4.2 – Number of cause specific deaths observed in the data (1980-2011) compared with expected number of deaths based on ISD reported years (1991-2011).

Primary COD field Year	Pneumonia		COPD		IHD		All causes	
	N	Row%	N	Row%	N	Row%	N	Col %
1980	2,082	6.9	1,294	4.3	8,782	29.2	30,129	3.28
1981	1,900	6.2	1,246	4.1	9,042	29.6	30,542	3.32
1982	2,253	7.3	1,339	4.4	8,928	29.1	30,716	3.34
1983	2,064	6.8	1,277	4.2	8,871	29.1	30,450	3.31
1984	2,013	6.8	1,295	4.4	8,812	29.7	29,682	3.23
1985	1,945	6.4	1,318	4.3	9,017	29.6	30,417	3.31
1986	2,110	7.0	1,358	4.5	8,660	28.8	30,088	3.27
1987	1,935	6.6	1,236	4.2	8,786	29.9	29,383	3.20
1988	1,874	6.4	1,264	4.3	8,564	29.0	29,507	3.21
1989	2,379	7.7	1,454	4.7	8,768	28.2	31,055	3.38
1990	1,931	6.7	1,210	4.2	7,962	27.8	28,666	3.12
1991 ^a	1,868	6.5	1,222	4.2	8,021	27.9	28,798	3.13
1992	1,850	6.4	1,212	4.2	7,797	27.0	28,844	3.14
1993	2,261	7.5	1,412	4.7	8,108	26.7	30,366	3.30
1994	1,890	6.8	1,196	4.3	7,220	25.9	27,911	3.04
1995	2,062	7.1	1,345	4.6	7,271	25.0	29,033	3.16
1996	2,087	7.4	1,227	4.4	6,896	24.6	28,088	3.06
1997	1,990	7.2	1,284	4.7	6,522	23.7	27,505	2.99
1998	1,949	7.1	1,341	4.9	6,414	23.3	27,516	2.99
1999 ^b	2,342	8.4	1,537	5.5	6,248	22.3	27,979	3.04
2000	1,200	4.2	1,496	5.2	6,025	20.9	28,775	3.13
2001	1,109	3.9	1,486	5.2	5,580	19.7	28,377	3.09
2002	1,272	4.4	1,526	5.3	5,569	19.4	28,721	3.12
2003	1,415	4.9	1,610	5.6	5,324	18.5	28,828	3.14
2004	1,153	4.2	1,430	5.2	5,011	18.1	27,656	3.01
2005	1,231	4.5	1,516	5.5	4,922	17.8	27,627	3.01
2006	1,284	4.7	1,487	5.4	4,654	16.9	27,498	2.99
2007	1,205	4.4	1,572	5.7	4,433	16.1	27,622	3.00
2008	1,243	4.5	1,496	5.5	4,243	15.4	27,474	2.99
2009	1,176	4.4	1,479	5.6	3,983	15.0	26,522	2.89
2010	1,124	4.2	1,369	5.1	4,015	15.1	26,648	2.90
2011	990	3.7	1,440	5.4	3,729	13.9	26,750	2.91
Total (80-11)	55,197	6.0	43,977	4.8	218,229	23.7	919,301	100
Total (91-11)	32,701	5.6	29,683	5.0	121,985	20.7	588,538	
Exp Total (80-11) ^c	49,445		46,019 ^d		193,530			
Exp Total (91-11) ^c	31,655		29,462 ^d		123,898			

^a – ISD reported death rates start in 1991. ^b – ICD10 replace ICD9 cause of death coding. ^c – Expected number of deaths based on ISD death rates and adjusted for the average ratio between study cite population and total Scottish population. ^d – ISD report not COPD specific rather all-encompassing Chronic lower respiratory diseases.

4.2 Summary of pollution exposure data

During 2011 eighty-eight monitors were running in Scotland, with eight new sites activated and five closed down since 2010.³³⁷ Tables C2a – C2c of Appendix C give background information for any monitor active during the 30 year study period for the five largest urban areas containing pollution monitoring; Aberdeen (Current Popⁿ=222,751), Dundee (147,197), Edinburgh (477,133), Glasgow (593,101), and Inverness (59,660).³²⁵ The number of active monitors has varied considerably since 1979, for example with Edinburgh hosting thirteen sites over the 30 year period nine of which ran concurrently until seven were deactivated in March 1982. Of the five urban areas Dundee was dropped due to a lack of sites with no AURN monitors present and all Black Carbon network decommissioned by 1982. Inverness contains no BS or SO₂ monitoring, however the location was kept to increase the sample size relating to PM₁₀ and NO₂. Table 4.3 gives summary statistics for the pollutants (BS, PM₁₀, PM_{2.5}, SO₂, and NO₂) in monitors chosen to be part of the study based on the protocol outlined in Chapter 3.4. Note, the weighted average interquartile range each pollutant respectively is 10.0, 11, 6, 17.2, and 21. Figure 4.1 gives a graphical representation of the start and finish date associated with monitors active for black smoke air pollution during the study. Note, three BS and SO₂ sites running concurrently in Glasgow were identified as suitable and kept in the data. For these three sites subject deaths in Glasgow were matched to the nearest monitor estimates based on distance between postcodes.

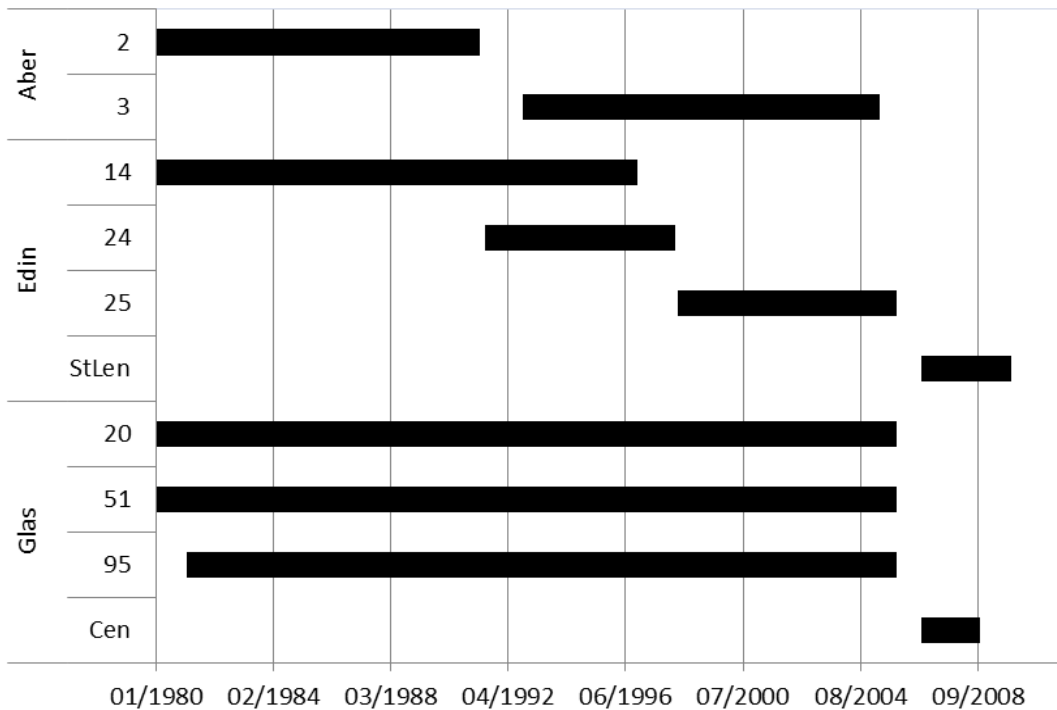
The period between 01/01/1979 and 31/12/2011 corresponds to 11,719 potential days with pollution data. Of the pollutants BS and SO₂ represent the largest available data source as monitoring networks began prior to January 1980 and continue until the late 2000s. The remaining pollutants became important monitoring targets in the 1990's and so the total number of available measurements was fewer and depended on the active period for each monitor type.

Table 4.3 - Summary statistics for pollution monitors used within the study

Poll (μgm^{-3})	Mon	Start (m/yr)	End (m/yr)	No. Exp ^a	Miss (%)	Med	IQR	Min	Max
BS	A2	Dec-79	May-91	4,197	10.9	7	9	1	322
	A3	Nov-92	May-05	4,859	27.0	3	4	1	98
	ED14	Dec-79	Nov-96	6,186	6.8	9	11	1	194
	ED24	Jul-91	Mar-98	497	5.4	3	4	1	51
	ED25	Apr-98	Dec-05	2,811	42.4	10	8	1	56
	EDSt Len	Oct-06	Dec-09	1,171	13.9	4	5	0.1	40
	G20	Dec-79	Dec-05	9,528	17.6	14	14	1	541
	G51	Dec-79	Dec-05	9,528	10.8	6	10	1	530
	G95	Feb-81	Dec-05	9,077	7.5	7	11	1	535
	GCen	Oct-06	Nov-08	788	16.9	4	3	0.1	45
PM10	ACen	Sep-99	Jun-12	4,669	7.3	15	10	1	94
	EDCen	Oct-92	Oct-03	4,014	11.0	22	12	5	102
	EDSt Len	Dec-03	Jan-12	3,102	10.9	16	8	3	154
	GCen	Jul-96	Jan-12	12,731	8.8	20	12	2	124
	INCen	May-07	Dec-07	4,002	22.3	13	10	3	89
PM2.5	ACen	Mar-09	Jun-12	1,217	16.7	6	6	0	47
	EDSt Len	Nov-08	Jan-12	1,324	4.1	9	5	2	54
	GCen	Dec-08	Jan-12	1,312	2.7	8	7	1	117
SO2	A2	Dec-79	May-91	4,197	11.5	27	19	0.1	151
	A3	Nov-92	May-05	4,712	22.8	12	13	0.1	102
	ACen	Jan-01	Sep-07	1,014	1.7	3	3	0.1	20
	ED14	Dec-79	Nov-96	6,186	8.2	37	24	0.1	422
	EDCen	Oct-92	Oct-03	2,531	9.2	5	5	0.1	65
	EDSt Len	Nov-03	Jan-12	3,128	3.0	3	3	0.1	63
	G20	Dec-79	Dec-05	9,528	17.6	32	25	0.1	355
	G51	Dec-79	Dec-05	9,516	10.3	26	17	0.1	448
	G95	Feb-81	Dec-05	9,077	7.1	26	20	0.1	308
	GCen	Jul-96	Jan-12	2,366	3.8	3	2	0.1	52
NO2	ACen	Sep-99	Jun-12	4,669	6.7	23	18	1	86
	EDCen	Oct-92	Oct-03	4,026	8.6	46	19	11	138
	ED St Len	Nov-03	Jan-12	3,128	3.4	24	16	0	155
	GCity	Jan-87	Mar-11	467	1.9	31	16	7	80
	GCen	Jul-96	Jan-12	22,705	4.0	49	24	8	230
	INCen	Jul-01	Jun-12	4,002	3.5	20	14	2	75

a. Number of expected days containing pollution measurements based on monitor start and finish dates
Ab = Aberdeen, Ed=Edinburgh G=Glasgow In = Inverness

Figure 4.1 - Graphical representation of the periods when each black smoke monitors employed in the study are active.



Pairwise correlation coefficients are reported in Table 4.4 for the five pollutants used in this study. The two particulate measurements PM10 and PM2.5 were strongly positively correlated at 0.83, with BS and SO2 similarly highly correlated at 0.57. The remaining comparisons were all moderately correlated, with a minimum correlation of 0.24 between SO2 and PM10. Correlations associated with year round temperature, summer only and winter only temperature have also been added indicating a gradual decrease in pollution as temperature increases, particularly in the winter months.

Table 4.4 – Pairwise correlation coefficients comparing the five pollutants and average temperature modelled in the study.

	Black Smoke	PM10	PM2.5	Sulphur Dioxide	Nitrogen Dioxide	Temperature (7am-11pm)
Black Smoke	1					
PM10	0.384	1				
PM2.5	0.475	0.827	1			
Sulphur Dioxide	0.570	0.242	0.430	1		
Nitrogen Dioxide	0.495	0.509	0.497	0.408	1	
Temperature (7am-11pm)	-0.298	-0.040	-0.249	-0.144	-0.280	1
Temperature (7am-11pm) - Winter	-0.295	-0.126	-0.206	-0.223	-0.281	1
Temperature (7am-11pm) - Summer	-0.021	0.126	0.046	-0.009	-0.015	1

4.3 Summary of temperature exposure data

The presence of clustering within pollution monitor needed to be accounted for in the main analysis. In order to reduce the complexity, a single temperature monitor with minimal missing data were intended for each city. In addition to inconsistent activity periods, the meteorological monitoring sites may also record measurements at hourly intervals during the day or at one single time point (usually 9am). To generate accurate daily temperature readings at least 15 hrs worth of measurements recorded per day were required. Centrally located monitors with measurements taken hourly were available for Aberdeen (ID=161) and Edinburgh (ID=246). Glasgow and Inverness each contained three monitors located within 10 miles of city centre that were active for some but not all of the period (see Table 4.5), or were inconsistent regarding the time of day the measurements were recorded (see Appendix C Table C3). The closest monitoring sites to Glasgow and Inverness containing complete data were found at Prestwick Airport (ID=1006) and Aviemore Park (ID=113), respectively. As Prestwick and Aviemore sites are 30 and 24 miles from Glasgow and Inverness, temperature readings were compared (Table 4.5) with those from the city centre at the commonly recorded time of 9am (Table C3 of Appendix C). A t-test comparison compared each city monitor with the surrogate monitor, suggesting the Prestwick monitor was on average 0.25°C to 1.12°C warmer than Glasgow city centre, and Aviemore park was between 0.8°C and 1.1°C colder than Inverness city centre.

Table 4.5 – Comparison statistics for available temperature monitors in Glasgow and Inverness based on 9am measurements.

Monitor	Start	End	N ^a	Mean	SD	Min	Max	Mean Diff ^b	P-val ^c
Glasgow 1006	01/12/1979	31/12/2011	10585	9.4	5.3	-11.6	24.0	-	-
Glasgow 972	01/03/1991	31/12/2011	6779	9.4	5.6	-18.5	24.1	0.25	<0.001
Glasgow 977	01/12/1979	31/08/1990	3656	8.2	5.3	-14.4	24.5	0.70	<0.001
Glasgow 978	01/12/1979	30/09/2005	9388	8.2	5.2	-23.4	24.0	1.12	<0.001
Inverness 113	01/07/1982	31/12/2011	10750	7.3	5.9	-19.9	24.7	-	-
Inverness 110	01/12/1979	31/12/2011	11361	8.0	5.5	-11.1	26.0	-0.80	<0.001
Inverness 115	01/12/1979	31/12/2011	11524	8.4	4.8	-9.5	23.2	-1.12	<0.001
Inverness 116	01/12/1979	31/12/2011	11466	8.4	5.4	-11.0	25.0	-1.06	<0.001

a. Number observations taken at 9 am

b. Mean of the difference in the 9am daily measurements

c. T-test comparison of 9am measurements between monitors (reference monitors 1006/113)

Summary statistics of the final chosen monitors linked to each city are reported in Table 4.6. Temperature statistics here refer to the daily average, minimum, and maximum, and percentage missing as observed between 7am and 11pm, the likely active period increasing subject exposure.

Table 4.6 – Daily (7am – 11pm) summary statistics for the temperature monitors used in this study within each city.

Monitor Temperature Statistic (7am-11pm) ^{a.}	Summary statistics across the study period				
	Mean	SD	Min	Max	Percent Missing ^{b.}
Aberdeen 161 (Dec 1979 - Dec 2011)					
Daily Average	9	4.9	-13.3	24.5	0.22%
Daily Minimum	6.2	4.8	-18.8	19.8	
Daily Maximum	11.4	5.3	-9.5	29.1	
Edinburgh 246 (Dec 1979 - Dec 2011)					
Daily Average	9.6	5.1	-12.7	24.5	0.29%
Daily Minimum	6.3	5.1	-16.8	19	
Daily Maximum	12.1	5.4	-8	30	
Glasgow 1006 (Dec 1979 - Dec 2011)					
Daily Average	10	4.9	-8.4	25.9	0.30%
Daily Minimum	7	5	-13.1	21	
Daily Maximum	12	5.1	-4	39	
Inverness 113 (Jul 2001 - Dec 2011)^{c.}					
Daily Average	8.1	5.6	-16.9	24.5	0.65%
Daily Minimum	4.7	5.3	-23.8	19.9	
Daily Maximum	10.8	6.1	-12.9	30.1	

a. Temperature measured each hour, summary statistics relate to the active hours of 7am to 11pm.
b. Missing $\geq 25\%$ of hours measured between 7am and 11pm
c. Full study period not required as pollution monitoring does not start until 2001 in Inverness

5 RESULTS – INITIAL PNEUMONIA INVESTIGATION

Prior to acquiring the complete dataset, an initial investigation of a truncated mortality dataset was performed. This analysis explored the association between black smoke and pneumonia mortality, based on the ‘any’ cause of death field (AFCOD), using a monitor from Edinburgh between January 1981 and March 1996. There were three aims of this analysis:

- To explore some of the techniques required to complete the main study.
- Explore how black smoke (BS) exposure influences mortality from pneumonia across a 30 day lag period,
- Determine if a greater association occurs in subjects who spent the exposure period in the ‘community’ (i.e. not in hospital) compared to those who spent some or all of the period in hospital.

The complete results generated here have been published in a peer review journal,³³⁸ which can be found in Appendix A. The results related to the rest of this study are reported here.

5.1 Introduction and differences in the methods

Daily average ‘black smoke’ air pollution data were obtained from a single centrally located background monitoring site in Edinburgh (Ed Mon 14 in Table 4.3) and hourly ambient air temperature (between 7am-11pm). Black smoke daily results for the month prior were modelled in the lag stratified model with firstly an average 1-30 day lag and then separately for 1-6, 7-12, 13-18, 19-24, and 25-30 days. Unlike the three temperature zones described in Chapter 7, temperature was previously shown to have here a double linear relationship with mortality that has a single change in risk (knot) occurring at 11°C. Two continuous temperature variables were created to represent these ‘high’ and ‘low’ temperature ranges:

High = $t-11$ if $t \geq 11^\circ\text{C}$ and 0 otherwise.

Low = $t-11$ if $t < 11^\circ\text{C}$ and 0 otherwise.

where t is the daytime mean temperature. Average temperature across lags 1–30 days for both “low” and “high” variables was included in all models as covariates. The ‘high’ and ‘low’ continuous temperature variables were expressed as a percentage increase in risk corresponding to a *decrease in 1°C of temperature* for any individual day during lag period.

A death was considered a community death from pneumonia (CDP) if the hospital admission data showed the subject had not been in hospital for any of the 30 days prior to death. A CDP is a special subgroup of the clinical term community acquired pneumonia (CAP) deaths, with the latter being based on based on the source of pneumonia. The association between BS and All pneumonia deaths (AP) was estimated before restricting to the subgroup (CDP) thought to have a greater potential exposure. Comparatively the CDP and non-CDP groups relate to zero, and 1-30 days in hospital. To test if a significant difference in exposure effects occurred, an interaction term was included to compare CDP and non-CDP for each BS lag term, and the Log-likelihood-ratio was used to test the difference. In addition to lag-stratified, a quadratic distributed lag estimated the lag time where the estimated effects were positive.

5.2 Results associated with Pneumonia mortality analysis

Missing pollution data for either a subjects’ case day or all of their control days resulted in 14,346 subjects eligible in a complete cases analysis here. Table 5.1 gives descriptive statistics for the daily average BS air pollution, air temperature and back ground demographic characteristics such as age and gender for both AP and CDP (52.5% of the AP) subject groups. Temperature ranged from -13°C to 25°C with a mean (and median) temperature of 9.4°C. Daily average black smoke pollution indicated a maximum of 194 μgm^{-3} , median 9 μgm^{-3} and interquartile range of 10 μgm^{-3} . Note, the summary statistics for each lag period also indicated a interquartile range of approximately 10 μgm^{-3} in each case.

Table 5.1 - Descriptive statistics of exposure and subjects for black smoke (BS) air pollution within Edinburgh between Jan 1981 and March 1996.

	Mean	S.D	Median	IQR	Min	Max
Daily Ave Air Temp(^o C)	9.4	5.1	9.4	8	-12.7	24.48
Daily Ave BS(μgm^{-3})	12.7	13.3	9	10	1	194
Lag 1-6 dys Ave BS(μgm^{-3})	12.9	10.8	9.3	9.5	1	95.2
Lag 7-12 dys Ave BS(μgm^{-3})	13	11.1	9.5	9.6	1	95.2
Lag 13-18 dys Ave BS(μgm^{-3})	13.1	11.3	9.5	9.7	1	95.2
Lag 19-24 dys Ave BS(μgm^{-3})	13.1	11.2	9.5	9.5	1	95.2
Lag 25-30 dys Ave BS(μgm^{-3})	13	11.2	9.5	9.3	1	95.2
Lag 1-30 dys Ave BS(μgm^{-3})	13.2	9.2	10.1	9.3	2.7	73.2
Age (CDP)	79.13	12.6	82	12	0	108
Age (Non-CDP)	79.21	11.92	81	13	0	108
Age (AP)	79.15	12.2	81	13	0	108

Categories	Gender		Age Grouped		Total
	Male	Female	<80	\geq 80	
CDP only Subjects (%)	3409(45.2)	4127(54.8)	3064(40.7)	4,472(59.3)	7536(52.5)
Non-CDP only Subjects (%)	3166(48.2)	3644(46.9)	3109(50.4)	3701(45.3)	6810(47.5)
All Pneumonia Subjects (%)	6575(45.8)	7771(54.2)	6173(43.0)	8,173(57.0)	14346(100)

BS = Black Smoke,
CDP = Community deaths from pneumonia
AP = All Deaths from pneumonia

The percentage change in relative risk (%RR), with 95% confidence intervals, are given in Table 5.2 for an $10 \mu\text{gm}^{-3}$ increase in BS on any individual day or a *temperature decrease* of 1°C . MODEL 1 refers to the model with BS averaged over a 1-30 days (lag 1-30) and MODEL 2 reports the 30 day lagged BS split into the five smaller lag periods. Concerns of collinearity due to strong correlation coefficients of around 0.7 for the five adjacent BS lag periods in model 2 were reduced by calculating the variance inflation factors ($\text{VIF} = 2.01\text{-}2.78$) indicating a very limited amount of correlation between predictors.³³⁹ The differences in %RR between AP and CDP along with corresponding significance levels are also given.

MODEL 1 reports the effects of exposure on any of the 30 days to be equal. An increase of $10\mu\text{gm}^{-3}$ black smoke pollution on any of the 30 days showed a small rise in AP percentage relative risk of 0.08% which increased to 0.19% in the CDP group, resulting in a %RR difference of -0.18% between CDP and non-CDP subjects (CDP-non-CDP %RR). When the 30 days was split into 5 lag periods (MODEL 2), the magnitude of the

effect (regardless of direction) was always larger in the CDP subjects compared to the AP. The largest changes in %RR between CDP and AP are seen in the 1-6, 7-12, and 13-18 day lags. This 18 day period prior to death appeared to be the high risk period, as an increase %RR was observed in 1-6, 7-12, and 13-18 day lags whereas a decrease was observed in the 19-24 and 25-30 day lags. The non-CDP subjects, report a decrease change in risk as BS increase for the shorter lags 1-6, and 7-12 days, before a delayed increase in risk.

Regardless of MODEL 1 or 2, low temperature effects changed very little a 1°C decrease corresponded to an increase in relative risk approximately the same for both AP (MODEL 1 = 0.20%), CDP (MODEL 1 = 0.22%), and non-CDP (MODEL 1 = 0.16%). In comparison, a 1°C decrease in high temperature shows a small decrease in risk in AP (MODEL 1=-0.05%) that increases in magnitude in the CDP (MODEL 1 = -0.20%) group. Non-CDP i.e. hospital based subjects appeared to show an increase in risk as temperature decreases towards 11°C i.e. warmer temperatures relate to a decrease in risk in those based in hospital.

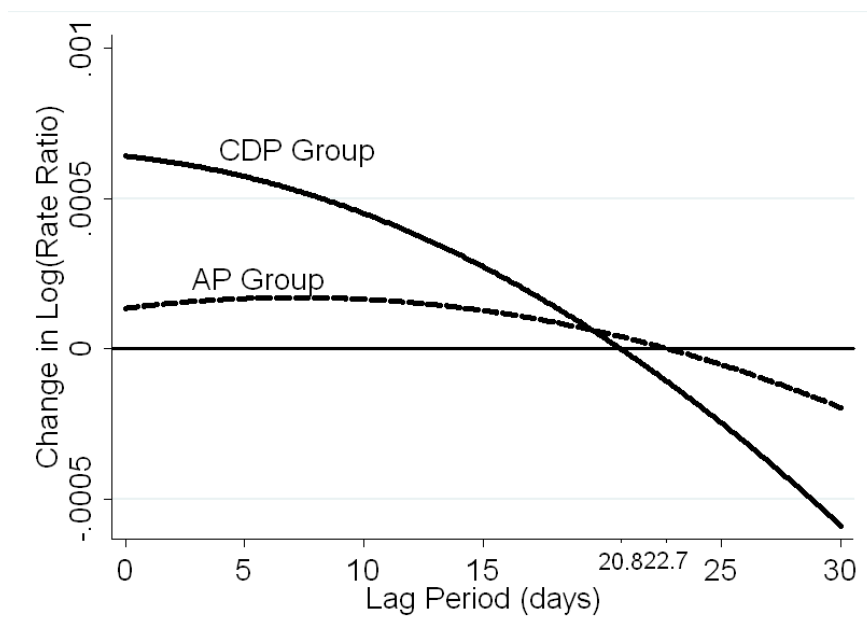
Table 5.2 – Results of the lag stratified black smoke air pollution percent change in risk for Pneumonia mortality, for all subjects, and split by community based only subjects and non-community based subjects only.

	AP		CDP only		Non - CDP only		CDP - AP Diff	CDP-Non CDP Diff	P-val	
	Lag (days)	% RR Change	95% C.I	% RR Change	95% C.I	% RR Change				95% C.I
MODEL 1 Black Smoke	1-30	0.08%	-0.17%,0.35%	0.19%	-0.16%,0.58%	0.01%	-0.34%,0.40%	0.11%	0.18%	0.285
Air Temp "Low"	1-30	0.20%	0.11%,0.29%	0.22%	0.09%,0.35%	0.16%	0.03%,0.30%	0.02%	0.06%	0.502
Air Temp "High"	1-30	-0.05%	-0.20%,0.10%	-0.20%	-0.40%,0.00%	0.11%	-0.11%,0.34%	0.15%	0.31%	0.059
MODEL 2 Black Smoke	1-6	0.12%	-0.37%,0.62%	0.56%	-0.14%,1.29%	-0.31%	-0.99%,0.41%	0.44%	0.87%	0.022
	7-12	0.05%	-0.42%,0.53%	0.32%	-0.33%,1.00%	-0.25%	-0.92%,0.45%	0.28%	0.57%	0.023
	13-18	0.40%	-0.08%,0.90%	0.71%	0.03%,1.42%	0.14%	-0.56%,0.86%	0.31%	0.57%	0.163
	19-24	-0.09%	-0.55%,0.38%	-0.16%	-0.79%,0.50%	0.05%	-0.63%,0.75%	0.06%	0.21%	0.272
	25-30	-0.11%	-0.57%,0.36%	-0.39%	-1.01%,0.25%	0.34%	-0.35%,1.06%	0.28%	0.73%	0.008
Air Temp "Low"	1-30	0.19%	0.10%,0.29%	0.19%	0.06%,0.32%	0.18%	0.05%,0.32%	0.00%	0.01%	0.487
Air Temp "High"	1-30	-0.05%	-0.19%,0.10%	-0.20%	-0.39%,0.01%	0.09%	-0.12%,0.33%	0.15%	0.29%	0.058

%RR Change - percentage change in Relative Risk, associated with an increase of 10µgm⁻³ BS or a decrease of 1°C, on any individual day within the lag period, with corresponding 95% Confidence Interval (95% C.I.)
Model 1 - One 30 day lag, Model 2 - The 30 days split into 5 lags of 6 days each fitted simultaneously
|CDP| – |AP| Diff - The difference in the magnitude of the effect size between AP and CDP (CDP-AP)
|CDP-Non CDP| Diff - The difference in the effect size between CDP and Non-CDP (CDP – Non CDP)

Table 5.2 indicated a larger effect occurring in the CDP group than the AP group that lasted approximately 18-24 days. Figure 5.1 plots the results of the quadratic lag distribution model, reporting the change in log rate ratio associated with the 30 day lag period for both AP and CDP. Results are comparable with Table 5.2, showing increased risk in CDP subjects occurring immediately before a decline that crossed zero at approximately 21 days, almost 2 days earlier than the more gradual AP decline in risk.

Figure 5.1 – Plot of the quadratic distributed lag model for all pneumonia (AP) and community deaths from pneumonia (CDP).



6 RESULTS - SIMULATION STUDY

The following chapter reports the results associated with the simulation study investigating the influence of missing exposure data in a standard ‘complete cases’ analysis and the suitability of the missing data technique ‘multiple imputation’.

6.1 Introduction to the simulation datasets and the true effect estimates

Two empirical datasets based on observational measurements and events were employed in this simulation study. The two datasets related to Edinburgh, Scotland.

1 - Black smoke pollution measurements from a single monitor located within Edinburgh, identified as the Edinburgh monitor 14, were employed as the exposure data. Monitor 14 was active for a total of 6180 days between 01/12/1979 and 31/10/1996 and contained in total 6.8% days with missing data. Of all the pollution exposure dataset available it contained the longest continuous time-period with no missing data. The period contained 1359 days from 07/07/1990 until 26/03/1994 with no missing values. In addition complete temperature data in the form of daily 7am to 11pm average was also available for the corresponding time period.

2 – A matching mortality dataset containing all non-accidental deaths was created for the same time period defined by the complete exposure dataset. This corresponding mortality dataset contained 31,694 deaths (‘case’ days) which in the time-stratified case-crossover design resulted in a total of 106,836 control days.

The complete observed exposure and temperature dataset and the mortality data were matched based on the date associated with the case or control days, and a conditional logistic regression model fitted. The model parameters were simple linear variables representing same day (lag 0) black smoke pollution and daily (7am to 11pm) average temperature. No attempt was made here to account for either a change in shape across the exposure range or a lagged relationship. From this model, the true effect estimates (S.E.) were calculated for BS pollution to be 0.0003302 (0.0007093) and daily average temperature to be -0.0012526 (0.0024583). This would correspond to a percentage

change in relative risk (%RR (95% C.I.)) of 0.33% (-1.05%,1.74%) per $10\mu\text{gm}^{-3}$ increase of black smoke pollution and -0.13% (-0.61%, 0.36%) per 1°C increase in 7am-11pm daily average temperature.

6.2 Simulating the missing data characteristics

Missing ‘black smoke’ pollution data were firstly simulated in the form of MCAR. Table 6.1 describes the total number (%) of missing data occurring in pollutant monitors included in the full study. With the exception of Edinburgh monitor 25 (42.4%) all monitors contained 27% or less missing data. To apply the MCAR characteristic, a pseudo-randomly selected pollution estimate was replaced with a missing value sequentially until 5% (n = 68 out of 1359), 10% (n = 136), 15% (n = 204), 20% (n = 272), and 25% (n = 340) of the dates were set as missing observations in five distinct datasets containing the artificially created missing data.

Table 6.1 - Description of the number (%) missing data for each pollutant and monitor in the study

Monitor Pollution	No. (%) Missing				
	BS	PM10	PM2.5	SO2	NO2
Ab 2	474(10.5)	-	-	513(11.3)	-
Ab 3	1313(27)	-	-	1073(22.8)	-
Ab Centre	-	340(7.3)	203(16.7)	17(1.7)	313(6.7)
Ed 14	418(6.4)	-	-	509(7.8)	-
Ed 24	27(5.4)	-	-	-	-
Ed 25	1193(42.4)	-	-	-	-
Ed Centre	-	441(11)	-	234(9.2)	346(8.6)
Ed St Leon	163(13.9)	337(10.9)	54(4.1)	94(3)	106(3.4)
Glas 20	1674(17)	-	-	1680(17)	-
Glas 51	1031(10.5)	-	-	985(10)	-
Glas 95	679(7.5)	-	-	648(7.1)	-
Glas Cen	133(16.9)	1114(8.8)	35(2.7)	90(3.8)	9(1.9)
Glas City	-	-	-	-	919(4)
Total	7105(14.2)	2232(9.1)	292(7.6)	5843(10.9)	1693(4.8)

- indicates exposure observations not available/suitable here

Once the MCAR scenarios had been applied, datasets were created that contained missing data under controlled characteristics (MAR). These replicated true missing data

distributions observed in pollution exposure data from the main study. The time-series nature of the exposure dataset meant that the observable MAR characteristics here related to time dependent factors such as long term time changes, seasonal, monthly, or day of the week differences.

Pollution measurement techniques changed considerably over the 30 years study period with the biggest differences being the switch from manual monitoring to automatic monitoring. It was thought that human error in collecting measurements during the earlier years of the study, or delays in fixing a broken automatic monitor would result in long term time-trends. In fact long-term time trends patterns only existed for certain pollutants when all data were combined, but did not exist within an individual monitor. As the simulation study (and the multiple imputation in the main analysis of the study) was performed within the monitor this was not investigated. It was also proposed, that day of the week or weekend specific characteristics may be present in the missing data patterns. In both cases no pattern was observed i.e. missing data appeared to be equally likely to occur on any day of the week.

Table 6.2 – Number (%) missing days for each pollution and monitor split by season (Winter/Summer)

Monitor	No. (season row %) Missing									
	BS		PM10		PM2.5		SO2		NO2	
	Win	Sum	Win	Sum	Win	Sum	Win	Sum	Win	Sum
Ab 2	146(31)	328(69)	-	-	-	-	155(30)	358(70)	-	-
Ab 3	647(49)	666(51)	-	-	-	-	460(43)	613(57)	-	-
Ab Cen	-	-	185(54)	155(46)	123(61)	80(39)	5(29)	12(71)	152(49)	161(51)
Ed 14	104(25)	314(75)	-	-	-	-	162(32)	347(68)	-	-
Ed 24	27(100)	0(0)	-	-	-	-	-	-	-	-
Ed 25	631(53)	562(47)	-	-	-	-	-	-	-	-
Ed Cen	-	-	284(64)	157(36)	-	-	172(74)	62(27)	221(64)	125(36)
Ed St Len	35(21)	128(79)	157(47)	180(53)	24(45)	30(56)	57(61)	37(39)	71(67)	35(33)
Glas 20	809(48)	865(52)	-	-	-	-	804(48)	876(52)	-	-
Glas 51	507(49)	524(51)	-	-	-	-	497(51)	488(49)	-	-
Glas 95	316(47)	363(54)	-	-	-	-	315(49)	333(51)	-	-
Glas Cen	88(66)	45(34)	405(36)	709(64)	13(37)	22(63)	49(54)	41(46)	5(56)	4(44)
Glas City	-	-	-	-	-	-	-	-	300(33)	619(67)
Total	3310 (47)	3795 (53)	1031 (46)	1201 (54)	160 (55)	132 (45)	2676 (46)	3167 (54)	749 (44)	944 (56)

- = indicates exposure observations not available/suitable here

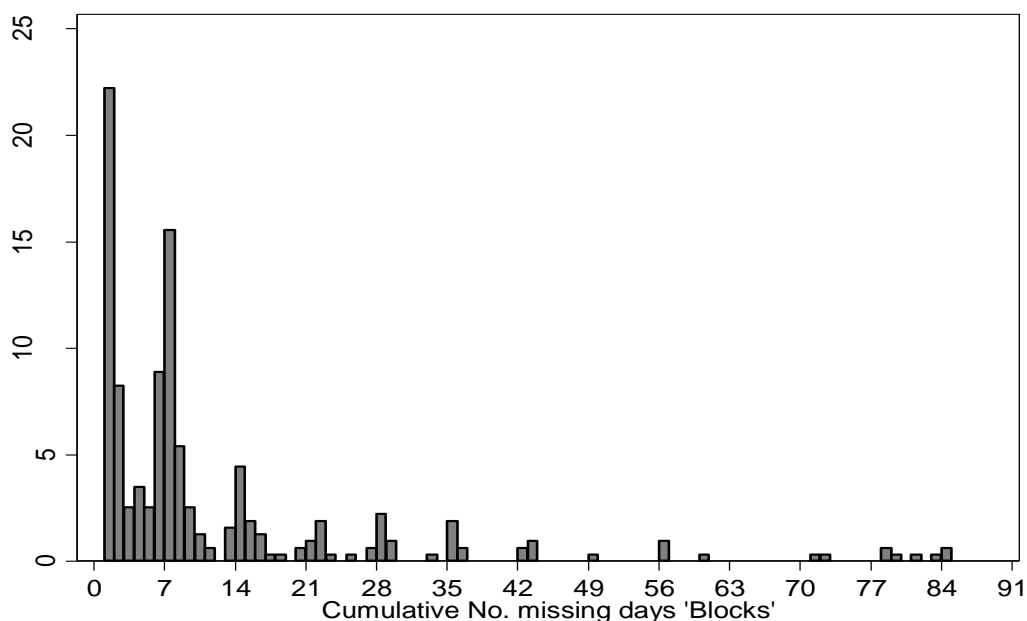
Sum = Summer months April to September, Win = Winter months October to March

Seasonal differences were present. Table 6.2 describes for each pollution monitor the percentage of missing data that occurred in both winter and summer, where summer was defined as April to September. Interestingly, BS and SO₂ appeared to report greater missing data in summer than winter, whereas PM₁₀, PM_{2.5}, and NO₂ were more likely to contain missing data in the winter months. This may be a reflection of manual vs automatic monitoring. Manual monitors might be susceptible to missed measurements or delays in maintenance during the summer holiday. Automatic monitors alternatively might succumb to bad weather such as extreme cold temperatures, whereas manual monitors are more robust, and results can be improvised by the attendee making the measurement. Irrespective of direction and with the exception of one monitor (Edinburgh monitor 24) the maximum ratio of missing data observed was 21:79 (Win:Sum) with the majority less than 25:75. Seasonal MAR characteristics in the form of 50:50, 40:60, and 25:75 (Win:Sum) ratios were therefore applied to the data.

As with observed pollution measurements, it was hypothesised that the probability of the presence of missing pollution measurement was also likely to contain autocorrelation i.e. a day following a missing observation had a greater probability of also containing a missing observation. These periods of cumulative days with missing observations were designated as missing 'blocks'. It was hypothesised that most blocks were due to delays in maintenance and so the majority would be fixed within a few days most likely the same day, or would take on average 7 to 14 days. This may follow a Poisson distribution with a mean around 7 days. Figure 6.1 illustrates the distributional properties when all black smoke monitors were combined (note truncated to 90 days). Three characteristics stand out: a high day zero percentage, long tails, and the appearance of a seven day cycle block length. As no day of the week pattern was present a seven day delay in maintenance appears present. Analysis of individual monitors indicated the seven day cycle (Ed 25 & Glas 20) and the one day delay (Ed St Leonards, Ed 24 & Aber 2) were specific to certain monitors. The within monitor average length varies between 5-27 days with a standard deviation of 5-48 days. Because of the wide variation and the limited time frame, the results from the original donor monitor (Ed14 mean = 10 and s.d.=15) were used to simulate the 'blocking' in a truncated (>0) normal distribution under two scenarios; blocking with a mean length of

7 and 14 days with a standard deviation of 8 days in each case. This was in order to replicate a relatively common 1 day fix and simultaneously the long tailed distribution.

Figure 6.1 - Graphical display illustrating the distributional properties of cumulative days with missing data designated as ‘blocks’ of missing days. Note all monitors combined.



Both the seasonal differences and the ‘blocking’ were applied initially individually and then whilst in combination, assuming three levels of total percentage missing (5%, 15%, and 25%) in each case. Table 6.3 describes the five MCAR and twenty-four MAR scenarios to be implemented in the simulation study.

Table 6.3 - Summary of the MAR and MCAR scenarios applied in simulation study

Characteristics		Details/Percentage Missing	5%	10%	15%	20%	25%
MCAR		Random	X	X	X	X	X
MAR	Season (Sum:Win)	60%:40%	X	-	X	-	X
		75%:25%	X	-	X	-	X
	Blocking (days)	Mean length=7, S.D.=8 days	X	-	X	-	X
		Mean length=14, S.D.=8 days	X	-	X	-	X
	Season (Sum:Win) & Blocking (days)	60%:40% & Mean length = 7, SD=8	X	-	X	-	X
		60%:40% & Mean length = 14, SD=8	X	-	X	-	X
		75%:25% & Mean length = 7, SD=8	X	-	X	-	X
		75%:25% & Mean length = 14, SD=8	X	-	X	-	X

X = Proposed simulation scenario

Sum = Summer months April to September, Win = Winter months October to March

6.3 Simulation study results - MCAR

Each missing data scenario and the subsequent missing data analysis as outlined in Chapter 3.8 was repeated 1000 times. In each repetition the pollutant and temperature effect estimates and corresponding standard errors were recorded and averaged across the simulations. Table 6.4 reports the complete cases and multiple imputation results for both pollution and temperature under the MCAR scenarios. The table reports for the effect estimates and the standard errors the absolute bias, the percentage bias, the standardised percentage bias, and a one-sample t-test comparison test p-value. Given the small initial size of the true effect a small amount of absolute bias can correspond to a large percentage bias. With some consideration to the overall percentage of missing data, a standardised percentage bias of 15% was thought to be acceptable which was based on the power calculation reported in Chapter 3.8.4.

As expected under these MCAR scenarios very little bias was observed in the main effect estimate for both pollution and temperature in either the complete cases or multiple imputation models. In a generalised linear model under MCAR the CC analysis is expected to report unbiased effect estimates. The similarity between CC and MI in the results reported here indicates that the missing observation random allocation procedure containing no hidden MAR component, and more importantly the imputation model parameters adequately included information relating to the value as well as its presence or not. Bias in the standard errors was greatly improved in the MI compared to the CC results, a reflection of the improvement in sample size increased. This improvement in accuracy was reflected in the Mean Square Error (MSE) term which, as the percentage of missing data increased, showed a 3 to 15% decrease in MSE for the MI result.

Missing data were not directly simulated in temperature. Any bias present in the results was due to missing observations occurring due to missing pollution data i.e. the complete cases procedure. Multiple imputation was therefore not applied as missing temperature values were returned when the missing pollution values were imputed. The temperature results from the simulations are unbiased which indicates that the underlying relationship between the temperature and pollution was being maintained in the imputation model.

Table 6.4 – Bias reported in pollution simulation results for complete cases (CC) and multiple imputation (MI) analysis under missing completely at random (MCAR) characteristics

Percent Missing	Model ^a	Effect Estimate				Standard Error Estimate	
		Bias	Bias% ^b	Std Bias% ^c	p-val ^d	Bias	Bias% ^b
<i>Black Smoke Pollution (True Effect size = 0.0003302, S.E. = 0.0007093)</i>							
5%	CC	0.000007	2.1	1.0	0.259	0.000025	3.5
	MI	0.000003	0.9	0.4	0.521	0.000014	2.0
10%	CC	-0.000005	-1.4	-0.6	0.601	0.000052	7.3
	MI	-0.000004	-1.2	-0.5	0.569	0.000031	4.3
15%	CC	0.000010	2.9	1.2	0.400	0.000081	11.5
	MI	0.000005	1.6	0.7	0.523	0.000046	6.5
20%	CC	-0.000021	-6.5	-2.6	0.143	0.000117	16.5
	MI	-0.000025	-7.6	-3.2	0.014	0.000067	9.5
25%	CC	0.000001	0.5	0.2	0.926	0.000154	21.8
	MI	-0.000022	-6.6	-2.7	0.050	0.000086	12.2
<i>Temperature (True Effect size = -0.0012516, S.E. = 0.0024583)</i>							
5%	CC	-0.000004	-0.4	-0.2	0.821	0.000083	3.4
	MI	-0.000002	-0.1	-0.1	0.571	0.000000	0.0
10%	CC	-0.000018	-1.4	-0.7	0.549	0.000179	7.3
	MI	-0.000009	-0.7	-0.4	0.039	0.000001	0.0
15%	CC	-0.000046	-3.7	-1.7	0.212	0.000281	11.4
	MI	-0.000006	-0.5	-0.2	0.260	0.000001	0.1
20%	CC	-0.000036	-2.9	-1.3	0.424	0.000398	16.2
	MI	-0.000028	-2.3	-1.2	<0.001	0.000002	0.1
25%	CC	-0.000042	-3.4	-1.4	0.417	0.000531	21.6
	MI	-0.000029	-2.3	-1.2	<0.001	0.000002	0.1

a. CC = Complete Cases analysis, MI = Multiple Imputation analysis

b. Bias percentage - difference between true and simulated value divided by the true

c. Standardised bias percentage - difference between true and simulated value, divided by standard error

d. P-value associated with a one sample t-test comparison with the true effect.

6.4 Simulation study results - MAR

The results of the simulation study when missing pollution values are applied under the MAR scenarios, season and blocking independently, are reported in Table 6.5. These MAR scenarios were applied under a total percentage missing of 5%, 15%, and 25%. In all but one scenario the bias was observed to tend towards the null (i.e. negative bias), indicating that under MAR the complete cases analysis and to a lesser extent the multiple imputation analysis was likely to underestimate of the true effect, reducing the importance of the result but also reducing the chance of a false positive occurring.

In both analysis techniques, the complete cases and multiple imputation the bias observed in the main effect estimate was considered acceptable ($\leq 15\%$) when the characteristic seasonal ratio was low (60:40) or the total percentage missing was small i.e. 75:25 with 15% or 5% missing data. Additionally, in all but the scenario summer:winter ratio of 75:25 at 25% total missing, the bias was improved when missing data were accounted for using the multiple imputation model.

Blocking i.e. cumulative sets of missing days appears to have had minimal influence on the pollution effect estimate under the complete cases analysis, with the bias in the results similar to MCAR. Missing data blocks that were generated by the combination of two random number generations, were unlikely to influence the main pollution effect here as the pollution effect is represented as a single day (lag 0) effect. The complete cases results were therefore comparable to the MCAR results. Blocking is likely to have a much greater influence when modelling lagged effects across longer time periods such as the 30 days in the main study. The results in fact show a greater amount of bias when using the multiple imputation model compared to the complete cases. Regardless of total percentage missing the bias appears to be less than that associated with the seasonal ratio and comparable between the two block lengths (mean 7 or 14 days). As total percentage missing increases the percentage bias (standardised bias percentage) associated with blocking also increases, peaking at the previously thought acceptable level of -15.0% (-6.1%) and -14.2% (-5.7%) for mean length 7 and 14 days respectively. In all but one scenario the bias observed has been within the acceptable level, and the bias associated with the standard error was always reduced in the MI model. As

confirmed by the percentage drop in accuracy estimate (MSE) being between 3-10% depending on the total percentage missing. Yet, the slight increase in bias compared to the MCAR result indicates that the multiple imputation model was not as strong when the presence of missing data (yes/no) was correlated with the presence of missing data in the previous day. In other words, including the adjacent days pollution estimates (which also contain missing data) as predictors may not be enough to adequately predict the exposure value of a missing day when in the presence missing blocks.

Table 6.5 – Bias reported in pollution simulation results for complete cases (CC) and multiple imputation (MI) analysis under missing at random (MAR) characteristics season and blocking

MAR Properties	Percent Missing	Model ^a	Effect Estimate				Standard Error Estimate	
			Bias	Bias% ^b	Std Bias% ^c	p-val ^d	Bias	Bias% ^b
<i>Black Smoke Pollution (True Effect size = 0.0003302, S.E. = 0.0007093)</i>								
Sum:Win 60:40 ratio	5%	CC	-0.000011	-3.4	-1.5	0.041	0.000020	2.9
		MI	-0.000005	-1.6	-0.7	0.216	0.000012	1.7
	15%	CC	-0.000019	-5.6	-2.4	0.073	0.000069	9.8
		MI	-0.000014	-4.1	-1.8	0.076	0.000041	5.8
	25%	CC	-0.000040	-12.1	-4.8	0.006	0.000125	17.6
		MI	-0.000041	-12.4	-5.2	0.000	0.000070	9.9
Sum:Win 75:25 ratio	5%	CC	-0.000021	-6.2	-2.8	<0.001	0.000016	2.3
		MI	-0.000008	-2.3	-1.0	0.053	0.000010	1.3
	15%	CC	-0.000049	-14.9	-6.4	<0.001	0.000053	7.4
		MI	-0.000029	-8.9	-4.0	<0.001	0.000031	4.4
	25%	CC	-0.000108	-32.6	-13.4	<0.001	0.000096	13.5
		MI	-0.000091	-27.7	-11.9	<0.001	0.000058	8.1
Blocks Mean=7 S.D.=8 days	5%	CC	-0.000001	-0.5	-0.2	0.826	0.000025	3.5
		MI	-0.000003	-0.8	-0.4	0.643	0.000018	2.6
	15%	CC	-0.000004	-1.3	-0.6	0.728	0.000082	11.5
		MI	-0.000021	-6.4	-2.7	0.044	0.000059	8.3
	25%	CC	-0.000014	-4.3	-1.6	0.437	0.000154	21.6
		MI	-0.000050	-15.0	-6.1	0.000	0.000105	14.9
Blocks Mean=14 S.D.=8 days	5%	CC	-0.000008	-2.4	-1.1	0.257	0.000025	3.5
		MI	-0.000011	-3.3	-1.5	0.070	0.000019	2.7
	15%	CC	0.000012	3.5	1.5	0.391	0.000083	11.7
		MI	-0.000018	-5.5	-2.4	0.100	0.000063	8.9
	25%	CC	-0.000001	-0.5	-0.2	0.932	0.000153	21.6
		MI	-0.000047	-14.2	-5.7	0.001	0.000110	15.5

a. CC = Complete Cases analysis, MI = Multiple Imputation analysis

b. Bias percentage - difference between true and simulated value divided by the true

c. Standardised bias percentage - difference between true and simulated value, divided by standard error

d. P-value associated with a one sample t-test comparison with the true effect.

The two MAR scenarios (seasonal and blocking) were applied whilst in combination and the results reported in Table 6.6. As stated, the influence of blocking was minimal on the main effect (lag 0) and the imputation model slightly under performed when blocking was present. Therefore bias associated with the complete cases results tended to be less than the MI results here and comparable in size to the bias associated with seasonal ratio alone.

Table 6.6 – Bias reported in pollution simulation results for complete cases (CC) and multiple imputation (MI) analysis under combined missing at random (MAR) characteristics season and blocking

MAR Properties	Percent Missing	Model ^a	Effect Estimate				Standard Error Estimate	
			Bias	Bias% ^b	Std Bias% ^c	p-val ^d	Bias	Bias% ^b
<i>Black Smoke Pollution (True Effect size = 0.0003302, S.E. = 0.0007093)</i>								
Blocks Mean=7 SD=8, & Sum:Win 60:40	5%	CC	-0.000009	-2.7	-1.2	0.177	0.000021	3.0
		MI	-0.000016	-4.7	-2.1	0.005	0.000016	2.2
	15%	CC	-0.000018	-5.3	-2.3	0.113	0.000068	9.5
		MI	-0.000034	-10.2	-4.5	0.000	0.000050	7.0
	25%	CC	-0.000016	-4.8	-2.1	0.168	0.000124	17.5
		MI	-0.000057	-17.3	-6.9	<0.001	0.000086	12.1
Blocks Mean=7 SD=8, & Sum:Win 75:25	5%	CC	-0.000031	-9.2	-4.2	<0.001	0.000017	2.4
		MI	-0.000035	-10.5	-4.8	<0.001	0.000012	1.7
	15%	CC	-0.000049	-15.0	-6.5	<0.001	0.000054	7.6
		MI	-0.000082	-24.8	-11.0	<0.001	0.000040	5.6
	25%	CC	-0.000079	-23.9	-9.8	<0.001	0.000095	13.4
		MI	-0.000164	-49.6	-21.1	<0.001	0.000067	9.5
Blocks Mean=14 SD=8, Sum:Win 60:40	5%	CC	-0.000005	-1.5	-0.7	0.431	0.000021	3.0
		MI	-0.000010	-2.9	-1.3	0.070	0.000016	2.2
	15%	CC	-0.000032	-9.6	-4.1	0.005	0.000068	9.6
		MI	-0.000052	-15.7	-6.8	<0.001	0.000051	7.2
	25%	CC	-0.000024	-7.4	-3.1	0.050	0.000122	17.3
		MI	-0.000091	-27.6	-11.0	<0.001	0.000084	11.9
Blocks Mean=14 SD=8, & Sum:Win 75:25	5%	CC	-0.000022	-6.7	-3.0	<0.001	0.000016	2.3
		MI	-0.000030	-9.2	-4.2	<0.001	0.000012	1.7
	15%	CC	-0.000059	-17.7	-7.7	<0.001	0.000053	7.4
		MI	-0.000084	-25.4	-11.2	<0.001	0.000039	5.5
	25%	CC	-0.000087	-26.2	-10.8	<0.001	0.000095	13.3
		MI	-0.000158	-47.7	-20.2	<0.001	0.000071	10.0

a. CC = Complete Cases analysis, MI = Multiple Imputation analysis
b. Bias percentage - difference between true and simulated value divided by the true
c. Standardised bias percentage - difference between true and simulated value, divided by standard error
d. P-value associated with a one sample t-test comparison with the true effect.

Missing data blocking was present in the observed main dataset and will affect the imputed values, so the improvement in accuracy and level of bias in the multiple imputation models is important to note. Especially as the combination of MAR characteristics exacerbates the bias.

In both 60:40 seasonal ratio scenarios the standardised percentage bias stayed below 15%, though the bias at 25% total percentage missing and 75:25 seasonal ratio does cross 20%. As when applied individually the block length appears to have caused little difference in the effect. This may become more important if the average block length becomes greater than 30 days, not only if a 30 day lag length is being modelled but also when the study design, the time-stratified case-crossover design, matches case and controls within a month. At 25% total missing, a 75:25 seasonal ratio, and mean block length at 14 days the percentage bias (standardised bias percentage) was at -47.4% (-20.2%). This large bias towards the null hypothesis indicates that results were strongly underestimating the true effect when missing data due to ‘blocking and season’ was present and the likelihood of a false negative result was increased. However, as subjects dropped due to missing data are included when multiple imputation modelling was employed, the sample size was increased and bias associated with standard error estimation reduced. It’s worth noting that even with the increased bias in the imputed analysis the MSE was still constantly lower than the complete cases analysis. The magnitude of the drop was smaller at less than 5% but the improvement in accuracy was still present even with the slightly greater bias.

The equivalent results for temperature can be found in Table 6.7 and Table 6.8. As temperature data were a covariate and only set to be missing if pollution data were missing, by imputing the missing pollution data the missing temperature data returns. Though still tending toward the null, the percentage bias associated with missing temperature data in a complete case analysis was much greater than seen in any pollution result. As temperature effects are greater in general than pollution, the greater amount of bias was more acceptable. Even so, at the lowest total percentage missing (5%) the bias percentage for season was almost double that seen in pollution at -11.9% and -15.5% for the 60:40 and 75:25 seasonal ratio respectively. When total percentage missing increases to 25%, this percentage bias increases to -54.9% and -109.6%,

respectively. At this level of bias a borderline switch in the direction of the effect is becoming more likely. As expected the results associated with MI returns the bias to approximately zero (<5%) indicating that any relationship between pollution and temperature was maintained within the imputed data even when pollution was subject to more extreme MAR characteristics.

Table 6.7 - Temperature simulation results for Complete Cases (CC) and Multiple Imputation (MI) under Missing at Random characteristics season and blocking

MAR Properties	Percent Missing	Model ^a	Effect Estimate				Standard Error Estimate	
			Bias	Bias% ^b	Std Bias% ^b	p-val ^c	Bias	Bias% ^b
Temperature (True Effect size = -0.0012516, S.E. = 0.0024583)								
Sum:Win 60:40 ratio	5%	CC	0.000149	-11.9	5.9	<0.001	0.000076	3.1
		MI	-0.000005	0.4	-0.2	0.066	0.000001	0.0
	15%	CC	0.000335	-26.7	12.4	<0.001	0.000249	10.1
		MI	-0.000015	1.2	-0.6	0.002	0.000002	0.1
	25%	CC	0.000688	-54.9	23.6	<0.001	0.000461	18.8
		MI	-0.000034	2.7	-1.4	<0.001	0.000004	0.2
Sum:Win 75:25 ratio	5%	CC	0.000194	-15.5	7.7	<0.001	0.000065	2.6
		MI	-0.000006	0.5	-0.2	0.025	0.000001	0.0
	15%	CC	0.000693	-55.3	25.9	<0.001	0.000213	8.7
		MI	-0.000021	1.6	-0.8	<0.001	0.000003	0.1
	25%	CC	0.001373	-109.6	48.3	<0.001	0.000386	15.7
		MI	-0.000061	4.9	-2.5	<0.001	0.000005	0.2
Blocks Mean=7 S.D.=8 days	5%	CC	-0.000030	2.4	-1.2	0.110	0.000086	3.5
		MI	-0.000008	0.7	-0.3	0.031	0.000000	0.0
	15%	CC	0.000012	-1.0	0.4	0.733	0.000286	11.6
		MI	-0.000027	2.2	-1.1	<0.001	0.000001	0.0
	25%	CC	0.000033	-2.7	1.1	0.503	0.000528	21.5
		MI	-0.000064	5.1	-2.6	<0.001	0.000003	0.1
Blocks Mean=14 S.D.=8 days	5%	CC	-0.000019	1.5	-0.8	0.303	0.000085	3.5
		MI	-0.000016	1.3	-0.6	0.000	0.000000	0.0
	15%	CC	0.000004	-0.3	0.1	0.910	0.000284	11.6
		MI	-0.000039	3.1	-1.6	<0.001	0.000003	0.1
	25%	CC	-0.000017	1.4	-0.6	0.726	0.000527	21.4
		MI	-0.000069	5.5	-2.8	<0.001	0.000011	0.4

a. CC = Complete Cases analysis, MI = Multiple Imputation analysis

b. Bias percentage/standardised bias percentage

c. P-value associated with a one sample t-test comparison with the true effect.

Table 6.8 - Temperature simulation results for Complete Cases (CC) and Multiple Imputation (MI) under combined Missing at Random characteristics season and blocking

MAR Properties	Percent Missing	Model ^a	Effect Estimate				Standard Error Estimate	
			Bias	Bias% ^b	Std Bias% ^b	p-val ^c	Bias	Bias% ^b
Temperature (True Effect size = -0.0012516, S.E. = 0.0024583)								
Blocks Mean=7 SD=8, & Sum:Win 60:40	5%	CC	0.000127	-10.2	5.0	<0.001	0.000076	3.1
		MI	-0.000014	1.1	-0.6	0.000	0.000000	0.0
	15%	CC	0.000323	-25.8	11.9	<0.001	0.000254	10.3
		MI	-0.000029	2.3	-1.2	<0.001	0.000001	0.0
	25%	CC	0.000654	-52.2	22.4	<0.001	0.000458	18.6
		MI	-0.000041	3.2	-1.6	<0.001	0.000002	0.1
Blocks Mean=7 SD=8, & Sum:Win 75:25	5%	CC	0.000240	-19.2	9.5	<0.001	0.000066	2.7
		MI	-0.000024	1.9	-1.0	<0.001	0.000001	0.0
	15%	CC	0.000761	-60.8	28.5	<0.001	0.000214	8.7
		MI	-0.000051	4.1	-2.1	<0.001	0.000007	0.3
	25%	CC	0.001404	-112.1	49.4	<0.001	0.000383	15.6
		MI	-0.000073	5.8	-3.0	<0.001	0.000009	0.4
Blocks Mean=14 SD=8, Sum:Win 60:40	5%	CC	0.000118	-9.4	4.6	<0.001	0.000077	3.1
		MI	-0.000010	0.8	-0.4	0.003	0.000000	0.0
	15%	CC	0.000339	-27.1	12.5	<0.001	0.000252	10.3
		MI	-0.000032	2.5	-1.3	<0.001	0.000003	0.1
	25%	CC	0.000648	-51.7	22.2	<0.001	0.000462	18.8
		MI	-0.000063	5.0	-2.6	<0.001	0.000006	0.2
Blocks Mean=14 SD=8, & Sum:Win 75:25	5%	CC	0.000258	-20.6	10.2	<0.001	0.000066	2.7
		MI	-0.000018	1.5	-0.7	<0.001	0.000001	0.0
	15%	CC	0.000767	-61.3	28.7	<0.001	0.000213	8.7
		MI	-0.000030	2.4	-1.2	<0.001	0.000006	0.2
	25%	CC	0.001339	-106.9	47.1	<0.001	0.000382	15.5
		MI	0.000001	-0.1	0.0	0.935	0.000019	0.8

a. CC = Complete Cases analysis, MI = Multiple Imputation analysis
b. Bias percentage/standardised bias percentage
c. P-value associated with a one sample t-test comparison with the true effect.

The results of this simulation study provide useful information not only within the air pollution context but also with respect to a time-series structured dataset. The bias observed in both the standard complete cases analysis, and when the missing data techniques ‘multiple imputation’ were applied, will provided a useful context to the results of the multiple imputation analysis in the main study (Chapter 8.4).

7 RESULTS – MODELLING TEMPERATURE

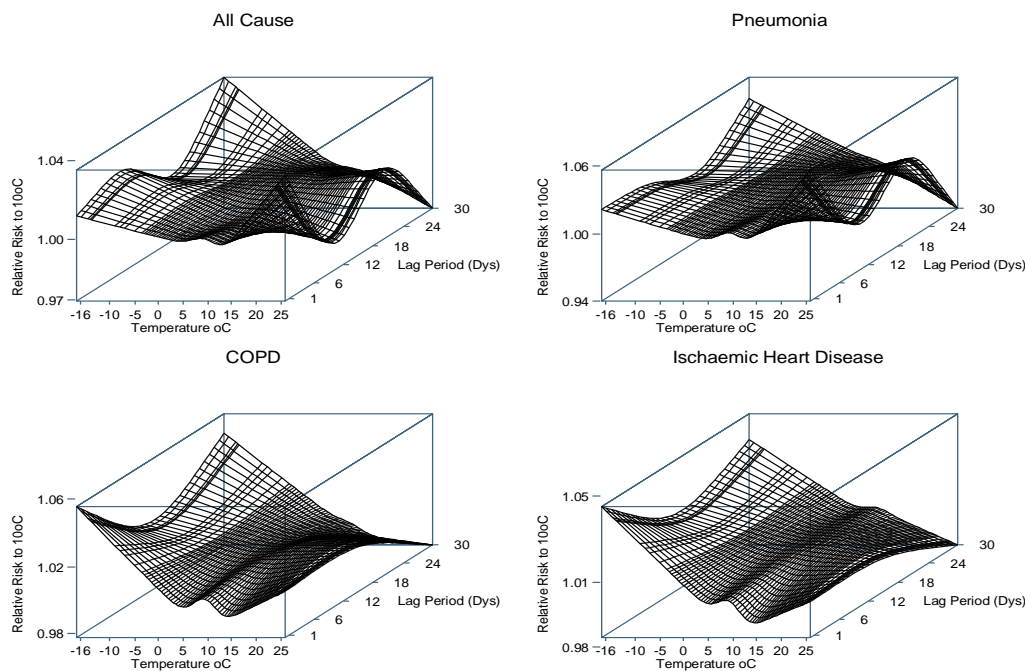
This chapter relates to an investigation of the temperature effects on the three causes of death. This chapter employs the main empirical dataset (Jan 1980 to Dec 2011) associated with the study that was described in full in Chapter 4.

7.1 Investigating temperature – mortality in Pneumonia, COPD, IHD

To explore any unique changes in risk for each cause of death the exposure range and lag period both were split into regions. Within each region a cubic function was fitted such that they connected at knot points, with the extreme regions constrained to be linear. The resulting cubic parameters for exposure and lag were combined in an interaction to form a distributed lag non-linear model. To allow flexibility, a varying number of knot points were fitted to both temperature and lag with the optimum combination identified using Akaike Information Criteria (AIC). The cause of death was identified in the mortality dataset under ‘any’ cause of death field (AFCOD). The all-cause of death knot combination of 5 for exposure and 4 for lag was chosen for each cause of death, except IHD which identified a 5:3 knot combination. Figure 7.1 plots for each cause of death the relative risk compared to lag 1 and the approximate mean temperature 10°C.

The plots in Figure 7.1 displayed a non-linear temperature-response effect that was persistent across the entire lag for all three causes of death. The non-linear relationship across the temperature exposure was largely in the form of a U shape for the immediate lag period (within 6 days). Though the shape changed across the 30 days, it was considered that a non-linear relationship with three distinct regions was generally observed. This was either in the form of a U shape throughout or the U shape transformed in and out of a reverse slanted J shape.

Figure 7.1 – Simultaneous plotting across the temperature range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields in a distributed lag non-linear model.



Knot no. combinations (temperature:lag) for All Cause(5:4), pneumonia (5:4), COPD (5:4), IHD (5,3)
Relative Risk comparable to lag 1 at approximate mean temperature (10°C)

In each case the outer temperature ranges (extreme cold or heat) show increased risk of mortality compared to the mild temperatures. Chronic obstructive pulmonary disease and IHD show similar patterns in relative risk, with the colder temperatures (<0°C) reporting stronger immediate risk than pneumonia that after a slight drop persists for the entire 30 day lag. Here warm temperature (>15°C) shows a modest immediate increase in risk that increases gradually and peaks (≈12 days) before dissipating. Pneumonia has a much stronger immediate increase in risk associated with the warm temperatures that quickly dissipates within 6 days before a brief but equally strong rebound at 18-24 days. A small immediate increase in pneumonia mortality risk was seen in cold temperatures that fluctuated slightly until 12 days when risk appeared to increase sharply. This strong rebound was only seen in extreme cold temperature which means it may be an artefact of a strong outlier or the small sample of days with extreme cold temperatures.

7.2 Bespoke temperature modelling for Pneumonia, COPD, and IHD

To improve the understanding and improve the comparisons between changes in risk across temperature exposure and lag period, the slightly less sophisticated models the lag stratified and distributed lag models, were applied. This began by splitting the temperature exposure range into three distinct ‘zones’ which as indicated by Figure 7.1 remained largely consistent regardless of changes across the lag period. Threshold modelling described in full in section 3.4.4, split the exposure range into three zones ‘cold temperature zone’, ‘mild temperature zone’, and a ‘warm temperature zone’ connected at knot points identified by the AIC. Table 7.1 describes the unique or ‘bespoke’ knot positions identified for each cause of death and applied to a lag stratified model. Table 7.1 further describes the percentage relative risk (%RR) associated with a 1°C increase in 7am-11pm average temperature on any single day, firstly within a 30 day lag period and then for the 30 days split into five lag periods of six days each set as 1-6, 7-12, 13-18, 19-24, 25-30 days.

As an example using pneumonia mortality, a 1°C increase in cold temperature (i.e. increasing towards the 1°C threshold) in the 1-6 day lag has a %RR (95% C.I.) of -0.46 (-0.66,-0.24). Indicating that one degree increase in the cold temperature range on an individual day corresponds to a reduced risk of death of 0.46%. Conversely, an increase of one degree in warm temperatures (i.e. >15°C) sees a non-significant increase in the risk for pneumonia within six days of 0.08% (-0.10%, 0.27%).

Table 7.1 – Percentage relative risk (%RR) associated with an increase in 1°C within each temperature zone associated with the lag stratified analysis bespoke for each cause of death.

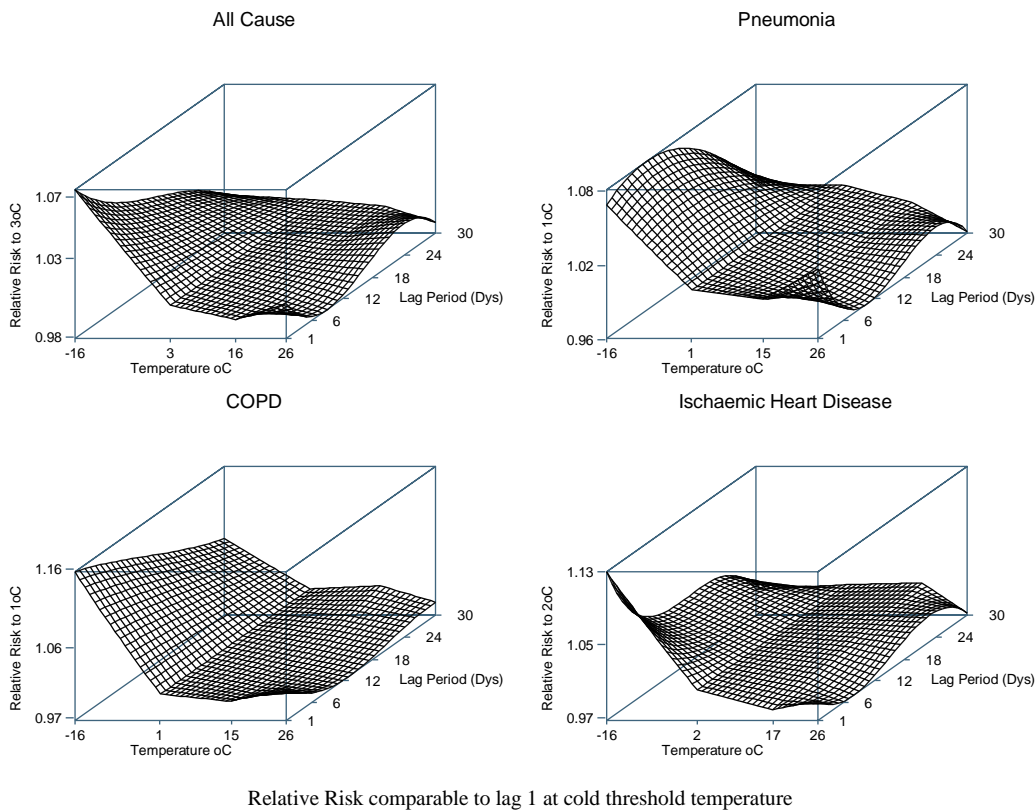
Lag Length	1oC increase within Temperature thresholds (%RR(95% C.I.)) ^a		
	<3°C	3-16°C	>16°C
All Cause			
30 Days	-0.20(-0.23,-0.16)	-0.06(-0.07,-0.05)	-0.04(-0.10,0.01)
1-6 Days	-0.40(-0.47,-0.34)	-0.08(-0.12,-0.07)	-0.07(-0.17,0.03)
7-12 Days	-0.18(-0.25,-0.12)	-0.08(-0.12,-0.07)	-0.08(-0.18,0.02)
13-18 Days	-0.22(-0.29,-0.15)	-0.03(-0.07,-0.02)	0.00(-0.10,0.10)
19-24 Days	-0.13(-0.20,-0.07)	-0.05(-0.07,-0.02)	-0.03(-0.12,0.07)
25-30 Days	-0.05(-0.12,0.03)	-0.02(-0.05,0.00)	-0.03(-0.13,0.07)
Pneumonia	<1°C	1-15°C	>15°C
30 Days	-0.35(-0.47,-0.23)	-0.13(-0.15,-0.11)	-0.06(-0.15,0.03)
1-6 Days	-0.46(-0.66,-0.24)	-0.12(-0.17,-0.05)	0.08(-0.10,0.27)
7-12 Days	-0.49(-0.71,-0.29)	-0.13(-0.20,-0.08)	-0.18(-0.37,-0.02)
13-18 Days	-0.37(-0.59,-0.15)	-0.15(-0.20,-0.10)	-0.07(-0.25,0.12)
19-24 Days	-0.35(-0.58,-0.13)	-0.12(-0.17,-0.07)	-0.07(-0.25,0.12)
25-30 Days	-0.08(-0.30,0.15)	-0.10(-0.15,-0.03)	-0.08(-0.27,0.10)
COPD	<1°C	1-15°C	>15°C
30 Days	-0.62(-0.78,-0.45)	-0.10(-0.13,-0.07)	-0.09(-0.22,0.04)
1-6 Days	-0.96(-1.27,-0.66)	-0.12(-0.20,-0.05)	0.12(-0.13,0.38)
7-12 Days	-0.59(-0.90,-0.29)	-0.13(-0.22,-0.05)	-0.20(-0.46,0.07)
13-18 Days	-0.66(-0.97,-0.34)	-0.13(-0.22,-0.05)	-0.05(-0.30,0.22)
19-24 Days	-0.51(-0.82,-0.17)	-0.08(-0.17,0.00)	-0.13(-0.39,0.12)
25-30 Days	-0.39(-0.71,-0.05)	0.00(-0.07,0.08)	-0.18(-0.46,0.07)
IHD	<2°C	2-17°C	>17°C
30 Days	-0.26(-0.33,-0.18)	-0.06(-0.08,-0.04)	-0.16(-0.29,-0.02)
1-6 Days	-0.61(-0.76,-0.47)	-0.13(-0.18,-0.10)	-0.25(-0.51,0.00)
7-12 Days	-0.13(-0.30,0.02)	-0.10(-0.13,-0.05)	-0.13(-0.39,0.12)
13-18 Days	-0.29(-0.44,-0.13)	-0.03(-0.08,0.02)	-0.17(-0.42,0.08)
19-24 Days	-0.17(-0.32,0.00)	-0.02(-0.07,0.02)	-0.02(-0.27,0.23)
25-30 Days	-0.10(-0.27,0.05)	0.00(-0.03,0.05)	-0.22(-0.47,0.03)

a. Percentage change in Relative Risk (95% Confidence interval) associated with a 1oC increase in temperature on an single day within the lag period

The lag stratified models indicated for all three causes of death a strong decrease in risk as temperature increased in the cold temperature range with a weaker effect in the mild temperatures. For all three causes a statistically significant ($p\text{-value}\leq 0.05$) decrease for the cold temperature range was seen up until at least 24 days post exposure, with an effect on COPD still significant 30 days later. Associations relating to cold temperatures were persistent for longer periods of time before tending towards the null effect with the respiratory diseases affected for longer than the IHD. COPD reported the greatest effect estimates in the cold temperature ranges regardless of the lag. Though non-significant, an increase in temperature effects during the warm temperature range showed an increased risk of respiratory mortality for the short lag (<6 days). This was strongest in COPD with a percentage relative risk (95% C.I.) for a 1°C increase above 15°C of 0.12% (-0.13%,0.38%). This was not present in IHD which produced a borderline significant decrease risk in the 1-6 day lag of -0.25% (-0.51%,0.00%).

The lack of an effect within the short-term lag (1-6 day) for IHD was unexpected, as the potential biological mechanisms indicate a very quick response should occur. Whereas the distributed lag non-linear (Figure 7.1) function was largely driven by the data meaning random variation may play a part in the shape, the lag stratified model averages across a 6 day lag, smoothing out any within lag variation. In between, the cubic distributed lag model allowed more flexibility within each lag period for each temperature zone. This means 1 or 2 day lagged effects that are present but are smoothed out by the lag stratified model may be observed. Again surface plots (Figure 7.2) were identified as the most straightforward method of observing the results. Figure 7.2 reports for each cause of death the relative risk related to the risk observed at the lag 1 for the cold-mild temperature threshold connection.

Figure 7.2 – Simultaneously plotting across the temperature range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields using a cubic distributed lag model.



The results here largely agree with those reported in lag stratified model, especially for the respiratory deaths. The cubic distributed lags replicated for the respiratory diseases the strong increased risk in cold temperatures that persisted for longer into the lag, along with the equally strong initial risk for ischaemic heart disease that had a rapid drop after 6 days. However, the cubic distributed lag for Ischaemic heart disease did display a slight increase in risk within the immediate 1-3 day lag period. The results were broadly in agreement with the lag stratified model, but lag periods with a length of 6 days that also do not include the same day effect (lag 0) may be too broad to identify effects that occur immediately within hours of exposure. This may be the case in ischaemic heart disease.

7.3 Comparing the temperature effects between causes of death

The analysis so far performed in Chapter 7 identified the cause of death based on ‘any’ cause of death field in the mortality dataset i.e. a subject was designated pneumonia if the ‘primary’ cause of death field or any of the ‘secondary’ cause of death fields contained pneumonia (see Chapter 1.8 and 3.9 for more details). This meant the potential for overlap was present as a subject may have COPD as the ‘primary’ cause but also ‘pneumonia’ as a secondary cause. To compare the change in temperature effects between the three causes of death any overlap in study participants needed to be removed. Therefore an analysis based only on the primary cause of death (PCOD) field was used with fixed threshold points for the three temperature zones modelled.

The three knot points were based on those identified for the general non-specific ‘all-cause’ of death i.e. $<3^{\circ}\text{C}$, $3\text{-}16^{\circ}\text{C}$, and $>16^{\circ}\text{C}$ as seen in Table 7.1. Note, ‘all-cause’ refers to all deaths from any underlying cause, rather than the previously defined ‘any’ cause of death field (AFCOD) that relates to the designated field within the mortality dataset that is partly dependent on the position within the chain of events leading up to death. A three factor categorical variable is then generated to represent the Primary Cause of Death (PCOD) set as pneumonia, COPD, or IHD. Again be aware that here the ‘primary cause of death’ simultaneously refers to the field in the dataset, and the specific cause of death defined thought to be the ‘underlying’ cause i.e. the cause that began the chain of events leading to death.

The z-test (see Chapter 3.9) compared the effect estimates produced by the three causes of death. To fully compare the three causes the test was repeated for the available combinations of pneumonia, COPD, and IHD. Table 7.2 reports the %RR associated with each cause of death separately whilst using the ‘all-cause’ of death knot points. The comparison test was then performed and p-value reported in the table firstly with respect to pneumonia (v_{Pneu}), and secondly with respect to COPD (v_{COPD}). Note, the effect estimates reported in Table 7.2 differ from the results in main temperature analysis (Table 7.1) as the knot points across the temperature exposure range have been fixed.

Table 7.2 – Cause of death specific percentage relative risk (%RR) associated with 1°C increase within each temperature zone (identified by all-cause of death), with comparison test result of three causes of death by lag period.

Temp range	Lag (Dys)	1°C increase within temperature thresholds (%RR (95% C.I.)) ^a					
		Pneumonia	COPD	vPneu	IHD	vPneu	vCOPD
<3°C	1-30	-0.29(-0.41,-0.15)	-0.52(-0.66,-0.37)	0.008	-0.21(-0.28,-0.14)	0.289	<0.001
	1-6	-0.22(-0.47,0.03)	-0.85(-1.13,-0.58)	0.001	-0.59(-0.71,-0.46)	0.010	0.076
	7-12	-0.46(-0.70,-0.20)	-0.54(-0.82,-0.25)	0.613	-0.07(-0.20,0.08)	0.007	0.002
	13-18	-0.30(-0.56,-0.05)	-0.52(-0.82,-0.24)	0.254	-0.24(-0.37,-0.08)	0.647	0.073
	19-24	-0.07(-0.32,0.20)	-0.30(-0.59,0.00)	0.233	-0.13(-0.27,0.02)	0.661	0.303
	25-30	-0.29(-0.56,-0.03)	-0.35(-0.64,-0.07)	0.742	-0.07(-0.22,0.08)	0.126	0.073
3-16°C	1-30	-0.20(-0.24,-0.15)	-0.12(-0.17,-0.07)	0.016	-0.05(-0.07,-0.03)	<0.001	0.015
	1-6	-0.08(-0.20,0.02)	-0.12(-0.24,0.02)	0.819	-0.13(-0.20,-0.08)	0.431	0.661
	7-12	-0.18(-0.29,-0.07)	-0.13(-0.25,0.00)	0.566	-0.08(-0.15,-0.03)	0.169	0.583
	13-18	-0.35(-0.47,-0.24)	-0.12(-0.25,0.00)	0.008	-0.02(-0.07,0.05)	<0.001	0.112
	19-24	-0.18(-0.30,-0.07)	-0.10(-0.24,0.02)	0.360	-0.02(-0.08,0.03)	0.009	0.214
	25-30	-0.15(-0.25,-0.03)	-0.08(-0.22,0.03)	0.477	0.02(-0.03,0.08)	0.007	0.115
>16°C	1-30	-0.11(-0.34,0.12)	-0.04(-0.30,0.21)	0.708	-0.16(-0.27,-0.05)	0.703	0.421
	1-6	0.10(-0.34,0.53)	0.28(-0.20,0.77)	0.591	-0.30(-0.51,-0.10)	0.103	0.033
	7-12	-0.51(-0.94,-0.05)	-0.22(-0.70,0.28)	0.390	-0.17(-0.37,0.05)	0.165	0.839
	13-18	0.03(-0.40,0.48)	-0.07(-0.56,0.43)	0.778	-0.15(-0.37,0.05)	0.444	0.727
	19-24	-0.02(-0.46,0.43)	-0.07(-0.56,0.43)	0.904	-0.05(-0.25,0.17)	0.906	0.967
	25-30	-0.18(-0.63,0.27)	-0.18(-0.68,0.33)	0.998	-0.13(-0.34,0.08)	0.845	0.860

a. Percentage change in Relative Risk (95% Confidence interval) associated with 1°C increase in temperature on a single day within the lag period

vPneu. p-value associated with z-test comparison with Pneumonia

vCOPD. p-value associated with z-test comparison with COPD

The reduced number of warm temperature days in Scotland may have played a part, as the only significant difference was observed in the shortest lag (1-6 days) between IHD and COPD. In the cold temperatures, significant differences tended to occur at 1-6 day lag where stronger effects were seen in COPD and IHD compared to pneumonia. The already described drop in risk observed in IHD after 6 days meant that the IHD effect was then significantly lower (lag 7-12) than the now similar COPD and Pneumonia. Mild temperatures were not significantly different between the three causes until after 7 days where IHD effect was lower than both, but significantly lower than pneumonia. Confirming respiratory diseases act similarly at cold and mild temperatures where effects were immediate and last the full lag, whereas IHD has an immediate effect that only last 6 days. This can be seen in the 30 day lag where the overall effect (IHD = -0.21% for cold and -0.05% for mild) is significantly lower than both pneumonia (-0.29% and -0.20%) and COPD (-0.52% and -0.12%).

7.4 Influence of hospital admission moderator on temperature effect

To determine if subjects located in hospital for some or all of the exposure period was causing incorrect estimation of the true temperature effect, the analysis was stratified by subjects' presence in hospital during exposure. Analysis here reverts back to the bespoke temperature models for each cause of death as identified by the AFCOD. A 'location' variable identified participants whose matched hospital admissions information indicated that during exposure they spent time in hospital for 'zero' days, '1-29' days, or 'all 30' days. This meant the 'zero' days in hospital group essentially represented those in the community during the entire exposure and the 'all 30' days those in hospital during exposure.

To determine if the effects observed for each cause of death were different by subject location during exposure the analysis was stratified by the three location groups. Again the first principles comparison z-test was performed firstly with respect to 'zero' days in hospital, and secondly with respect to '1-29' days in hospital as the reference category. Table 7.3, 7.4, and 7.5 reports for pneumonia, COPD, and IHD respectively the lag stratified models %RR associated with a 1°C increase within each temperature zone when further stratified by hospital admissions status.

7.4.1 Adjusting for hospital admission exposure – Pneumonia mortality

The greater the time spent in the community during exposure, the greater the effect cold temperatures had on pneumonia mortality risk. In all lag periods, an increase in temperature towards the cold-mild temperature connection point 1°C saw a greater decrease in risk for those in the community (at least 1 day) rather than in hospital (all 30 days). This was significant for the short-term lags 1-6 days and 7-12 days. Though a non-significant effect was observed, those who spent all 30 days in hospital were more likely to show an increase in risk as temperature increases towards 1°C (30 Day lag = 0.21% (-0.13%,0.55%)) compared to those spending at least some time in the community (zero, 1-29 days in hospital). As Table 7.3 indicates for pneumonia, within mild temperatures any increase relates to a decrease in risk that, in terms of magnitude was similar regardless of the hospital admission status. With respect to warm

temperatures, when all data were combined (%RR=0.08, 95% C.I.=-0.10,0.27 shown previously in Table 7.1) a small increase risk was observed in the short lag 1-6 days (%RR=0.13, 95% C.I.=-0.10,0.36). This was shown here to be present only in those with zero days in hospital. The remaining lag periods displayed the decreased risk reported already in the cold and mild temperatures with no clear hospital admission pattern.

Table 7.3 – Pneumonia specific percentage relative risk (%RR) and comparison test of hospital admission status during exposure (Zero, 1-29, all 30 days) associated with an increase in 1°C within each temperature zone

Temp Range	Lag (Dys)	Pneumonia 1°C increase within Temperature thresholds (%RR(95% C.I.)) ^a					
		Zero Days	1-29 Days	P-val vZero	30 Days	P-val vZero	P-val v1-29d
<1°C	1-30	-0.47(-0.62,-0.31)	-0.20(-0.38,-0.03)	0.015	0.21(-0.13,0.55)	<0.001	0.040
	1-6	-0.51(-0.80,-0.25)	-0.35(-0.70,-0.02)	0.445	0.22(-0.46,0.88)	0.048	0.143
	7-12	-0.75(-1.04,-0.49)	-0.12(-0.47,0.22)	0.004	0.13(-0.54,0.78)	0.016	0.512
	13-18	-0.42(-0.71,-0.13)	-0.34(-0.70,0.00)	0.741	0.07(-0.61,0.74)	0.195	0.288
	19-24	-0.34(-0.63,-0.05)	-0.40(-0.76,-0.08)	0.707	-0.15(-0.80,0.49)	0.620	0.477
	25-30	-0.24(-0.52,0.07)	0.07(-0.27,0.41)	0.189	0.64(-0.02,1.28)	0.020	0.140
1-15°C	1-30	-0.12(-0.15,-0.09)	-0.14(-0.18,-0.11)	0.325	-0.12(-0.19,-0.06)	0.909	0.618
	1-6	-0.07(-0.15,0.00)	-0.17(-0.27,-0.10)	0.057	-0.15(-0.32,0.00)	0.322	0.822
	7-12	-0.13(-0.22,-0.07)	-0.15(-0.25,-0.08)	0.621	-0.02(-0.18,0.13)	0.236	0.146
	13-18	-0.17(-0.25,-0.10)	-0.12(-0.22,-0.03)	0.365	-0.18(-0.35,-0.03)	0.897	0.491
	19-24	-0.12(-0.20,-0.05)	-0.12(-0.22,-0.03)	0.915	-0.08(-0.25,0.07)	0.761	0.717
	25-30	-0.07(-0.15,-0.02)	-0.12(-0.20,-0.02)	0.540	-0.17(-0.34,-0.02)	0.328	0.572
>15°C	1-30	-0.01(-0.13,0.11)	-0.13(-0.27,0.01)	0.195	-0.08(-0.34,0.18)	0.629	0.732
	1-6	0.13(-0.10,0.36)	0.00(-0.29,0.27)	0.455	-0.12(-0.63,0.40)	0.389	0.711
	7-12	-0.22(-0.47,0.02)	-0.15(-0.44,0.12)	0.686	-0.03(-0.56,0.48)	0.520	0.709
	13-18	-0.03(-0.27,0.20)	-0.12(-0.40,0.17)	0.651	-0.05(-0.58,0.48)	0.953	0.824
	19-24	0.13(-0.12,0.36)	-0.32(-0.61,-0.05)	0.016	-0.30(-0.83,0.22)	0.146	0.929
	25-30	-0.10(-0.35,0.15)	-0.05(-0.34,0.23)	0.796	0.10(-0.44,0.61)	0.528	0.650

a. Percentage change in Relative Risk (95% Confidence interval) associated with 1°C increase in temperature on a single day within the lag period

vZero. p-value associated with z-test comparison with zero days in hospital

v1-29. p-value associated with z-test comparison with 1-29 days in hospital

7.4.2 Adjusting for hospital admission during exposure – COPD mortality

Similar to pneumonia, more time spent in the community during exposure indicated a greater effect of cold temperatures on COPD mortality risk. In COPD this increased risk at colder temperatures only occurred in the short lag periods (lag 1-18), whereas for lag 18 or above the greater decrease risk occurred in those located in hospital for all 30 days. Even so, increased risk at colder temperatures when located in the community was only significantly different during lag 13-18 when compared between zero and 1-29 days. Interestingly the overall 30 day lag length reported similar effect sizes for all three locations (-0.60%, -0.62%, and -0.70%). indicating that for COPD total increased risk over 30 days is consistent but acting differently within the 30 days depending on location.

The effect sizes and differences in effect sizes between the three hospital admission groups were smaller for the mild temperatures. Yet an increase risk was observed for those with ‘all 30’ days that was significantly different to both ‘zero’ and ‘1-29’ days in hospital during the 1-30 and 19-24 day lag periods. Interestingly the short-term effects (<12 days) of increased mild temperatures on COPD appear to improve health (i.e. decreased risk) in those located in the community (‘zero’ days in hospital) during exposure (Table 7.4). Whereas, those located in hospital showed slight increases in %RR that were significantly different from the ‘1-29’ day hospital admission (at 1-6 days lag) and both ‘1-29’ and ‘zero’ in the 7-12 days lag. The decrease in risk occurring in mild temperatures persists for almost the full lag period (Zero,1-29 days), however in the 30 days in hospital a brief decrease in risk occurs before returning to an increased risk at 19-24 days (%RR=0.43 (0.15,0.69)).

Significant differences between the three locations during exposure were not present for warm temperature (>15°C). As with pneumonia, short-term increased COPD risk at warm temperatures for all data combined (Table 7.1 – lag 1-6 = 0.12%) appeared to be driven by those subjects in the community (Table 7.4 – lag 1-6=0.20%), with most subsequent lags (6-30 days) indicating a decrease in risk occurring as warm temperature increased for those in hospital.

Table 7.4 - COPD specific percentage relative risk (%RR) and comparison test of hospital admission status during exposure (Zero, 1-29, all 30 days) associated with an increase in 1°C within each temperature zone.

Temp Range	Lag (Dys)	COPD - 1oC increase within Temperature thresholds (%RR(95% C.I.)) ^a					
		Zero Days	1-29 Days	P-val vZero	30 Days	P-val vZero	P-val v1-29d
<1°C	1-30	-0.60(-0.83,-0.37)	-0.62(-0.86,-0.38)	0.874	-0.70(-1.27,-0.12)	0.696	0.762
	1-6	-1.01(-1.43,-0.61)	-0.89(-1.36,-0.46)	0.685	-0.20(-1.34,0.94)	0.180	0.259
	7-12	-0.71(-1.15,-0.30)	-0.42(-0.90,0.03)	0.346	-0.58(-1.70,0.56)	0.806	0.806
	13-18	-0.24(-0.71,0.20)	-1.06(-1.52,-0.63)	0.009	-0.05(-1.24,1.13)	0.770	0.113
	19-24	-0.56(-1.03,-0.12)	-0.42(-0.90,0.03)	0.663	-1.04(-2.13,0.05)	0.405	0.283
	25-30	-0.32(-0.80,0.13)	-0.40(-0.87,0.05)	0.817	-1.24(-2.24,-0.24)	0.086	0.115
1-15°C	1-30	-0.08(-0.13,-0.04)	-0.13(-0.18,-0.08)	0.100	0.05(-0.06,0.16)	0.034	0.003
	1-6	-0.12(-0.24,0.00)	-0.13(-0.25,-0.02)	0.796	0.05(-0.24,0.31)	0.280	0.224
	7-12	-0.12(-0.24,-0.02)	-0.15(-0.29,-0.03)	0.696	0.07(-0.22,0.33)	0.224	0.155
	13-18	-0.13(-0.25,-0.02)	-0.15(-0.27,-0.03)	0.809	-0.20(-0.49,0.07)	0.611	0.710
	19-24	-0.08(-0.20,0.03)	-0.12(-0.22,0.02)	0.966	0.43(0.15,0.69)	0.001	0.001
	25-30	0.12(0.00,0.23)	-0.12(-0.25,-0.02)	0.004	-0.17(-0.44,0.10)	0.051	0.750
>15°C	1-30	-0.05(-0.23,0.12)	-0.12(-0.31,0.07)	0.604	-0.20(-0.62,0.23)	0.527	0.741
	1-6	0.20(-0.15,0.54)	0.02(-0.37,0.38)	0.477	-0.15(-1.03,0.70)	0.451	0.723
	7-12	-0.10(-0.46,0.23)	-0.27(-0.66,0.10)	0.516	0.18(-0.70,1.06)	0.557	0.355
	13-18	0.08(-0.27,0.43)	-0.22(-0.61,0.17)	0.254	-0.34(-1.22,0.54)	0.375	0.800
	19-24	-0.10(-0.46,0.25)	-0.18(-0.59,0.18)	0.708	-0.47(-1.36,0.40)	0.414	0.549
	25-30	-0.44(-0.80,-0.08)	0.08(-0.30,0.48)	0.047	-0.08(-0.96,0.77)	0.465	0.712

a. Percentage change in Relative Risk (95% Confidence interval) associated with 1oC increase in temperature on an single day within the lag period

vZero. p-value associated with z-test comparison with zero days in hospital

v1-29. p-value associated with z-test comparison with 1-29 days in hospital

7.4.3 Adjusting for hospital admission during exposure – IHD mortality

Previous results (Table 7.1) for ischaemic heart disease indicated a persistent decreasing risk for all lag periods at different rates across the exposure range even above 17°C, though Figure 7.2 did indicate a slight increase in risk present in the first few 1-3 days. Regardless of location during exposure no increase in risk occurred during a warm temperature increase at any lags within 18 days. A non-significant increase was observed at lag 19-24 for those in hospital for at least one day ('1-29' and 'all 30' days), and these were not found to be significantly different from zero days. Though not significantly different across the three locations, at lag 1-6 days the decrease in risk for those in hospital (all 30 days = -0.80%) was greater than both 1-29 (-0.24%) and zero days (-0.29%). This stronger decrease in risk in the 'all 30' days persisted up until the end of lag 1-18.

Though significant differences in the effect sizes between hospital admissions are present in the cold and mild temperatures, no clear pattern stands out. Previously the main analysis (Table 7.1) indicated an immediate risk for cold temperatures that quickly dropped after lag 1-6 and held steady across the remaining lag periods (7-30 days). This appears to be replicated in those in the community with a %RR of -0.71% (lag 1-6) that then holds steady between 0.10%-0.15%. Those identified spending 'all 30' days in hospital also displayed a strong immediate risk, however a second strong period of risk was observed for colder temperatures between 13-18 day lag. The risk associated with '1-29' days in hospital appeared to follow a pattern representing a middle ground between the two groups at each lag period.

Table 7.5 – Ischaemic heart disease specific percentage relative risk (%RR) and comparison test of hospital admission status during exposure (Zero, 1-29, all 30 days) associated with an increase in 1°C within each temperature zone.

Temp Range	Lag (Dys)	IHD - 1oC increase within Temperature thresholds (%RR(95% C.I.)) ^a					
		Zero Days	1-29 Days	P-val vZero	30 Days	P-val vZero	P-val v1-29d
<2°C	1-30	-0.25(-0.34,-0.15)	-0.28(-0.41,-0.14)	0.689	-0.30(-0.62,0.02)	0.728	0.889
	1-6	-0.71(-0.89,-0.54)	-0.39(-0.66,-0.13)	0.043	-0.61(-1.24,-0.02)	0.774	0.482
	7-12	-0.13(-0.34,0.03)	-0.15(-0.42,0.13)	0.954	-0.03(-0.68,0.59)	0.761	0.750
	13-18	-0.18(-0.39,0.00)	-0.51(-0.78,-0.24)	0.057	-0.51(-1.15,0.12)	0.323	0.981
	19-24	-0.10(-0.30,0.08)	-0.30(-0.59,-0.03)	0.229	-0.56(-1.20,0.08)	0.169	0.467
	25-30	-0.10(-0.30,0.10)	-0.13(-0.40,0.13)	0.828	0.12(-0.54,0.75)	0.548	0.495
2-17°C	1-30	-0.05(-0.07,-0.03)	-0.08(-0.11,-0.05)	0.126	0.00(-0.07,0.07)	0.165	0.040
	1-6	-0.13(-0.20,-0.10)	-0.15(-0.24,-0.07)	0.928	-0.08(-0.27,0.10)	0.549	0.536
	7-12	-0.10(-0.17,-0.05)	-0.10(-0.18,-0.02)	0.958	0.05(-0.15,0.23)	0.177	0.186
	13-18	-0.02(-0.08,0.03)	-0.07(-0.15,0.02)	0.399	-0.03(-0.22,0.15)	0.957	0.726
	19-24	0.00(-0.07,0.05)	-0.07(-0.15,0.00)	0.152	0.15(-0.05,0.33)	0.150	0.039
	25-30	0.02(-0.05,0.07)	-0.02(-0.10,0.05)	0.439	-0.08(-0.29,0.10)	0.291	0.519
>17°C	1-30	-0.21(-0.37,-0.04)	-0.07(-0.31,0.16)	0.350	-0.36(-0.90,0.18)	0.545	0.290
	1-6	-0.29(-0.59,0.02)	-0.24(-0.68,0.20)	0.858	-0.80(-1.85,0.25)	0.331	0.308
	7-12	-0.20(-0.51,0.10)	-0.03(-0.47,0.40)	0.534	-0.35(-1.38,0.69)	0.781	0.573
	13-18	-0.24(-0.54,0.07)	-0.07(-0.52,0.38)	0.549	-0.37(-1.42,0.67)	0.799	0.596
	19-24	-0.07(-0.37,0.23)	0.05(-0.39,0.49)	0.651	0.12(-0.90,1.13)	0.739	0.919
	25-30	-0.25(-0.58,0.03)	-0.13(-0.58,0.31)	0.631	-0.51(-1.54,0.53)	0.639	0.495

a. Percentage change in Relative Risk (95% Confidence interval) associated with 1oC increase in temperature on an single day within the lag period

vZero. p-value associated with z-test comparison with zero days in hospital

v1-29. p-value associated with z-test comparison with 1-29 days in hospital

In summary, across the three causes of death clear patterns of variation according to hospital status during admission were most obvious for cold temperatures in the respiratory causes of death. Zero days in hospital i.e. the subject was community based only, revealed the strongest decrease in risk as temperature increased by 1°C within the cold temperature zone. Subjects who spent 1-29 exposure days in hospital also reported a decrease in risk than those who were there for zero days, though with a weaker effect magnitude. Significant differences were less likely in the longer lag periods (19-30 days) due to the effects observed in 'zero' days in hospital group dissipating faster, or a delay in peak risk occurring in those in hospital for at least one day. In these cold temperatures percentage relative risk tended towards the null effect, or showed a non-significant increase in risk when subjects had spent 'all 30' days in hospital i.e. hospital only based subjects.

Ischaemic heart disease appeared to be much less affected by the subject's location i.e. any increase in cold temperatures at any lag had a relatively consistent effect regardless of the subject location. It may be that a 6 day lag length is too wide, as indicated by Figure 7.2, and is masking sudden changes in relative risk. To briefly explore this, a small set of analyses were performed on those subjects with zero days in hospital only. It was still felt they were more likely to provide a greater chance of a true relationship between temperature and IHD mortality. The 1-6 day lag period was split into two and then three shorter lag periods of equal length, followed by all 1-6 days analysed as daily lags 1,2,3, and then 3-6 days. In each of these lag period combinations similar results to Table 7.5 were found, with a decrease or no change in risk when temperature increased above 17°C. The final analysis included same day (lag 0) in the model. In this case the warm temperatures (>17°C) displayed a borderline or better significant increase in risk as temperature increased by one degree all subjects %RR (95% C.I.) = 1.20% (0.23%,2.18%) and those with zero exposure days in hospital 1.18% (-0.01%,2.38%). In all subjects a small and non-significant increase in risk was also observed for a one degree increase (0.06% (-0.14%,0.27%) within the mild temperatures 2-17°C.

8 RESULTS – MAIN STUDY ANALYSIS OF POLLUTION

The following reports the results of the main analysis of this study. Involving the complete empirical dataset (Jan 1980 to Dec 2011), it aims to model the influence of five pollutants on the three specific causes of death as defined by the ‘any’ cause of death field. The analysis then attempts to improve the accuracy of any pollution-mortality relationship by investigating bias due to subject location during exposure, missing data, and influential outliers. Note, all fitted models adjust for temperature parameters identified for each specific cause of death in Chapter 7 that matches the model structure reported i.e. lag stratified models are adjusted by temperature in the lag stratified format.

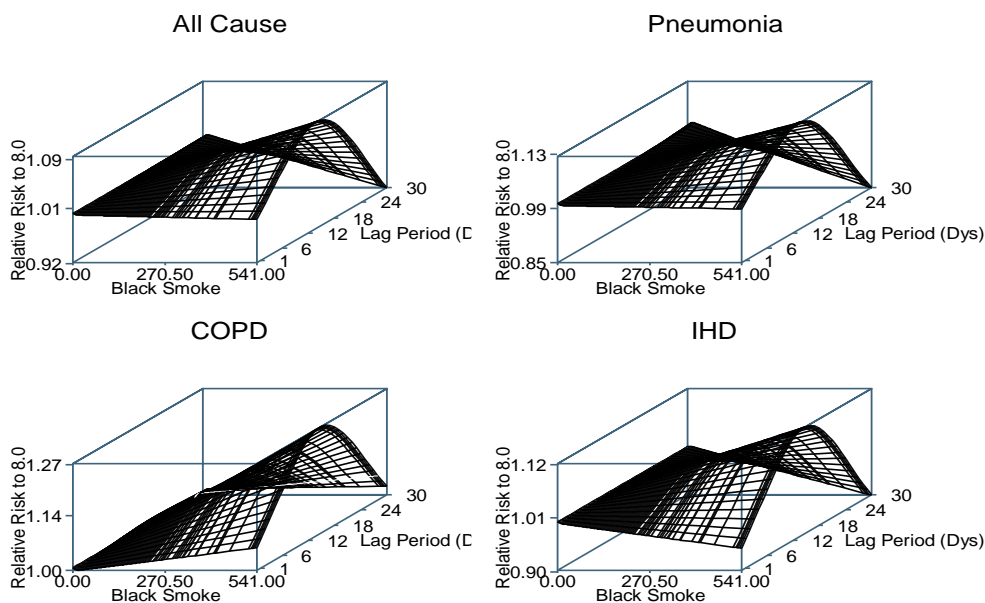
8.1 Investigating pollution–mortality: Pneumonia, COPD, and IHD

Analysis of the particulate (BS PM10, and PM2.5) and gaseous pollutants (SO2 and NO2) followed a similar pattern to the one described for temperature (see Chapter 7). Initially distributed lag non-linear models using natural cubic splines with differing combinations of the number of knot points applied to the exposure range and to the lag length were generated. Each combination of exposure lag knot numbers were fitted one by one and the most optimum combination for each cause of death was identified using the AIC value. In all but two cases, the cause specific deaths identified a combination of 3 knots points for exposure and 3 for the lag period. The exceptions occurred for PM2.5 - COPD and SO2 – pneumonia which required an exposure:lag combination of 3:4 and 4:3 respectively. Figures 8.1, 8.2, and 8.3 illustrate the relative risk observed simultaneously across the exposure and lag period when compared to the rounded pollutant exposure mean at lag 1. The knot combinations per pollutant exposure and lag for each AFCOD are also provided.

Plotting the relative risk produced by the distributed lag non-linear model indicates the presence of a linear exposure-response relationship, not only immediately after exposure, but also at longer lag periods extending to 30 days. Figure 8.1 plots the relative risk associated with the cause of death identifiers due to a change in black

smoke air pollution. Note, though not a main objective, the ‘all-causes’ of death result was shown for completeness. The change in risk for ‘all causes’ appears to be driven by the relationship observed in IHD; the cause of death with the largest sample size. Of the causes specific analysis, all three report a linear exposure-response that persists across the complete 30 day lag period. The change in risk is a similar pattern in all three causes of death, with a gradual increase in risk over the first week peaking between 12 to 18 days before a gradual return towards baseline. Increased risk of COPD mortality began immediately (at lag 1) and went on to show the strongest increase in risk, with peak relative risk of 1.27 at maximum BS compared to 1.13 and 1.12 in pneumonia and IHD respectively. In contrast to the earlier analysis (Chapter 5) the risk of pneumonia mortality associated with BS (lag 1) rose more gradually in the first 6 days. Ischaemic heart disease also showed a delay with even a slight decrease in risk for the first 1-2 days before rising similarly to the other causes of death.

Figure 8.1 – Simultaneously plotting across the ‘black smoke’ range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields in a distributed lag non-linear model.



Knot no. combinations (pollution:lag) for All Cause(3:3), pneumonia (3:3), COPD (3:3), IHD (3:3)
 Relative Risk comparable to lag 1 at approximate mean pollution ($8.0\mu\text{gm}^{-3}$)

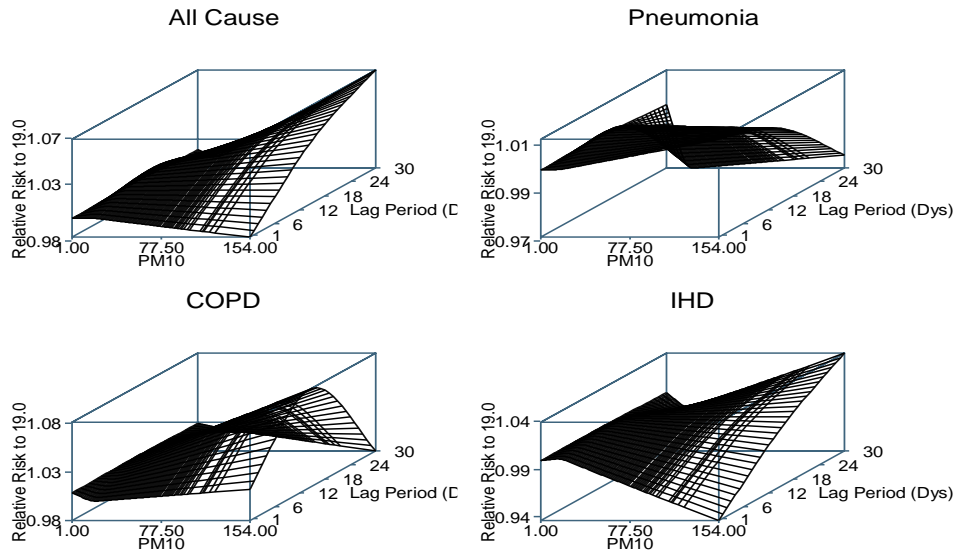
The size specific particulate matter fractions (PM10 & PM2.5) were investigated and reported in Figure 8.2. Though results were similar across the two particulate sizes, a

COD specific set of relationships were observed, unlike BS. In both PM10 and PM2.5 the effect on pneumonia was minimal, though a small increase in risk was observed that quickly dissipated. Of the three causes of death, chronic obstructive pulmonary disease again reported the strongest effect sizes with a delayed peak at a relative risk of 1.08 and 2.14 for maximum PM10 and PM2.5 exposure. Similar to pneumonia, the relative risk observed in ischaemic heart disease was weaker than for black smoke. Risk was observed to increase continuously over the lag period. A much longer delay between initial exposure and peak risk in IHD mortality may exist here, before returning to baseline. In all three cause of death, the effect sizes were observed to be much larger for PM2.5 than for PM10. This was particularly apparent in COPD, but also occurred to a lesser extent in Pneumonia and IHD.

The gaseous pollutants, sulphur dioxide and nitrogen dioxide, are reported in Figure 8.3 and contrast with the particulate pollutants. Previously the particulates had the strongest association with COPD whereas the gaseous pollutants, particularly SO₂, appear to have the greatest effect on risk of mortality in pneumonia and IHD. In both SO₂ and NO₂ the pattern of change in risk follows that described for BS on pneumonia and IHD. Except here SO₂ risk was greater, peaking at ≈ 1.18 relative risk for max SO₂ exposure at a later point in the lag period (18-24 days). SO₂ also appears to have a greater effect (≈ 1.26) on IHD than any other pollutants reported here. The effect on COPD was a comparatively mild but gradual increase in risk over the entire 30 days with peak risk appearing to occur after 30 days. The nitrogen dioxide-lagged mortality relationship was similar in shape to BS for each COD respectively. A gradual increase in risk occurred, peaking between 12 and 18 days before returning to baseline.

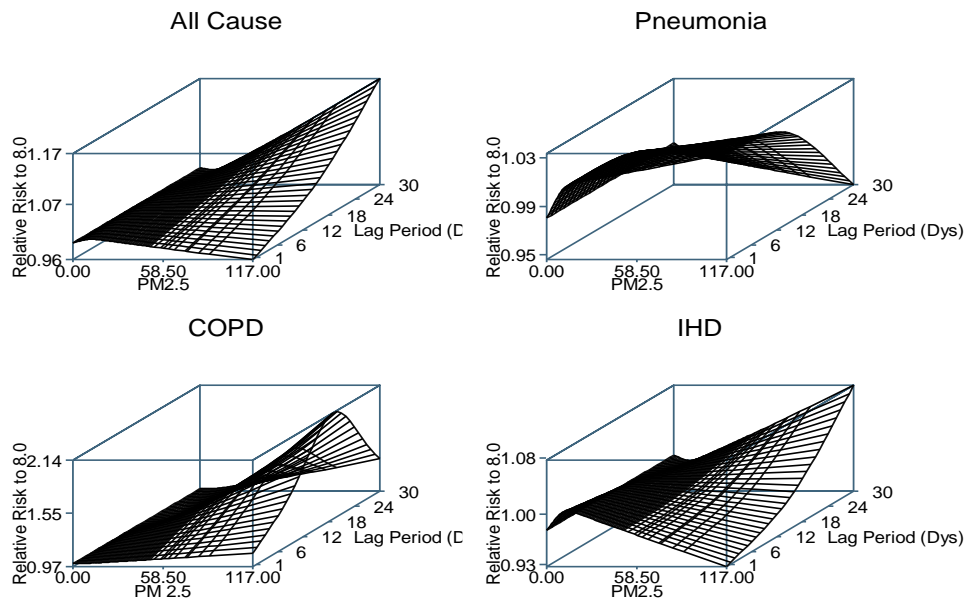
Figure 8.2 – Simultaneously plotting across the ‘Particulate Matter’ range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields in a distributed lag non-linear model

Particulate Matter 10 ($\mu\text{g}\text{m}^{-3}$) – PM10



Knot no. combinations (pollution:lag) for All Cause(3:3), pneumonia (3:3), COPD (3:3), IHD (3:3)
 Relative Risk comparable to lag 1 at approximate mean pollution ($19.0\mu\text{g}\text{m}^{-3}$)

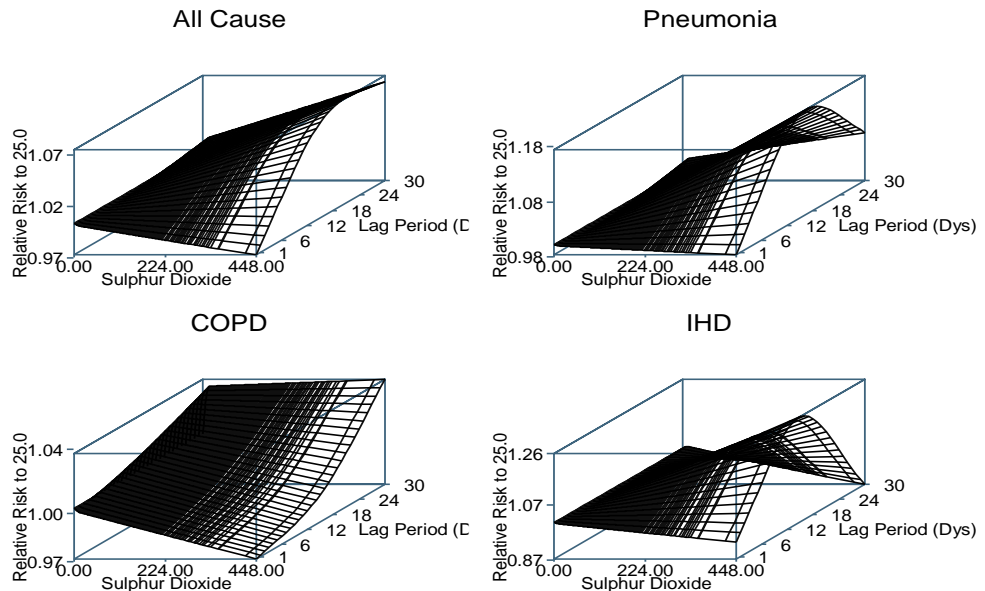
Particulate Matter 2.5 ($\mu\text{g}\text{m}^{-3}$) – PM2.5



Knot no. combinations (pollution:lag) for All Cause(3:3), pneumonia (3:3), COPD (3:4), IHD (3:3)
 Relative Risk comparable to lag 1 at approximate mean pollution ($8.0\mu\text{g}\text{m}^{-3}$)

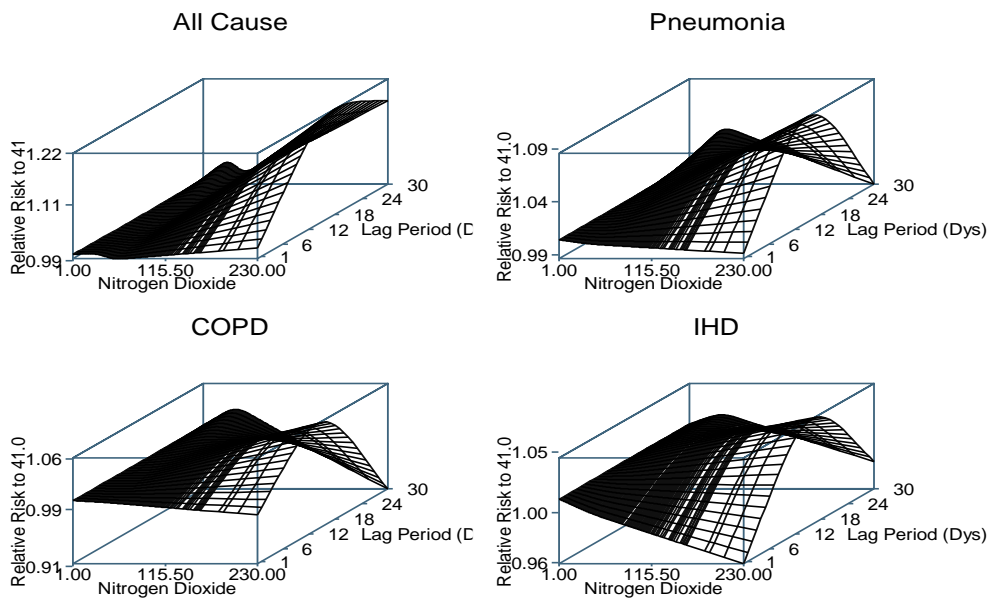
Figure 8.3 – Simultaneously plotting across the ‘Gaseous Pollutants’ range and lag period the change in risk for all-cause, and pneumonia, COPD, IHD mortality identified by any cause of death fields in a distributed lag non-linear model

Sulphur Dioxide ($\mu\text{g}\text{m}^{-3}$) – SO₂



Knot no. combinations (pollution:lag) for All Cause(3:3), pneumonia (4:3), COPD (3:3), IHD (3:3)
Relative Risk comparable to lag 1 at approximate mean pollution ($25.0\mu\text{g}\text{m}^{-3}$)

Nitrogen Dioxide ($\mu\text{g}\text{m}^{-3}$) – NO₂



Knot no. combinations (pollution:lag) for All Cause(4:3), pneumonia (3:3), COPD (3:3), IHD (3:3)
Relative Risk comparable to lag 1 at approximate mean pollution ($41.0\mu\text{g}\text{m}^{-3}$)

8.2 Modelling lagged pollution effect: Pneumonia, COPD, and IHD

A linear exposure-response relationship was assumed appropriate based on Figure 8.1, 8.2, and 8.3. The lag stratified model was then applied to the data with a single continuous linear variable representing the entire exposure range for each pollutant within each lag period. Table 8.1 reports for pneumonia, COPD and IHD the percentage relative risk associated with an increase in each pollutant of $10\mu\text{gm}^{-3}$ on any single day within the lag period. The analysis in Table 8.1 reports the results when cause of death is defined by 'any' cause of death field. To compare the three causes of death, a three-way categorical variable representing pneumonia, COPD and IHD was generated based on the PCOD ('primary' cause of death see Chapter 7.3 for more details). The first principles comparison test between for each lag period was performed and the p-values reported in Table 8.1 when compared against pneumonia, and COPD. The corresponding full results, %RR and 95% confidence intervals, relating to the PCOD only are available in Appendix D.

The largest increase in percentage relative risk due to an increase in particulate pollution occurred for COPD mortality. The three particulate pollutants (BS, PM10, and PM2.5) and COPD reported %RR of 0.08%, 0.15%, 0.47% for the 30 day lag, respectively. A $10\mu\text{gm}^{-3}$ increase in daily average PM2.5 related to the largest %RR (95% C.I) in COPD mortality of 1.05% (0.14%,2.01%) at lag 1-6 days. Note, given PM2.5 exposure interquartile range is distinct from BS and PM10 the effect on the population experiencing an IQR increase would relate here to 0.63% (0.08%,1.20%). Increased risk of pneumonia and IHD mortality only occurred after a delay and were weaker than those seen in COPD. In all three particulate pollutants IHD again showed a decrease in risk (%RR -0.13%, -0.42%, and -0.16%, respectively) for the immediate lag (1-6 days). Comparing all three particulate pollutants, the magnitude of the effect estimates was greatest when associated with PM2.5 regardless of the cause of death. However, confidence intervals were also observed to be wider reflecting the smaller sample sizes in the PM2.5 analysis.

Table 8.1 – Cause of death specific percentage relative risk (%RR) associated with 10µgm⁻³ increase per pollutant reported by lag stratified analysis, with comparison test result between three causes of death by lag period.

Lag Period/ Pollutant	Any Cause of Death - %RR (95%.C.I.) ^a					
	Pneumonia	COPD	vPneu	IHD	vPneu	vCOPD
Black Smoke (per 10 µgm⁻³ increase, note IQR = 10)						
30 Days	0.02(-0.05,0.08)	0.08(-0.04,0.21)	0.393	-0.02(-0.07,0.04)	0.459	0.175
1-6 Days	-0.09(-0.21,0.04)	-0.01(-0.31,0.30)	0.636	-0.13(-0.34,0.09)	0.728	0.518
7-12 Days	-0.01(-0.16,0.13)	-0.04(-0.23,0.15)	0.814	-0.10(-0.25,0.04)	0.379	0.610
13-18 Days	0.02(-0.17,0.21)	0.14(-0.15,0.43)	0.504	0.07(-0.03,0.18)	0.621	0.688
19-24 Days	-0.11(-0.24,0.02)	0.21(-0.12,0.55)	0.074	0.02(-0.08,0.12)	0.115	0.276
25-30 Days	0.03(-0.10,0.15)	0.18(0.00,0.37)	0.170	-0.03(-0.15,0.09)	0.520	0.058
PM 10 (per 10 µgm⁻³ increase, note IQR = 11)						
30 Days	0.01(-0.09,0.12)	0.15(-0.05,0.36)	0.254	-0.08(-0.17,0.02)	0.215	0.045
1-6 Days	0.11(-0.13,0.34)	-0.01(-0.32,0.31)	0.571	-0.42(-0.61,-0.23)	0.001	0.029
7-12 Days	-0.19(-0.43,0.05)	0.29(-0.18,0.77)	0.074	-0.08(-0.40,0.23)	0.592	0.200
13-18 Days	0.03(-0.20,0.27)	0.05(-0.26,0.38)	0.923	0.03(-0.16,0.23)	0.997	0.916
19-24 Days	-0.06(-0.36,0.25)	0.07(-0.24,0.40)	0.565	-0.12(-0.32,0.07)	0.714	0.300
25-30 Days	0.07(-0.17,0.30)	0.16(-0.17,0.50)	0.650	0.20(-0.07,0.47)	0.469	0.867
PM 2.5 (per 10 µgm⁻³ increase, note IQR = 6)						
30 Days	0.27(-0.07,0.66)	0.47(-0.27,1.38)	0.669	-0.15(-0.79,0.65)	0.316	0.262
1-6 Days	0.36(-0.35,1.11)	1.05(0.14,2.01)	0.256	-0.16(-0.75,0.46)	0.283	0.032
7-12 Days	0.28(-0.58,1.19)	0.08(-1.59,1.95)	0.844	0.38(-0.67,1.51)	0.889	0.779
13-18 Days	-0.19(-1.08,0.74)	0.32(-0.53,1.20)	0.423	0.23(-0.35,0.82)	0.445	0.866
19-24 Days	0.32(-0.36,1.03)	0.71(-0.13,1.59)	0.488	-0.27(-0.85,0.32)	0.196	0.061
25-30 Days	-0.56(-2.05,1.09)	-0.24(-1.55,1.19)	0.766	0.38(-0.36,1.15)	0.300	0.442
Sulphur Dioxide (per 10 µgm⁻³ increase, note IQR = 17)						
30 Days	0.03(-0.01,0.08)	0.00(-0.07,0.06)	0.364	-0.03(-0.06,0.01)	0.038	0.521
1-6 Days	0.03(-0.08,0.14)	0.03(-0.19,0.26)	0.973	-0.13(-0.28,0.03)	0.117	0.264
7-12 Days	-0.03(-0.14,0.08)	-0.02(-0.25,0.21)	0.931	-0.03(-0.11,0.06)	0.954	0.954
13-18 Days	0.10(0.00,0.21)	0.09(-0.18,0.37)	0.931	0.06(-0.03,0.14)	0.504	0.816
19-24 Days	0.01(-0.09,0.12)	0.06(-0.09,0.22)	0.616	0.00(-0.12,0.11)	0.839	0.515
25-30 Days	0.08(-0.03,0.19)	-0.02(-0.18,0.14)	0.306	-0.07(-0.19,0.04)	0.057	0.601
Nitrogen Dioxide (per 10 µgm⁻³ increase, note IQR = 21)						
30 Days	0.02(-0.05,0.08)	0.12(-0.05,0.30)	0.291	-0.04(-0.15,0.06)	0.341	0.119
1-6 Days	-0.24(-0.59,0.12)	-0.24(-0.64,0.17)	0.998	-0.27(-0.53,-0.01)	0.885	0.897
7-12 Days	-0.08(-0.23,0.06)	0.07(-0.14,0.27)	0.237	-0.10(-0.22,0.02)	0.832	0.155
13-18 Days	0.08(-0.11,0.27)	0.23(-0.11,0.58)	0.447	0.06(-0.06,0.19)	0.914	0.373
19-24 Days	-0.03(-0.24,0.18)	0.23(-0.18,0.66)	0.262	0.08(-0.14,0.31)	0.452	0.533
25-30 Days	0.12(-0.02,0.26)	0.15(-0.05,0.34)	0.822	0.03(-0.09,0.15)	0.332	0.306

a. Percentage Relative Risk (95% Confidence Interval) per 10µgm⁻³ increase of pollutant on any single day within the lag period.

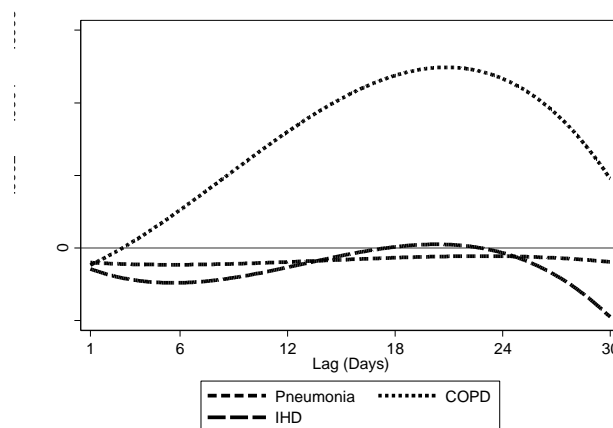
vPneu. p-value associated with z-test comparison with Pneumonia

vCOPD. p-value associated with z-test comparison with COPD

Any change in risk associated with gaseous pollutants tended to be observed after 12 days. Here SO₂ and NO₂ concentrations were associated with pneumonia, COPD, and IHD mortality the strongest between 13-18 days post exposure (SO₂ %RR (95% C.I) = 0.10% (0.00%,0.21%), 0.09% (-0.18%,0.37%), and 0.06%(-0.03%,0.14%), NO₂ %RR (95% C.I) = 0.08% (-0.11%,0.27%), 0.23% (-0.11%,0.58%), and 0.06%(-0.06%,0.19%)). Nitrogen dioxide in all three causes of death displayed a strong initially decreasing %RR that then increased to a peak risk in the 13-18 day lag. This was strongest in COPD with a %RR of 0.23%(-0.11%,0.58%) or if adjusting for the IQR of 21µgm⁻³ 0.48% (-0.23%,1.22%). Whereas SO₂ peaked for pneumonia at 0.10% (0.00%,0.21%), which for an IQR increase of 17µgm⁻³ gives 0.17% (0.00%,0.36%).

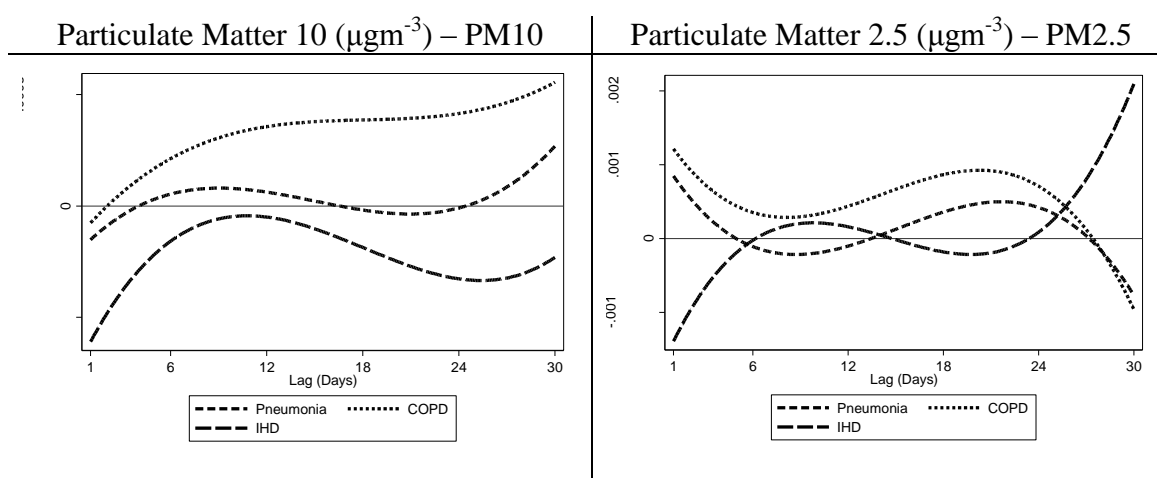
The association with the pollutant concentration on cause of death was then studied with a cubic polynomial function in the distributed lag model in order to observe any, if present, change in risk was occurring over shorter lag lengths. The cubic function allows for a flexible but smooth change in risk over the lag period. Using the parameter effect estimates, Figure 8.4, 8.5, and 8.6 are plots of the cause specific change in logged rate ratio across the lag period for black smoke, the particulates (PM10 and PM2.5), and the gaseous pollutants (SO₂ and NO₂) respectively. As with the lag stratified %RR, all plots represents the change in risk when pollutant concentration increases by 10µgm⁻³. These plots allow for an easy comparison of the cause specific change in risk over the lag period.

Figure 8.4 – Plotting the change in mortality risk described by a cubic distributed lag model associated with a 10µgm⁻³ increase in black smoke on pneumonia, COPD, and IHD mortality reported in any cause of death fields.



For all three particulate metrics the influence on risk of COPD mortality was again observed to be the strongest of the three causes of death, and of the three pollutants, PM2.5 was most strongly associated with all three causes of death. Of the three particulate metrics only PM2.5 showed slight increase risk for IHD occurring between 6 and 12 days post exposure. Both PM10 and PM2.5 displayed slight increases in risk for pneumonia, with for PM2.5 it was immediate (<6 days) whereas for PM10 there was a gradual increase that lasted approximately 15 days into the lag. In some cases a strong sudden increase in risk occurred towards the end of the lag (>25 days), this is more likely due to forced cubic polynomial shape that accounted for earlier changes in the lag period, and not a true effect. Ischaemic heart disease appears to not be associated by PM10 concentrations. The decrease in risk for IHD (BS & PM10 - Figure 8.4 & 8.5) persists even when greater flexibility would allow for a change in direction within lag lengths of two or three days.

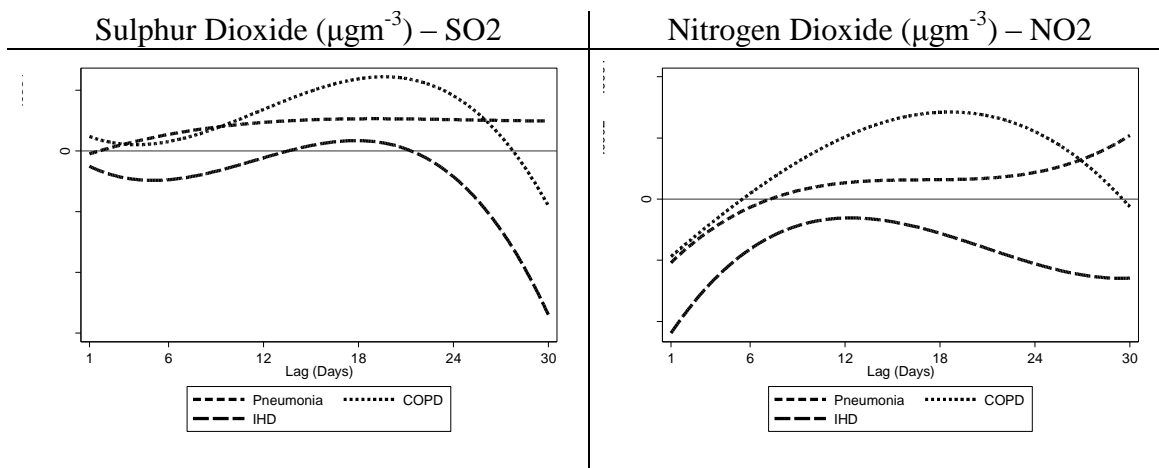
Figure 8.5 - Plotting the change in mortality risk described by a cubic distributed lag model associated with a $10\mu\text{g}\text{m}^{-3}$ increase in particulate matter concentrations on pneumonia, COPD, and IHD mortality reported in any cause of death fields.



Even when the effect sizes associated with gaseous pollutants were weaker (Figure 8.6) than the effects associated with particulates exposure, there appeared to be a greater influence on COPD compared to pneumonia and IHD. In contrast to the lag stratified model the greater COPD association is also present in SO₂. The risk associated with sulphur dioxide was the smallest of all pollutants covered. However, an increased risk was observed for both COPD and pneumonia across the entire 30 day lag and a small increase for IHD between 12 and 24 days. The curve displayed for SO₂ at lags 1-6

reports a slight incline upwards towards lag 1, indicating that a possible increase in risk occurred at the same day lag 0, though the shape may just be representing the forced cubic curve.

Figure 8.6 - Plotting change in mortality risk described by a cubic distributed lag model associated with a $10\mu\text{g}\text{m}^{-3}$ increase in gaseous pollutant concentrations on pneumonia, COPD, and IHD mortality reported in any cause of death fields.



Nitrogen dioxide, for all three causes of death indicated a decrease in risk occurred during the short-term lags (<6 days). In the two respiratory diseases this rose after 6 days since exposure and became an increase in risk as the pollution increased. Ischaemic heart disease corresponds to a decrease in risk as NO2 increased, that even though it approaches the baseline risk, it was persistent across the entire 30 days post exposure.

8.3 Accounting for hospital admission during pollution exposure

To gain the most direct relationship between ambient exposure and the subject, the analysis was repeated whilst stratifying by the hospital admission location during exposure i.e. 'zero', '1-29', or 'all 30' days in hospital. The results are reported in Table 8.2, 8.3, and 8.4 for pneumonia, COPD, and IHD respectively. Figure 8.7, 8.8, and 8.9 (see Figure D1-D3 Appendix D for gaseous pollutants) plot the relationship identified in the cubic distributed lag models. The p-values associated with the first principles comparison test of location status during exposure indicated if effects are significantly different from zero days, and 1-29 days in hospital. For ease of comparison, each table also reports the results from the main analysis (Table 8.1).

8.3.1 Pollution-pneumonia mortality: accounting for hospital status.

Some evidence was present for an increase in risk due to particulates caused pneumonia mortality in community based subjects (Table 8.2 and Figure 8.7). In all three particulate pollutants, the lag 1-30 indicated a greater increased risk in the 'zero' days than any day ('1-29' or 'all 30' days) spent in hospital. Though significant differences were few with predominantly BS %RR (95% C.I.) at 7-12 days -0.28% (-0.61%,0.05%) compared to 'zero' days, and PM_{2.5} 0.92% (0.25%,1.70%) compared to '1-29' and 'all 30' days being significantly different. The cubic function (Figure 8.7) indicated a greater increase in risk per increase in pollutant concentration during short lags for PM₁₀. These were superseded by those in hospital for all at least 1 day ('1-29' & 'all 30' days). The reverse was present for fine particulate matter (PM_{2.5}) where 'zero' and '30 days' in hospital had an immediate risk that lasted for 12 days, but in the community based subjects it was only greater after 19 days. The gaseous pollutants, SO₂ and NO₂, showed no significant differences in pneumonia risk between the three hospital statuses. An increase of 10µgm⁻³ of daily average SO₂ on any day within the 1-6 day lag period corresponded to a %RR increase in pneumonia mortality of 0.11% (-0.20%, 0.42%) in those in hospital for all 30 days. Whereas those with zero days exposure spent in hospital reported a %RR of 0.05% (-0.09%, 0.18%). Similarly for NO₂, with zero days in hospital corresponding to a slight decrease in risk until 25-30 days later, whereas 30 days in hospital reported an increase risk after 13-18 days.

Table 8.2 – Percentage relative risk (%RR) associated with an increase in 10µgm⁻³ within each pollutant associated with the lag stratified analysis for Pneumonia, split by hospital admission status during exposure (Zero, 1-29, and All 30 days).

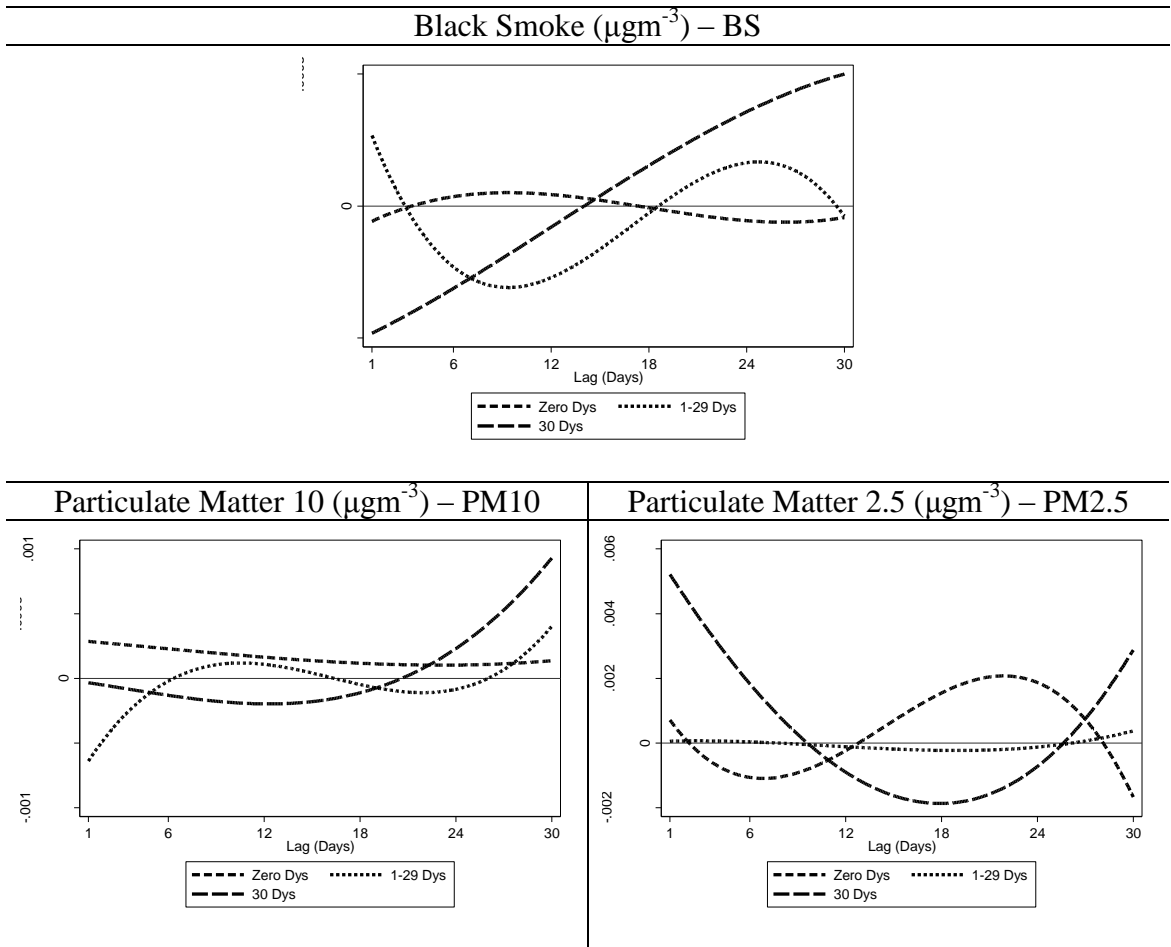
Lag (Dys)	Pneumonia %RR (95%.C.I.) ^a - By Hospital Admission Status						
	All Subjects	Zero	1-29 Days	P-val vZero	All 30 Days	P-val vZero	P-val v1-29d
Black Smoke (per 10 µgm-3 increase, note IQR = 10)							
30	0.02(-0.05,0.08)	0.02(-0.04,0.08)	0.01(-0.10,0.13)	0.944	-0.03(-0.18,0.12)	0.538	0.632
1-6	-0.09(-0.21,0.04)	-0.06(-0.25,0.14)	-0.06(-0.32,0.20)	0.960	-0.17(-0.55,0.21)	0.599	0.653
7-12	-0.01(-0.16,0.13)	0.09(-0.06,0.25)	-0.28(-0.61,0.05)	0.043	0.02(-0.41,0.45)	0.746	0.273
13-18	0.02(-0.17,0.21)	0.10(-0.13,0.32)	-0.23(-0.70,0.25)	0.229	0.07(-0.51,0.67)	0.946	0.430
19-24	-0.11(-0.24,0.02)	-0.15(-0.30,0.00)	0.06(-0.27,0.40)	0.256	-0.11(-0.55,0.35)	0.851	0.560
25-30	0.03(-0.10,0.15)	-0.03(-0.18,0.12)	0.16(-0.17,0.51)	0.299	0.05(-0.34,0.45)	0.713	0.661
PM 10 (per 10 µgm-3 increase, note IQR = 11)							
30	0.01(-0.09,0.12)	0.04(-0.12,0.21)	0.00(-0.17,0.17)	0.717	-0.02(-0.31,0.30)	0.722	0.912
1-6	0.11(-0.13,0.34)	0.19(-0.39,0.79)	-0.02(-0.38,0.36)	0.557	-0.07(-0.72,0.62)	0.571	0.899
7-12	-0.19(-0.43,0.05)	-0.12(-0.46,0.22)	-0.29(-0.66,0.09)	0.536	-0.16(-0.83,0.53)	0.923	0.755
13-18	0.03(-0.20,0.27)	-0.06(-0.39,0.29)	0.25(-0.12,0.63)	0.238	-0.24(-0.89,0.44)	0.636	0.216
19-24	-0.06(-0.36,0.25)	0.09(-0.25,0.43)	-0.23(-0.84,0.41)	0.387	0.20(-0.46,0.89)	0.771	0.362
25-30	0.07(-0.17,0.30)	-0.05(-0.39,0.29)	0.16(-0.21,0.53)	0.421	0.29(-0.38,0.98)	0.377	0.734
PM 2.5 (per 10 µgm-3 increase, note IQR = 6)							
30	0.27(-0.07,0.66)	0.92(0.25,1.70)	-0.03(-0.46,0.46)	0.024	-0.07(-1.17,1.58)	0.241	0.953
1-6	0.36(-0.35,1.11)	0.35(-0.86,1.66)	0.15(-0.82,1.18)	0.807	2.33(-0.51,5.68)	0.226	0.167
7-12	0.28(-0.58,1.19)	0.87(-0.33,2.16)	0.00(-0.92,0.96)	0.268	1.15(-3.84,8.07)	0.927	0.694
13-18	-0.19(-1.08,0.74)	-0.06(-1.64,1.67)	0.02(-1.61,1.83)	0.945	-1.42(-3.75,1.32)	0.386	0.363
19-24	0.32(-0.36,1.03)	1.51(0.33,2.76)	-0.35(-1.25,0.60)	0.016	-1.60(-3.92,1.13)	0.041	0.373
25-30	-0.56(-2.05,1.09)	-0.99(-2.91,1.2)	-0.60(-2.46,1.50)	0.790	0.90(-1.62,3.85)	0.271	0.376
Sulphur Dioxide (per 10 µgm-3 increase, note IQR = 17)							
30	0.03(-0.01,0.08)	0.05(-0.01,0.10)	-0.02(-0.13,0.09)	0.307	0.05(-0.10,0.21)	0.967	0.477
1-6	0.03(-0.08,0.14)	0.05(-0.09,0.18)	-0.05(-0.30,0.20)	0.507	0.11(-0.20,0.42)	0.715	0.434
7-12	-0.03(-0.14,0.08)	-0.04(-0.17,0.10)	0.04(-0.20,0.29)	0.567	-0.10(-0.41,0.22)	0.718	0.481
13-18	0.10(0.00,0.21)	0.09(-0.04,0.22)	0.06(-0.24,0.37)	0.888	0.22(-0.11,0.54)	0.473	0.502
19-24	0.01(-0.09,0.12)	0.04(-0.11,0.19)	0.01(-0.22,0.25)	0.854	-0.15(-0.45,0.16)	0.273	0.402
25-30	0.08(-0.03,0.19)	0.06(-0.07,0.19)	0.09(-0.14,0.33)	0.821	0.15(-0.22,0.53)	0.659	0.796
Nitrogen Dioxide (per 10 µgm-3 increase, note IQR = 21)							
30	0.02(-0.05,0.08)	-0.02(-0.18,0.14)	0.06(-0.11,0.24)	0.481	-0.03(-0.22,0.18)	0.985	0.521
1-6	-0.24(-0.59,0.12)	-0.23(-0.56,0.11)	-0.21(-0.81,0.41)	0.954	-0.32(-0.74,0.10)	0.728	0.760
7-12	-0.08(-0.23,0.06)	-0.05(-0.25,0.14)	-0.13(-0.37,0.12)	0.648	-0.12(-0.56,0.32)	0.774	0.989
13-18	0.08(-0.11,0.27)	-0.06(-0.51,0.42)	0.16(-0.09,0.41)	0.431	0.32(-0.31,0.98)	0.353	0.645
19-24	-0.03(-0.24,0.18)	-0.04(-0.23,0.16)	-0.05(-0.30,0.19)	0.919	-0.23(-0.76,0.31)	0.507	0.556
25-30	0.12(-0.02,0.26)	0.04(-0.15,0.23)	0.25(0.01,0.49)	0.180	0.11(-0.30,0.54)	0.736	0.597

a. Percentage Relative Risk (95% Confidence Interval) per 10µgm-3 increase of pollutant on any single day within the lag period.

vZero. p-value associated with z-test comparison with zero days in hospital

v1-29. p-value associated with z-test comparison with 1-29 days in hospital

Figure 8.7 – Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on pneumonia split by hospital admission during exposure (zero, 1-29, and all 30 days).



8.3.2 Pollution-COPD mortality: accounting for hospital status.

Over the 30 day lagged average, both particulates fraction PM10 and PM2.5 indicated a greater increase in risk of COPD mortality (Table 8.3) than in those with less than 30 days in hospital (PM10 = 0.21%/0.19% and PM2.5 = 0.55%/0.53% for zero/1-29 hospital days). For black smoke those who spent all 30 days in hospital were more at risk of mortality. Within the 30 day lag period the coarse particulates (BS & PM10) were comparable, both reported in all three hospital admission status a delay before a strong increase risk was observed between 7-18 day lag. Figure 8.8 also shows the effect estimates themselves are similar across the 30 days, with only those in hospital for all 30 days showing a greater magnitude. Note, the sample size for 'all 30' days was smaller than the other groups, contributing to the effect estimates. Though they followed the same general pattern given the effect sizes and wide confidence intervals for PM2.5 the sample size appears to be strongly influencing the 'all 30' day in hospital effect estimates. If we compare community only and 1-29 days in hospital vs, zero the effect estimates in PM2.5 appear to be slightly greater. With a positive %RR in all but the 25-30 day lag, the increase risk also appears to continue longer into the lag period. PM2.5 and COPD mortality relationship in Figure 8.8 indicates a similar relationship between 'zero' days and '1-29' days in hospital during exposure. Those with 'all 30' days displayed the same pattern as PM10 in PM2.5 except with a greater effect magnitude.

Smaller effect sizes, though similar patterns, were observed for the two gaseous pollutants (Figure D3 Appendix D). In both cases the community based subjects reported a delay before a period (between 12-30 lag) of risk stronger than the other hospital admission groups. This may indicate that in a chronic condition those in the community are more robust to gaseous pollutants requiring a delay before they can have an impact.

Figure 8.8 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on COPD split by hospital admission during exposure (zero, 1-29, and all 30 days).

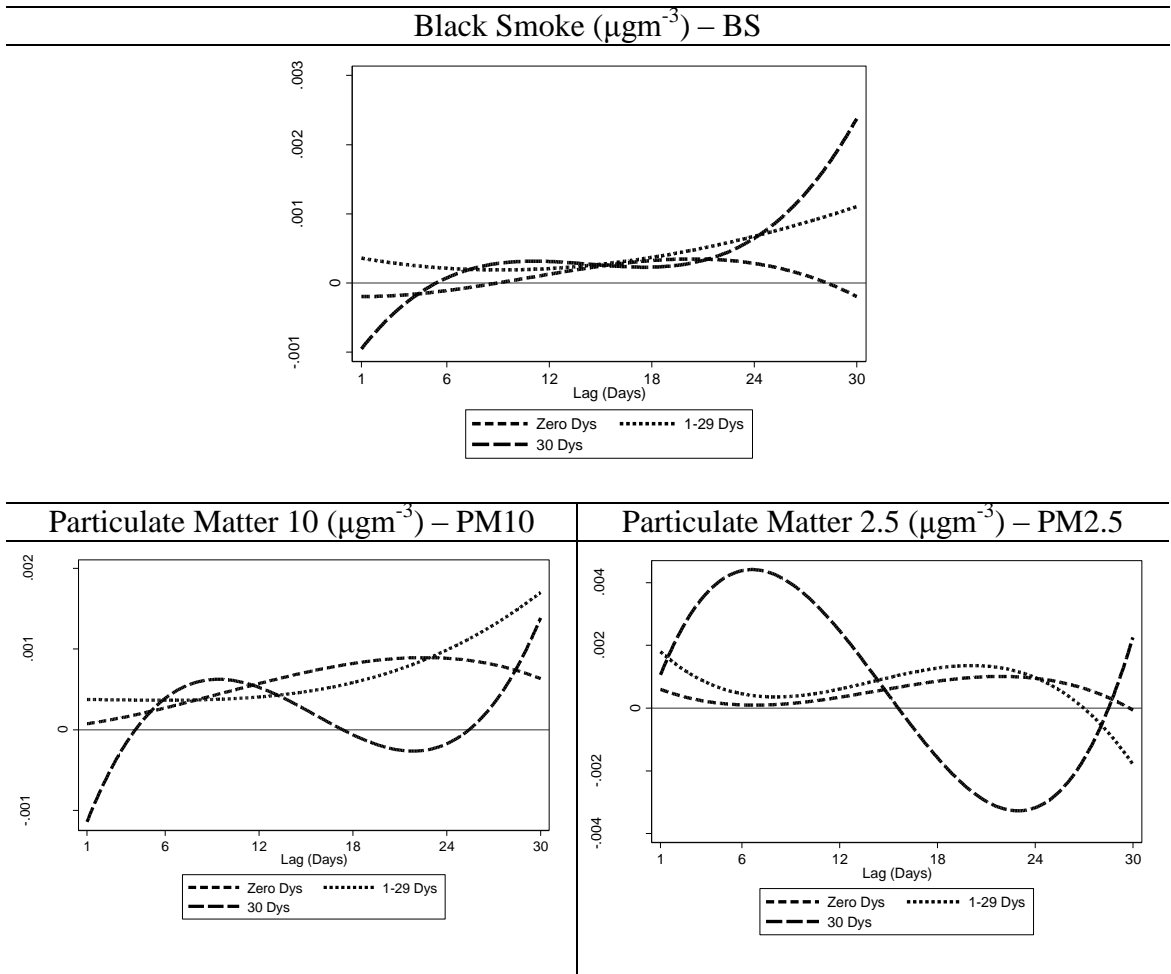


Table 8.3 - Percentage relative risk (%RR) associated with an increase in 10µgm⁻³ within each pollutant associated with the lag stratified analysis for COPD, split by hospital admission status during exposure (Zero, 1-29, and All 30 days).

Lag (Dys)	COPD %RR (95% C.I.) ^a - By Hospital Admission Status						
	All Subjects	Zero	1-29 Days	P-val vZero	All 30 Days	P-val vZero	P-val v1-29d
Black Smoke (per 10 µgm-3 increase, note IQR = 10)							
30	0.08(-0.04,0.21)	0.03(-0.09,0.14)	0.09(-0.04,0.22)	0.483	0.20(-0.09,0.52)	0.271	0.483
1-6	-0.01(-0.31,0.30)	-0.08(-0.40,0.24)	0.29(-0.20,0.79)	0.215	-0.27(-1.00,0.50)	0.661	0.230
7-12	-0.04(-0.23,0.15)	-0.18(-0.40,0.04)	0.18(-0.17,0.54)	0.083	0.88(0.09,1.70)	0.010	0.114
13-18	0.14(-0.15,0.43)	0.39(-0.04,0.83)	-0.17(-0.60,0.27)	0.075	-0.14(-1.13,0.92)	0.359	0.951
19-24	0.21(-0.12,0.55)	0.08(-0.23,0.39)	0.23(-0.10,0.57)	0.514	0.26(-1.02,1.65)	0.790	0.964
25-30	0.18(0.00,0.37)	0.10(-0.12,0.33)	0.36(0.02,0.70)	0.223	0.34(-0.44,1.15)	0.576	0.964
PM 10 (per 10 µgm-3 increase, note IQR = 11)							
30	0.15(-0.05,0.36)	0.21(-0.03,0.47)	0.19(-0.20,0.63)	0.934	0.00(-0.58,0.72)	0.566	0.636
1-6	-0.01(-0.32,0.31)	0.08(-0.41,0.58)	-0.08(-0.54,0.39)	0.644	-0.08(-1.04,0.94)	0.775	0.999
7-12	0.29(-0.18,0.77)	0.29(-0.57,1.19)	0.14(-0.59,0.91)	0.804	0.94(-0.87,2.96)	0.540	0.440
13-18	0.05(-0.26,0.38)	-0.18(-0.66,0.32)	0.42(-0.38,1.26)	0.216	-0.21(-1.32,0.98)	0.956	0.382
19-24	0.07(-0.24,0.40)	0.38(-0.11,0.89)	0.06(-0.58,0.72)	0.445	-0.30(-1.26,0.72)	0.237	0.557
25-30	0.16(-0.17,0.50)	-0.20(-0.68,0.30)	0.49(0.03,0.96)	0.046	-0.02(-1.29,1.36)	0.803	0.482
PM 2.5 (per 10 µgm-3 increase, note IQR = 6)							
30	0.47(-0.27,1.38)	0.55(-0.18,1.45)	0.53(-0.06,1.23)	0.969	-1.26(-2.68,3.18)	0.290	0.291
1-6	1.05(0.14,2.01)	0.65(-0.83,2.28)	1.27(0.06,2.57)	0.549	1.51(-2.95,7.43)	0.749	0.927
7-12	0.08(-1.59,1.95)	-0.52(-1.89,0.97)	0.38(-1.32,2.26)	0.439	-2.56(-7.68,5.48)	0.565	0.423
13-18	0.32(-0.53,1.20)	-0.09(-1.49,1.44)	0.54(-0.58,1.74)	0.512	-0.60(-3.85,3.47)	0.800	0.569
19-24	0.71(-0.13,1.59)	1.53(-0.49,3.80)	0.83(-0.28,2.02)	0.571	-3.99(-6.65,-0.62)	0.007	0.010
25-30	-0.24(-1.55,1.19)	0.81(-2.18,4.42)	-0.73(-1.79,0.40)	0.365	0.56(-2.81,4.74)	0.919	0.505
Sulphur Dioxide (per 10 µgm-3 increase, note IQR = 17)							
30	0.00(-0.07,0.06)	0.02(-0.09,0.14)	-0.01(-0.21,0.20)	0.777	0.01(-0.22,0.25)	0.921	0.896
1-6	0.03(-0.19,0.26)	0.10(-0.10,0.30)	-0.04(-0.48,0.41)	0.580	-0.22(-0.79,0.37)	0.306	0.619
7-12	-0.02(-0.25,0.21)	-0.03(-0.23,0.17)	-0.04(-0.60,0.53)	0.980	0.17(-0.41,0.76)	0.516	0.610
13-18	0.09(-0.18,0.37)	0.22(0.02,0.42)	-0.13(-0.66,0.42)	0.238	0.32(-0.27,0.94)	0.740	0.268
19-24	0.06(-0.09,0.22)	0.10(-0.09,0.30)	0.03(-0.26,0.33)	0.687	-0.23(-0.80,0.35)	0.282	0.425
25-30	-0.02(-0.18,0.14)	-0.04(-0.23,0.16)	0.12(-0.17,0.42)	0.394	-0.89(-2.12,0.45)	0.208	0.144
Nitrogen Dioxide (per 10 µgm-3 increase, note IQR = 21)							
30	0.12(-0.05,0.30)	-0.07(-0.38,0.27)	0.20(-0.13,0.56)	0.277	0.21(-0.12,0.58)	0.255	0.959
1-6	-0.24(-0.64,0.17)	-0.42(-0.98,0.17)	0.01(-0.76,0.82)	0.390	-0.08(-0.76,0.62)	0.465	0.863
7-12	0.07(-0.14,0.27)	-0.21(-0.72,0.31)	0.04(-0.36,0.44)	0.460	1.14(0.22,2.10)	0.012	0.032
13-18	0.23(-0.11,0.58)	0.31(0.02,0.61)	0.35(-0.01,0.70)	0.895	0.03(-0.88,0.99)	0.571	0.538
19-24	0.23(-0.18,0.66)	0.22(-0.06,0.52)	0.09(-0.23,0.41)	0.527	0.27(-0.90,1.53)	0.943	0.772
25-30	0.15(-0.05,0.34)	0.05(-0.22,0.33)	0.30(0.00,0.60)	0.240	0.04(-0.61,0.72)	0.979	0.495

a. Percentage Relative Risk (95% Confidence Interval) per 10µgm-3 increase of pollutant on any single day within the lag period.

vZero. p-value associated with z-test comparison with zero days in hospital

v1-29. p-value associated with z-test comparison with 1-29 days in hospital

8.3.3 Pollution-IHD mortality: accounting for hospital status.

Inspection of the association between PM10 and PM2.5 concentration and ischaemic heart disease mortality in Figure 8.9 would appear to indicate that the relative shape across the lag for each hospital admission status was near identical regardless of the particulate size. In all three cases (zero, 1-29, and 30 days in hospital), a decrease risk was seen initially followed by a rise after 6 days towards increases in risk (1-29 days in hospital) or a fluctuation around to baseline change in risk, before a second stronger rise after 18-24 days. These relationships across the lag period are largely replicated in the lag stratified model (Table 8.4), with the exception of PM2.5 which reports a relatively strong increase in risk at 1-6 days, (%RR (95% C.I.) = 0.89% (-1.89%, 4.19%)) unlike the decrease in risk reported in the cubic distributed lag model. This again maybe related to the sample size. Black smoke reports small differences between the three hospital admission designations. The risk of IHD mortality for those in the community for all 30 days showed a small but immediate (lag 1) increase in risk per increase in BS (see Figure 8.9) that then fluctuates around the baseline for the remaining 30 days. The '1-29' days and 'all 30' days in hospital groups were not consistent for BS, with the first an increase in risk only occurring after 15 days in the '1-29' days in hospital group.

The shape and magnitude of risk reported in community only based subjects is remarkably similar between BS and SO2 both in the lag stratified (Table 8.4) and distributed lag plots (Figure D3 of Appendix D). The lag stratified model reported a significant increase in percentage relative risk of 0.10% (0.00%, 0.20%) in lag 13-18 days for SO2. The same subject group also showed an increase in IHD risk after 6 days associated with the traffic related pollutant NO2 that lasted the entire lag period. This was observed in the lag stratified model with the strongest increase in risk associated with a gaseous pollutant (NO2) of 0.23% at 19-24 day lag. Sulphur dioxide only reports a strong risk of IHD that last the entire lag period in those who spent all 30 days in hospital. Though this was contradicted in the lag stratified model where only increase risk was present during the 7-12 and 13-18 lag periods. Effect sizes are overall smaller for both SO2 and NO2.

Table 8.4 – Percentage relative risk (%RR) associated with an increase in 10µgm⁻³ within each pollutant associated with the lag stratified analysis for IHD, split by hospital admission status during exposure (Zero, 1-29, and all 30 days).

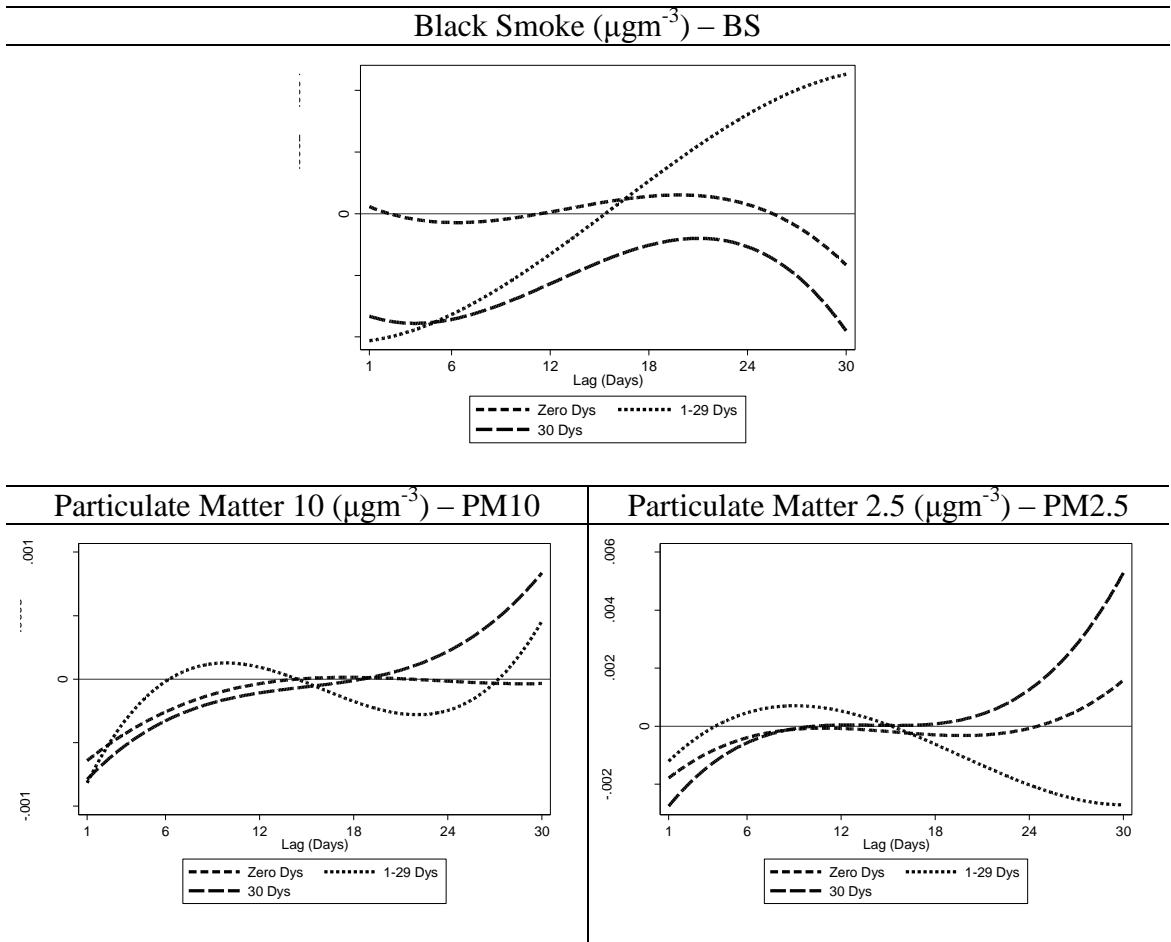
Lag (Dys)	IHD %RR (95%.C.I.)a - By Hospital Admission Status						
	All Subjects	Zero	1-29 Days	P-val vZero	All 30 Days	P-val vZero	P-val v1-29d
Black Smoke (per 10 µgm-3 increase, note IQR = 10)							
30	-0.02(-0.07,0.04)	0.00(-0.06,0.06)	-0.02(-0.11,0.06)	0.624	-0.19(-0.53,0.19)	0.314	0.388
1-6	-0.13(-0.34,0.09)	-0.05(-0.27,0.17)	-0.25(-0.58,0.08)	0.315	-0.27(-0.78,0.25)	0.436	0.953
7-12	-0.10(-0.25,0.04)	-0.09(-0.24,0.07)	-0.19(-0.42,0.05)	0.486	0.02(-0.58,0.64)	0.742	0.536
13-18	0.07(-0.03,0.18)	0.13(-0.01,0.27)	-0.09(-0.42,0.24)	0.233	-0.20(-0.88,0.52)	0.380	0.796
19-24	0.02(-0.08,0.12)	0.03(-0.09,0.15)	-0.01(-0.28,0.27)	0.799	0.03(-0.49,0.57)	0.996	0.894
25-30	-0.03(-0.15,0.09)	-0.03(-0.15,0.10)	0.10(-0.19,0.39)	0.440	-0.49(-1.09,0.14)	0.157	0.097
PM 10 (per 10 µgm-3 increase, note IQR = 11)							
30	-0.08(-0.17,0.02)	-0.11(-0.23,0.01)	-0.06(-0.21,0.10)	0.585	0.04(-0.46,0.63)	0.592	0.739
1-6	-0.42(-0.61,-0.23)	-0.34(-0.59,-0.08)	-0.52(-0.84,-0.19)	0.390	-0.58(-1.27,0.13)	0.515	0.866
7-12	-0.08(-0.40,0.23)	-0.23(-0.49,0.02)	0.15(-0.46,0.77)	0.258	-0.37(-1.24,0.55)	0.780	0.357
13-18	0.03(-0.16,0.23)	-0.01(-0.26,0.25)	0.04(-0.29,0.37)	0.835	0.26(-1.37,2.07)	0.759	0.800
19-24	-0.12(-0.32,0.07)	-0.10(-0.35,0.16)	-0.21(-0.54,0.12)	0.580	0.15(-0.59,0.93)	0.542	0.383
25-30	0.20(-0.07,0.47)	0.06(-0.20,0.32)	0.36(0.00,0.72)	0.183	0.38(-0.35,1.13)	0.418	0.958
PM 2.5 (per 10 µgm-3 increase, note IQR = 6)							
30	-0.15(-0.79,0.65)	-0.01(-0.46,0.51)	0.06(-1.09,1.82)	0.927	-1.22(-2.51,2.06)	0.349	0.365
1-6	-0.16(-0.75,0.46)	-0.38(-1.21,0.49)	0.02(-1.23,1.37)	0.613	0.89(-1.89,4.19)	0.413	0.597
7-12	0.38(-0.67,1.51)	0.20(-0.61,1.06)	0.91(-1.85,4.18)	0.650	-2.00(-6.16,3.83)	0.418	0.345
13-18	0.23(-0.35,0.82)	-0.18(-0.97,0.64)	1.46(-0.93,4.21)	0.213	-0.53(-3.28,2.78)	0.826	0.330
19-24	-0.27(-0.85,0.32)	0.00(-0.81,0.86)	-0.48(-1.34,0.43)	0.432	0.06(-2.64,3.29)	0.970	0.725
25-30	0.38(-0.36,1.15)	0.35(-0.47,1.22)	0.35(-0.55,1.30)	0.991	0.99(-2.61,5.52)	0.756	0.755
Sulphur Dioxide (per 10 µgm-3 increase, note IQR = 17)							
30	-0.03(-0.06,0.01)	-0.02(-0.06,0.03)	-0.07(-0.14,0.01)	0.241	0.02(-0.14,0.19)	0.683	0.346
1-6	-0.13(-0.28,0.03)	-0.08(-0.24,0.08)	-0.16(-0.35,0.03)	0.508	-0.12(-0.84,0.64)	0.922	0.908
7-12	-0.03(-0.11,0.06)	-0.03(-0.17,0.11)	-0.08(-0.30,0.15)	0.710	0.12(-0.59,0.85)	0.695	0.608
13-18	0.06(-0.03,0.14)	0.10(0.00,0.20)	-0.11(-0.32,0.11)	0.083	0.15(-0.25,0.57)	0.813	0.268
19-24	0.00(-0.12,0.11)	-0.01(-0.11,0.09)	0.04(-0.14,0.22)	0.669	-0.21(-0.73,0.33)	0.473	0.396
25-30	-0.07(-0.19,0.04)	-0.10(-0.20,0.01)	0.01(-0.18,0.19)	0.350	-0.10(-0.74,0.56)	0.983	0.752
Nitrogen Dioxide (per 10 µgm-3 increase, note IQR = 21)							
30	-0.04(-0.15,0.06)	0.01(-0.18,0.21)	-0.08(-0.27,0.11)	0.484	-0.50(-0.93,0.01)	0.062	0.125
1-6	-0.27(-0.53,-0.01)	-0.22(-0.46,0.03)	-0.22(-0.56,0.13)	0.999	-0.73(-1.78,0.39)	0.367	0.378
7-12	-0.10(-0.22,0.02)	-0.10(-0.25,0.05)	-0.04(-0.26,0.19)	0.663	-1.08(-2.14,0.07)	0.094	0.079
13-18	0.06(-0.06,0.19)	0.15(-0.05,0.35)	-0.09(-0.43,0.25)	0.226	-0.27(-0.97,0.46)	0.276	0.667
19-24	0.08(-0.14,0.31)	0.23(-0.17,0.64)	-0.05(-0.29,0.18)	0.233	-0.16(-0.63,0.33)	0.229	0.702
25-30	0.03(-0.09,0.15)	0.05(-0.09,0.19)	0.02(-0.20,0.23)	0.802	-0.34(-1.18,0.54)	0.387	0.436

a. Percentage Relative Risk (95% Confidence Interval) per 10µgm-3 increase of pollutant on any single day within the lag period.

vZero. p-value associated with z-test comparison with zero days in hospital

v1-29. p-value associated with z-test comparison with 1-29 days in hospital

Figure 8.9 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on IHD split by hospital admission during exposure (zero, 1-29, and all 30 days).



8.4 Influence of outliers and missing data

Two factors thought to have a strong influence on the results were the presence of influential outliers, and the presence of missing data. With non-negative pollution exposure data, outliers refer to the days with extreme concentration measurements. Due to systematic differences between monitors regarding measurement of the true exposure value and to be consistent with the definition of an extreme outlier across pollutants, an outlier was defined as being those measurements located within the greatest 1% of the data range for each monitor. These outliers were removed and the analysis repeated.

As indicated in the simulation study, missing pollution exposure data may be causing underestimation of the true effect in a complete cases (CC) analysis. The data were reanalysed with missing pollution exposure data imputed using the multiple imputation procedure described in Chapters 3.5 and 5. Note, this was performed in the complete exposure data and not those with outliers removed as removing the top 1% would be akin to removing data due to the observation themselves i.e. a form of missing not at random (MNAR). The multiple imputation procedure employed a predictive mean matching window with 10 nearest neighbour observations to randomly select the imputed value. A burn in period of 50 iterations was chosen to generate 10 imputed datasets for each pollutant. Ten imputations were considered appropriate after a trial run of the particulate pollutants indicated the relative efficiency was consistently around 99%, but with a minimum of 97.5% when the fraction of missing information was at approximately 25% in the black smoke monitors with a limited number of exposure days.

The main analysis, the complete cases with 100% of the available exposure data (CC – 100% Exposure) is also reported in Table 8.5, 8.6 and 8.7 to aid any comparisons. Table 8.5, 8.6 and 8.7 report for pneumonia, COPD and IHD the results of the two analyses; a complete cases analysis with outliers removed (CC – 99% Exposure) in this case the greatest 1% removed, and the multiple imputed complete data (MI – 100% Exposure). In all three cases the analysis reported here is based on those who were in the community for all 30 days i.e. ‘zero’ days in hospital during exposure, as they are thought to represent an exposure effect with reduced interference.

8.4.1 Outliers and missing data - Pneumonia

Table 8.5 reports results from the lag stratified model, and Figure 8.10 plots the cubic distributed lag relationship associated between pneumonia and both the particulates and gaseous pollutants. The cubic function plots represent the change in risk during the 30 days following a single unit increase in the pollutant. In other words a value above the x axis at lag day 6 represents an increase per unit increase in pollutant in mortality risk six days after exposure.

When compared to the main analysis, removal of an extreme outlier does not appear to have had a strong influence on effect estimates with respect to the particulates (PM10 and PM2.5) and the two gaseous pollutants. In general, the lag shape and effect estimates have changed very little in these pollutants with only a slight difference in PM10 (Figure 8.10) that is not reflected in the %RR in Table 8.5. Black smoke does indicate a potential difference. The main analysis reports a peak risk in the first 12 days which, when outliers are removed, becomes a period of decrease risk before peak increase risk after 18 days. The lag stratified effect estimates appear to indicate %RR effect estimates in similar directions just with a greater magnitude. For example, lags 1-6 and 13-18 the %RR alters from -0.06% to -0.18%, and 0.10% to 0.18% when the outliers were removed.

Multiple imputation (MI) of missing pollution measurements appeared to have little influence on lag shape as indicated by the cubic function plots. Only PM10 indicated a slight change from the main analysis, where the previously almost linear relationship was returned to a cubic lag shape, observed when all subjects were combined (Figure 8.5) and in the effect estimates of the lag stratified model (Table 8.5). Of the remaining pollutants, %RR effect estimates appeared largely similar to the main analysis with some cases of slightly increased magnitude. Though with range of the confidence intervals the previously negative short-term lag period effect estimates (i.e. a decrease in risk as pollution increased) for 'black smoke' became equivalent to baseline when MI was applied (lags 1-6 and 7-12 respectively, were -0.06% and 0.09% before and 0.04% and 0.10% after MI). A reduction, though sometimes small, was observed in the width of the confidence intervals in all lagged effects indicated a small improved precision.

Table 8.5 - Percentage relative risk (%RR) associated with an increase in 10µgm-3 within each pollutant associated with the lag stratified analysis for Pneumonia, results of the analysis investigating outliers and missing data.

Lag Period/ Pollutant	Pneumonia %RR (95%.C.I.) ^a - Zero Days in Hospital		
	CC - 100% Exposure	CC - 99% Exposure	MI 100% Exposure
Black Smoke (per 10 µgm-3 increase, note IQR = 10)			
30 Days	0.02(-0.04,0.08)	-0.02(-0.11,0.08)	0.02(-0.03,0.08)
1-6 Days	-0.06(-0.25,0.14)	-0.18(-0.48,0.13)	0.04(-0.12,0.20)
7-12 Days	0.09(-0.06,0.25)	-0.18(-0.41,0.07)	0.10(-0.04,0.23)
13-18 Days	0.10(-0.13,0.32)	0.18(-0.16,0.53)	0.07(-0.07,0.21)
19-24 Days	-0.15(-0.30,0.00)	-0.05(-0.28,0.19)	-0.09(-0.22,0.04)
25-30 Days	-0.03(-0.18,0.12)	-0.22(-0.45,0.02)	0.04(-0.10,0.17)
PM 10 (per 10 µgm-3 increase, note IQR = 11)			
30 Days	0.04(-0.12,0.21)	0.05(-0.15,0.26)	0.02(-0.08,0.13)
1-6 Days	0.19(-0.39,0.79)	0.21(-0.40,0.84)	0.11(-0.20,0.41)
7-12 Days	-0.12(-0.46,0.22)	-0.17(-0.56,0.23)	-0.15(-0.46,0.16)
13-18 Days	-0.06(-0.39,0.29)	-0.12(-0.51,0.28)	-0.09(-0.49,0.33)
19-24 Days	0.09(-0.25,0.43)	0.09(-0.31,0.49)	0.26(-0.09,0.62)
25-30 Days	-0.05(-0.39,0.29)	-0.04(-0.44,0.37)	-0.05(-0.54,0.46)
PM 2.5 (per 10 µgm-3 increase, note IQR = 6)			
30 Days	0.92(0.25,1.70)	1.04(-0.05,2.49)	0.75(0.08,1.55)
1-6 Days	0.35(-0.86,1.66)	0.58(-1.05,2.38)	0.48(-0.78,1.83)
7-12 Days	0.87(-0.33,2.16)	0.98(-0.61,2.73)	0.99(-0.25,2.31)
13-18 Days	-0.06(-1.64,1.67)	-0.16(-1.94,1.84)	0.11(-1.08,1.39)
19-24 Days	1.51(0.33,2.76)	1.46(-0.18,3.28)	1.60(0.37,2.92)
25-30 Days	-0.99(-2.91,1.20)	-0.72(-3.24,2.28)	-0.99(-2.85,1.12)
Sulphur Dioxide (per 10 µgm-3 increase, note IQR = 17)			
30 Days	0.05(-0.01,0.10)	0.05(-0.02,0.12)	0.05(-0.01,0.10)
1-6 Days	0.05(-0.09,0.18)	0.07(-0.10,0.25)	0.06(-0.07,0.19)
7-12 Days	-0.04(-0.17,0.10)	-0.10(-0.26,0.07)	-0.01(-0.13,0.12)
13-18 Days	0.09(-0.04,0.22)	0.08(-0.08,0.25)	0.09(-0.03,0.22)
19-24 Days	0.04(-0.11,0.19)	0.10(-0.07,0.27)	0.01(-0.11,0.14)
25-30 Days	0.06(-0.07,0.19)	0.03(-0.14,0.20)	0.07(-0.06,0.19)
Nitrogen Dioxide (per 10 µgm-3 increase, note IQR = 21)			
30 Days	-0.02(-0.18,0.14)	-0.02(-0.13,0.10)	-0.03(-0.19,0.13)
1-6 Days	-0.23(-0.56,0.11)	-0.28(-0.61,0.06)	-0.17(-0.50,0.17)
7-12 Days	-0.05(-0.25,0.14)	-0.13(-0.35,0.08)	-0.16(-0.33,0.01)
13-18 Days	-0.06(-0.51,0.42)	-0.04(-0.51,0.44)	-0.04(-0.44,0.36)
19-24 Days	-0.04(-0.23,0.16)	0.07(-0.14,0.29)	0.04(-0.24,0.33)
25-30 Days	0.04(-0.15,0.23)	0.07(-0.15,0.28)	0.10(-0.06,0.26)

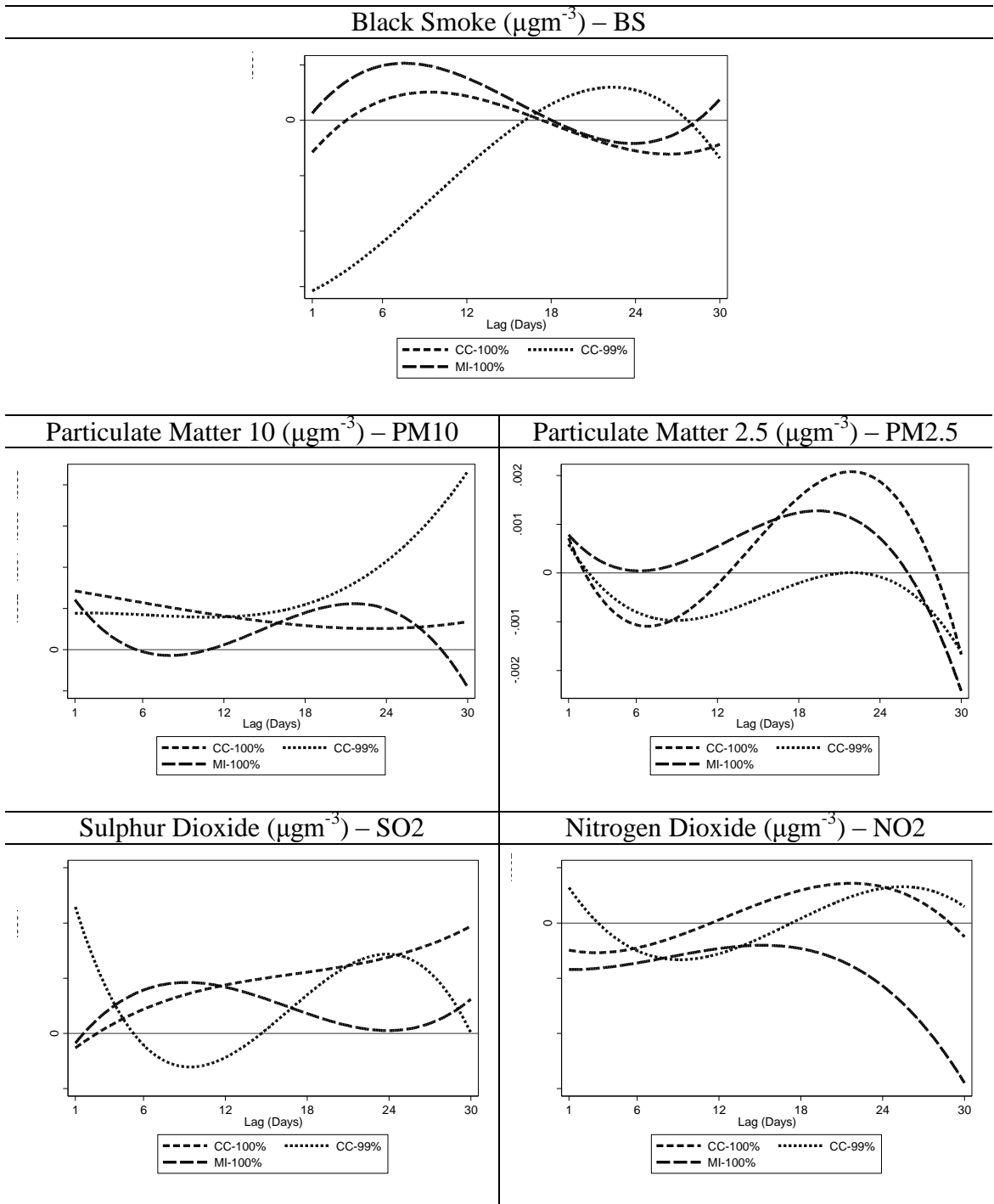
a. Percentage Relative Risk (95% Confidence Interval) per 10µgm-3 increase of pollutant on any single day within the lag period.

CC 100% Exposure Data. Complete Cases with 100% exposure data (repeat of main analysis).

CC 99% Exposure Data. Complete Cases with greatest 1% exposure data removed.

MI 100% Exposure Data. Multiple Imputed missing exposure data using complete observed exposure data

Figure 8.10 – Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on pneumonia mortality, repeated for analyses investigating outliers and missing data.



8.4.2 Outliers and missing data - COPD

Similar conclusions could be made for the effect associated with COPD (Table 8.6 & Figure 8.11) as were made for pneumonia when outliers were removed. In all but PM10 and possibly NO2 the cubic plots indicated a similar but slightly exaggerated change in shape across the lag period before and after the outliers were removed. In NO2 the shape remains consistent however when the outlier is removed the effect estimate for the entire 30 days becomes negative i.e. a decrease in risk. Note, positive %RRs were observed for NO2 in the lag stratified model in the longer lags with even a significant increase at lag 19-24 of 0.34% (95% C.I. = 0.02%, 0.66%). In PM10, NO2 and the remaining pollutants the effect estimates largely held constant (Table 8.6) before and after the outlier was removed. Where the biggest influence appears to have been in PM2.5 where the longer delayed effects appeared to become greater and significant (lag 19-24 increased to 2.94% (0.56%,5.61%).

Similarly to when the outlier was removed, multiple imputation only appears to influence the cubic plots of black smoke by smoothing out the previously cubic curve into an almost linear relationship, and NO2 by removing the previously positive risk. The effect estimates of the lag stratified model are again closer to the main complete cases analysis. In the black smoke, the %RR were consistently positive throughout the 30 day lag period, with a peak period of COPD mortality risk occurring at lag 13-18 days (%RR = 0.28%). Though much smaller and non-significant in this case, NO2 once again displays a period of increase risk between lag 19-24 days. The PM2.5 significant risk that occurred when outliers were removed has returned to be non-significant and similar to the main complete cases analysis. Even though the width of the confidence intervals reduced, in only PM10 did a previously non-significant percentage relative risk become significant. This occurred in the PM10 %RR (95% C.I) associated with a $10\mu\text{gm}^{-3}$ increase on an individual day within the 30 day lag (lag 1-30) and the lag 19-24 day was 0.15% (0.00%,0.30%) and 0.48% (0.02%, 0.95%) respectively.

Table 8.6 – Percentage relative risk (%RR) associated with an increase in 10µg⁻³ within each pollutant associated with the lag stratified analysis for COPD, results of the analysis investigating outliers and missing data.

Lag Period/ Pollutant	COPD %RR (95%.C.I.) ^a - Zero Days in Hospital		
	CC - 100% Exposure	CC - 99% Exposure	MI 100% Exposure
Black Smoke (per 10 µg⁻³ increase, note IQR = 10)			
30 Days	0.03(-0.09,0.14)	0.09(-0.15,0.35)	0.03(-0.05,0.11)
1-6 Days	-0.08(-0.40,0.24)	-0.20(-0.69,0.30)	0.00(-0.19,0.20)
7-12 Days	-0.18(-0.40,0.04)	-0.15(-0.51,0.22)	0.01(-0.17,0.19)
13-18 Days	0.39(-0.04,0.83)	0.55(-0.07,1.20)	0.28(-0.05,0.61)
19-24 Days	0.08(-0.23,0.39)	0.33(-0.04,0.70)	0.04(-0.16,0.24)
25-30 Days	0.10(-0.12,0.33)	0.17(-0.20,0.54)	0.11(-0.09,0.30)
PM 10 (per 10 µg⁻³ increase, note IQR = 11)			
30 Days	0.21(-0.03,0.47)	0.20(-0.08,0.51)	0.15(0.00,0.30)
1-6 Days	0.08(-0.41,0.58)	0.06(-0.50,0.63)	-0.12(-0.55,0.32)
7-12 Days	0.29(-0.57,1.19)	0.28(-0.74,1.37)	0.36(-0.09,0.82)
13-18 Days	-0.18(-0.66,0.32)	-0.20(-0.77,0.38)	-0.21(-0.66,0.24)
19-24 Days	0.38(-0.11,0.89)	0.38(-0.20,0.98)	0.48(0.02,0.95)
25-30 Days	-0.20(-0.68,0.30)	-0.20(-0.77,0.38)	0.21(-0.22,0.65)
PM 2.5 (per 10 µg⁻³ increase, note IQR = 6)			
30 Days	0.55(-0.18,1.45)	0.57(-0.37,1.80)	0.48(-0.26,1.40)
1-6 Days	0.65(-0.83,2.28)	0.52(-1.43,2.71)	0.61(-0.93,2.30)
7-12 Days	-0.52(-1.89,0.97)	-0.10(-1.89,1.91)	-0.52(-1.92,1.01)
13-18 Days	-0.09(-1.49,1.44)	-0.43(-2.22,1.57)	-0.02(-1.46,1.55)
19-24 Days	1.53(-0.49,3.80)	2.92(0.56,5.61)	1.29(-0.18,2.89)
25-30 Days	0.81(-2.18,4.42)	0.39(-2.98,4.60)	0.73(-2.41,4.57)
Sulphur Dioxide (per 10 µg⁻³ increase, note IQR = 17)			
30 Days	0.02(-0.09,0.14)	0.03(-0.09,0.15)	0.03(-0.06,0.11)
1-6 Days	0.10(-0.10,0.30)	0.07(-0.19,0.33)	0.11(-0.08,0.31)
7-12 Days	-0.03(-0.23,0.17)	-0.04(-0.29,0.22)	-0.04(-0.23,0.16)
13-18 Days	0.22(0.02,0.42)	0.17(-0.09,0.43)	0.15(-0.04,0.34)
19-24 Days	0.10(-0.09,0.30)	0.32(0.06,0.59)	0.05(-0.14,0.24)
25-30 Days	-0.04(-0.23,0.16)	0.05(-0.21,0.31)	0.00(-0.19,0.20)
Nitrogen Dioxide (per 10 µg⁻³ increase, note IQR = 21)			
30 Days	-0.07(-0.38,0.27)	-0.05(-0.36,0.29)	-0.06(-0.29,0.19)
1-6 Days	-0.42(-0.98,0.17)	-0.42(-1.00,0.19)	-0.39(-0.98,0.22)
7-12 Days	-0.21(-0.72,0.31)	-0.21(-0.72,0.31)	-0.21(-0.62,0.22)
13-18 Days	0.31(0.02,0.61)	0.27(-0.05,0.59)	0.00(-0.25,0.24)
19-24 Days	0.22(-0.06,0.52)	0.34(0.02,0.66)	0.09(-0.42,0.62)
25-30 Days	0.05(-0.22,0.33)	0.02(-0.29,0.33)	0.09(-0.14,0.32)

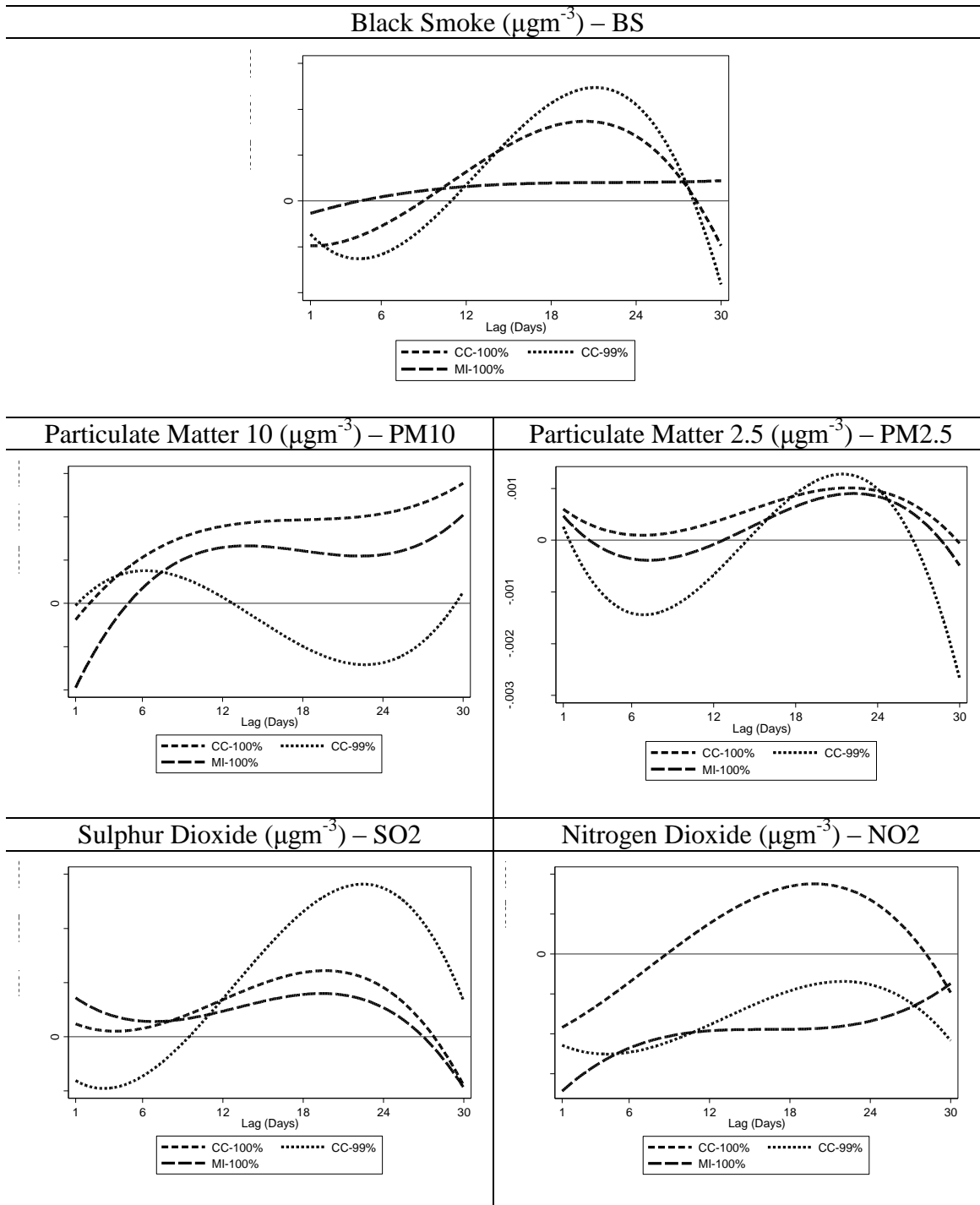
a. Percentage Relative Risk (95% Confidence Interval) per 10µg⁻³ increase of pollutant on any single day within the lag period.

CC 100% Exposure Data. Complete Cases with 100% exposure data (repeat of main analysis).

CC 99% Exposure Data. Complete Cases with greatest 1% exposure data removed.

MI 100% Exposure Data. Multiple Imputed missing exposure data using complete observed exposure data

Figure 8.11 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on COPD mortality, repeated for analyses investigating outliers and missing data.



8.4.3 Outliers and missing data – Ischaemic Heart Disease

Though the change in shape across the lag period was similar within the two respiratory diseases, some departures in shape and effect magnitude could be seen across the three analyses; complete cases, outlier removed, and multiple imputation. For ischaemic heart disease, other than at the extreme outer range of the 30 day lag (e.g. black smoke in <5 or >25 day lag periods) very little difference was seen between the three analyses for the five pollutants (Figure 8.12). Only PM10 showed a slight departure in the effect magnitude when the outlier was removed, when between 6 and 30 days the cubic function indicated a slightly greater increase in risk per unit increase of PM10.

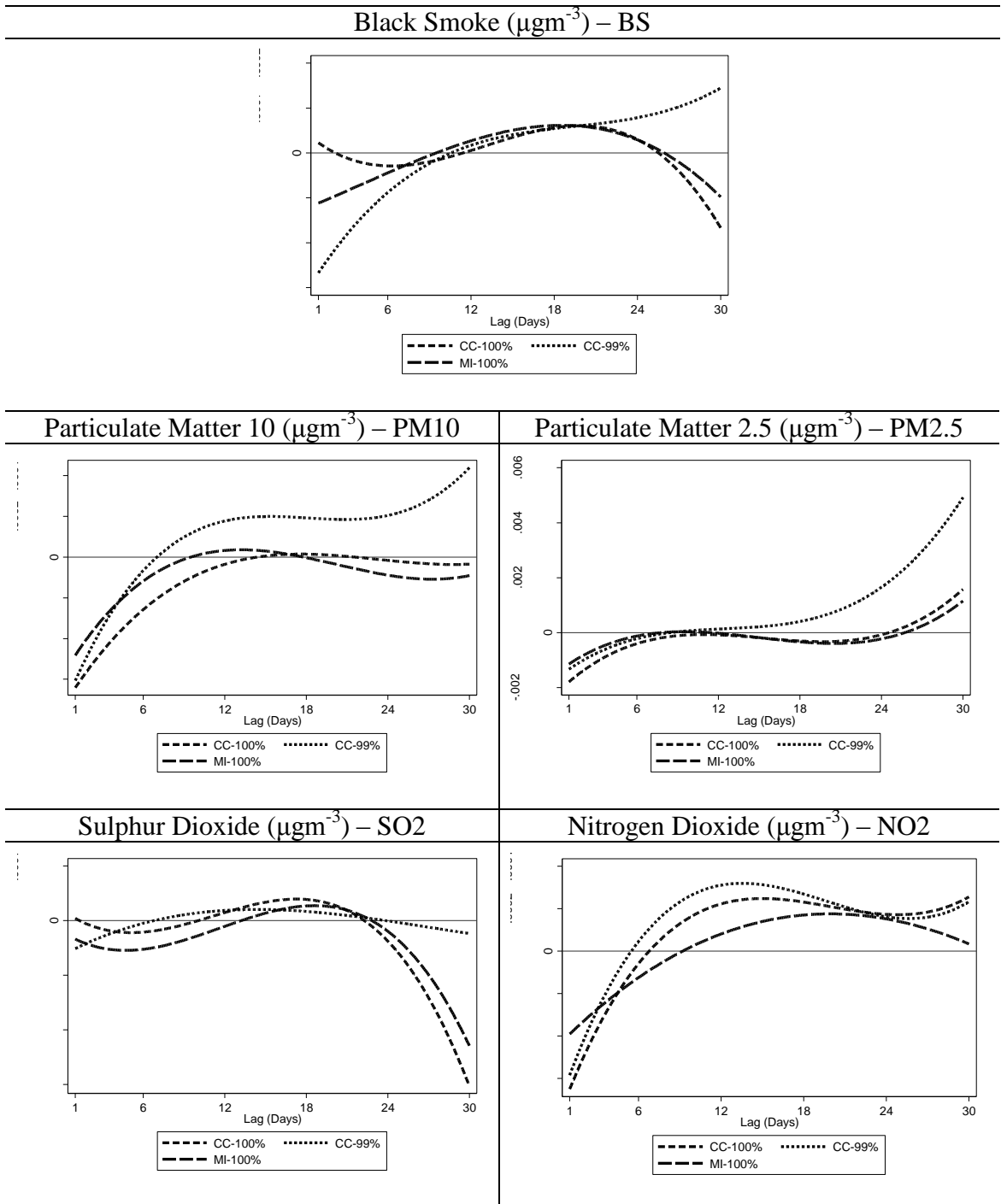
The limited variation across the three methods in the cubic distributed lag plots was reflected in the percentage relative risk (%RR) of the lag stratified model (Table 8.7); the direction of the effect estimate appeared to remain consistent regardless of which analysis methods was applied. The magnitude of the effect estimates also appeared consistent across pollutants, with only PM2.5 showing relatively large difference. At lag periods 1-6, 13-18, and 19-24 respectively, the percentage relative risk magnitude increased from -0.38% to -1.01%, -0.18% to -0.38%, and 0.00% to 0.25% when the outlier was removed compared to the main complete cases analysis. The multiple imputation results, even in PM2.5, remained consistent across with the main analysis. Even though the width of the confidence intervals again tightened around the estimate of the change in percentage relative risk it was not enough to alter any conclusions regarding any effect significance.

Table 8.7 – Percentage relative risk (%RR) associated with an increase in 10µgm⁻³ within each pollutant associated with the lag stratified analysis for IHD, results of the analysis investigating outliers and missing data.

Lag Period/ Pollutant	IHD %RR (95%.C.I.) ^a - Zero Days in Hospital		
	CC - 100% Exposure	CC - 99% Exposure	MI 100% Exposure
Black Smoke (per 10 µgm⁻³ increase, note IQR = 10)			
30 Days	0.00(-0.06,0.06)	-0.04(-0.14,0.06)	-0.01(-0.05,0.03)
1-6 Days	-0.05(-0.27,0.17)	-0.14(-0.46,0.18)	-0.04(-0.18,0.10)
7-12 Days	-0.09(-0.24,0.07)	-0.06(-0.24,0.12)	-0.03(-0.14,0.08)
13-18 Days	0.13(-0.01,0.27)	0.25(-0.02,0.52)	0.10(-0.01,0.22)
19-24 Days	0.03(-0.09,0.15)	0.06(-0.15,0.27)	-0.01(-0.11,0.09)
25-30 Days	-0.03(-0.15,0.10)	0.01(-0.25,0.27)	-0.02(-0.19,0.16)
PM 10 (per 10 µgm⁻³ increase, note IQR = 11)			
30 Days	-0.11(-0.23,0.01)	-0.11(-0.25,0.03)	-0.08(-0.16,0.00)
1-6 Days	-0.34(-0.59,-0.08)	-0.38(-0.66,-0.09)	-0.28(-0.51,-0.04)
7-12 Days	-0.23(-0.49,0.02)	-0.24(-0.53,0.06)	0.01(-0.23,0.25)
13-18 Days	-0.01(-0.26,0.25)	0.04(-0.26,0.34)	-0.04(-0.27,0.20)
19-24 Days	-0.10(-0.35,0.16)	-0.13(-0.43,0.17)	-0.02(-0.26,0.23)
25-30 Days	0.06(-0.20,0.32)	0.08(-0.22,0.38)	-0.07(-0.29,0.16)
PM 2.5 (per 10 µgm⁻³ increase, note IQR = 6)			
30 Days	-0.01(-0.46,0.51)	-0.10(-0.58,0.45)	-0.16(-0.71,0.51)
1-6 Days	-0.38(-1.21,0.49)	-1.01(-2.03,0.08)	-0.44(-1.29,0.46)
7-12 Days	0.20(-0.61,1.06)	0.13(-0.90,1.24)	0.09(-0.74,0.97)
13-18 Days	-0.18(-0.97,0.64)	-0.38(-1.42,0.74)	-0.16(-0.97,0.69)
19-24 Days	0.00(-0.81,0.86)	0.25(-0.83,1.40)	-0.03(-0.86,0.85)
25-30 Days	0.35(-0.47,1.22)	0.08(-0.96,1.20)	0.35(-0.49,1.24)
Sulphur Dioxide (per 10 µgm⁻³ increase, note IQR = 17)			
30 Days	-0.02(-0.06,0.03)	0.00(-0.06,0.05)	-0.02(-0.06,0.02)
1-6 Days	-0.08(-0.24,0.08)	-0.10(-0.30,0.10)	-0.05(-0.17,0.08)
7-12 Days	-0.03(-0.17,0.11)	0.05(-0.10,0.19)	-0.05(-0.15,0.05)
13-18 Days	0.10(0.00,0.20)	0.07(-0.05,0.20)	0.11(0.02,0.21)
19-24 Days	-0.01(-0.11,0.09)	0.01(-0.11,0.14)	-0.03(-0.13,0.07)
25-30 Days	-0.10(-0.20,0.01)	0.00(-0.14,0.15)	-0.09(-0.18,0.01)
Nitrogen Dioxide (per 10 µgm⁻³ increase, note IQR = 21)			
30 Days	0.01(-0.18,0.21)	0.00(-0.19,0.21)	0.02(-0.15,0.20)
1-6 Days	-0.22(-0.46,0.03)	-0.20(-0.36,-0.05)	-0.16(-0.36,0.05)
7-12 Days	-0.10(-0.25,0.05)	-0.11(-0.27,0.05)	-0.03(-0.16,0.10)
13-18 Days	0.15(-0.05,0.35)	0.13(-0.04,0.29)	0.15(-0.07,0.37)
19-24 Days	0.23(-0.17,0.64)	0.26(-0.13,0.65)	0.23(-0.18,0.66)
25-30 Days	0.05(-0.09,0.19)	0.00(-0.16,0.15)	0.00(-0.12,0.12)

^a. Percentage Relative Risk (95% Confidence Interval) per 10µgm⁻³ increase of pollutant on any single day within the lag period.
CC 100% Exposure Data. Complete Cases with 100% exposure data (repeat of main analysis).
CC 99% Exposure Data. Complete Cases with greatest 1% exposure data removed.
MI 100% Exposure Data. Multiple Imputed missing exposure data using complete observed exposure data

Figure 8.12 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in particulate pollutants on IHD mortality, repeated for analyses investigating outliers and missing data.



As stated in the introduction, the analyses reported here (Chapters 8.4.1 to 8.4.3) were repeated for all study subjects i.e. any subject regardless of the location, community or hospital, during the exposure period. The equivalent Table 8.5, 8.6, and 8.7 and Figure 8.10, 8.11, and 8.12 can be found in Appendix D under Table D2a, to D2c and Figures D4 – D6 for the three causes of death pneumonia, COPD, and ischaemic heart diseases respectively.

Compared to the results presented here for community based subjects only, the equivalent results for all subjects in the study appear to show a far greater amount of variation between the three analysis methods. Though some patterns in risk across the lag period have been replicated in the cubic functions there does not appear to be a clear pattern in differences between the standard complete cases, the outliers removed, and the multiple imputation analysis. As with the community based subjects the lag stratified model appeared to produce a more consistent set of effect estimates across the three analyses. Even so, the variation in the direction and magnitude of the effect estimates between the three analyses was far greater, even in ischaemic heart disease which of the three causes of death, appeared to be the most consistent in the community based subjects.

8.5 Sensitivity analysis – Analysis and data checks

The results of the main analysis created some questions regarding their reliability. It was therefore proposed that an investigation of the results be performed. This came in two parts a comparison with similar work, and an investigation of the between monitor variability.

Part 1: Three papers published between 2005-2010 Carder et al. analysed the data for Scotland between January 1981 and December 2001 focusing on the effect of temperature on mortality.^{13,272,273} The 2008 paper investigated the interaction effect between BS particulate pollution and temperature. To help confirm the analysis performed in this study was correct an attempt was made to replicate the results using this data. However, a number of differences are present between the two studies some of which cannot be replicated.

Table 8.8 reports the results associated with the analysis after undergoing stage by stage changes that reflected as many differences between the analysis performed in this study and that performed by Carder et al (2008). Starting with the complete dataset used in this thesis with the ‘primary’ cause of death field rather than ‘any’ cause of death field. It replicated the respiratory category by combining pneumonia and COPD, and the cardiovascular category was represented by the IHD subjects. One by one the following additional changes were made.

- Same day exposure (lag 0) was included as a parameter.
- Three bespoke linear temperature covariates (Note pneumonia and COPD both had 1°C & 15°C knot points) were replaced with two linear split at 11°C.
- The random effects adjustment for between monitor clustering was removed.
- The date of the analysis was restricted to Jan 1981 and Dec 2001
- The analysis was repeated for one monitor Edinburgh 14.

At each step the results are reported along with those reported by Carder et al in Table 8.8. As all of factors that differ between the two analyses could not be fully taken into account, the effect estimates still contained some systematic differences between the two studies. Even so, with the exception of lag 0 %RR on respiratory disease (-0.34% vs 0.50%), the results did appear to be similar once all factors were taken into account, providing some confidence in the results produced in this thesis.

Table 8.8 – Results (%RR) associated with of a sensitivity analysis attempting to replicate previous work (Carder et al (2008)) done on the similar dataset.

Additional Matching Characteristics	%RR (95%.C.I.)a per 10 $\mu\text{g}\text{m}^{-3}$ increase in Black Smoke					
	0 Days	1-6 Days	7-12 Days	13-18 Days	19-24 Days	25-30 Days
Pneumonia & COPD Combined \approx Respiratory Disease in Carder et al (2008) paper						
Study Result - PCOD	-	-0.04(-0.25,0.16)	-0.05(-0.29,0.18)	-0.04(-0.20,0.12)	-0.04(-0.22,0.14)	0.00(-0.29,0.29)
Lag 0	-0.35(-0.91,0.21)	-0.03(-0.34,0.28)	-0.09(-0.33,0.15)	-0.05(-0.21,0.12)	0.01(-0.20,0.23)	-0.02(-0.35,0.32)
Temperature (<11oC/> 11oC)	-0.36(-0.91,0.19)	-0.02(-0.33,0.30)	-0.02(-0.29,0.25)	-0.01(-0.17,0.16)	0.04(-0.19,0.28)	0.07(-0.27,0.43)
RE removed	-0.38(-0.93,0.18)	0.03(-0.12,0.19)	0.01(-0.14,0.17)	-0.03(-0.18,0.12)	-0.01(-0.16,0.13)	0.12(-0.03,0.27)
Jan81-Dec2001	-0.32(-0.90,0.26)	0.04(-0.12,0.21)	0.00(-0.16,0.16)	0.01(-0.15,0.18)	-0.04(-0.20,0.11)	0.17(0.01,0.33)
Ed Mon 14 Only	-0.34(-0.93,0.24)	0.09(-0.08,0.26)	0.02(-0.14,0.19)	0.05(-0.12,0.22)	-0.05(-0.21,0.11)	0.17(0.01,0.33)
Carder et al. 2008 (Respiratory)	0.50(-0.05,1.00)	0.11(-0.02,0.28)	0.22(0.08,0.32)	0.15(0.03,0.31)	0.08(-0.13,0.22)	0.22(0.05,0.35)
Ischaemic Heart Disease \approx Cardiovascular Disease in Carder et al (2008) paper						
Study Result - PCOD	-	-0.09(-0.31,0.14)	-0.07(-0.19,0.05)	0.08(-0.08,0.24)	0.06(-0.05,0.17)	-0.03(-0.16,0.11)
Lag 0	-0.08(-0.87,0.73)	-0.08(-0.31,0.16)	-0.09(-0.20,0.03)	0.10(-0.09,0.28)	0.04(-0.07,0.16)	-0.02(-0.17,0.13)
Temp Adjust (<11oC/> 11oC)	0.01(-0.78,0.80)	-0.05(-0.31,0.22)	-0.08(-0.19,0.03)	0.12(-0.07,0.30)	0.06(-0.05,0.17)	0.00(-0.16,0.15)
RE removed	0.27(-0.11,0.66)	0.01(-0.10,0.12)	-0.08(-0.19,0.02)	0.11(0.01,0.22)	0.06(-0.04,0.16)	-0.01(-0.12,0.09)
Jan81-Dec2001	0.29(-0.11,0.70)	0.03(-0.09,0.14)	-0.06(-0.18,0.05)	0.13(0.02,0.25)	0.05(-0.06,0.17)	0.03(-0.09,0.14)
Ed Mon 14 Only	0.37(-0.04,0.78)	0.03(-0.08,0.15)	-0.06(-0.17,0.05)	0.14(0.02,0.26)	0.06(-0.05,0.17)	0.03(-0.08,0.15)
Carder et al. 2008 (Cardiovascular)	0.18(-0.12,0.48)	-0.05(-0.13,0.03)	0.01(-0.10,0.10)	0.08(-0.02,0.12)	-0.05(-0.12,0.02)	0.08(-0.01,0.10)

a. Percentage Relative Risk (95% Confidence Interval) per 10 $\mu\text{g}\text{m}^{-3}$ increase of pollutant on any single day within the lag period.
Lag 0. Inclusion of same day (Day 0) parameter in the model.
Temp. Temperature adjustment altered to two continuous linear threshold variables above and below 11oC.
RE Adjustment for multiple monitors using Random Effects model (removed).

Part 2: To determine the potential influence of between monitors variation the results associated with each individual ‘black smoke’ monitor included in this study were extracted before a composite main effect was calculated. Table 8.9 reports monitor specific lag stratified effect estimates for the three cause of deaths based on ‘any’ cause of death field. The %RRs (95% C.I.) are reported for both the lag 1-30 day average, and the 30 days split into 5 lag periods of 6 days each.

The magnitude of the effects largely depends on the sample size. Those monitors are only active for a short period (Ed24, Ed St Leonards, and Glasgow Central) or with large amounts of missing data (Ab3 and Ed25) appear to be unstable and corresponding confidence intervals are large. Even when cause specific mortality rate was relatively high; the size of the %RR (95% C.I.) in these monitors was often still large e.g. IHD lag 1-6 = -4.42% (-7.01%, 1.13%). The magnitude of the results from monitors with large complete exposure data (Ab 2, Ed14 & Glasgow 20, 51, 95) are closer to the expected effect size for a single day average increase. Often the direction of the effect do disagree, for example monitors Ab2 and Ed 14 relating to pneumonia mortality give lag 1-6 days %RR of -0.31% and 0.22%. Even those monitors running concurrently for the same time period in the same city (Glasgow 20, 51, and 95) with a large sample size disagree in terms of the direction of the effect.

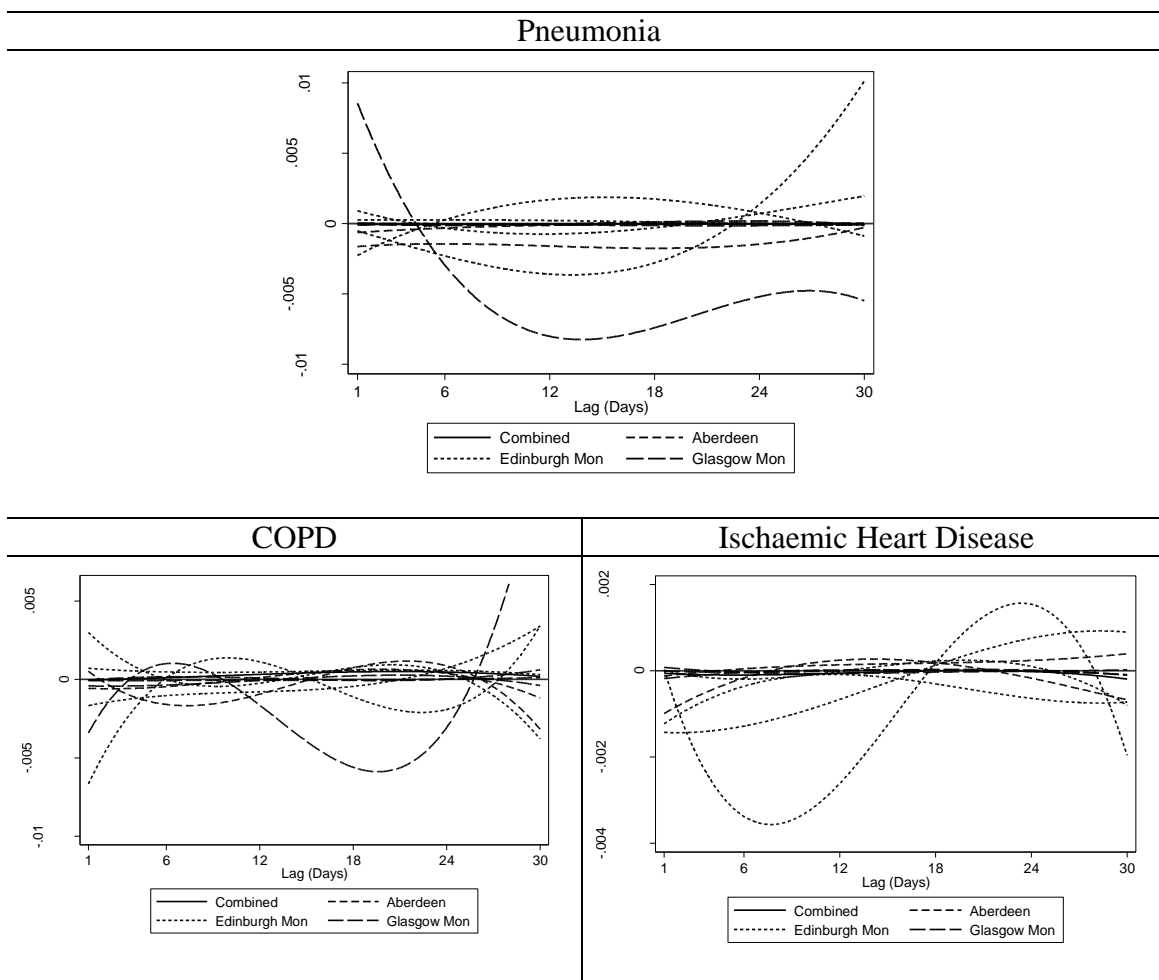
Table 8.9 –Percentage relative risk (%RR) associated with increase in black smoke within each individual monitor used in the main analysis.

Mon / Lag	%RR (95%.C.I.)a per 10 µgm ⁻³ increase in Black Smoke Pollution				
	1-6 Days	7-12 Days	13-18 Days	19-24 Days	25-30 Days
Pneumonia					
Ab2	-0.31(-0.71,0.10)	-0.10(-0.60,0.42)	0.24(-0.30,0.79)	-0.15(-0.68,0.40)	0.09(-0.45,0.64)
Ab3	-1.27(-2.77,0.40)	-0.80(-2.26,0.81)	-0.47(-1.93,1.13)	-1.10(-2.51,0.45)	-0.37(-1.86,1.27)
Ed14	0.22(-0.15,0.61)	-0.15(-0.51,0.23)	0.17(-0.20,0.55)	-0.23(-0.59,0.14)	-0.14(-0.50,0.24)
Ed24	0.09(-4.39,6.20)	-2.41(-5.55,1.63)	-2.13(-5.36,2.03)	2.02(-2.22,7.48)	4.53(0.07,10.18)
Ed25	-0.04(-1.28,1.30)	-0.95(-2.16,0.35)	-1.39(-2.68,0.01)	0.96(-0.39,2.42)	-0.12(-1.39,1.24)
EdStL	0.30(-2.29,3.35)	2.21(-0.60,5.50)	0.19(-2.55,3.46)	0.34(-2.29,3.46)	-0.88(-3.23,1.88)
Gl20	-0.11(-0.31,0.09)	0.06(-0.15,0.27)	0.18(-0.02,0.39)	-0.19(-0.39,0.02)	0.13(-0.07,0.34)
Gl51	-0.12(-0.35,0.10)	0.10(-0.12,0.33)	-0.21(-0.43,0.02)	-0.08(-0.30,0.13)	-0.08(-0.30,0.15)
Gl95	0.03(-0.34,0.41)	-0.17(-0.53,0.21)	0.04(-0.30,0.40)	0.05(-0.28,0.39)	0.11(-0.22,0.44)
GlCen	0.69(-3.31,5.87)	1.16(-2.25,5.39)	-0.04(-2.81,3.29)	1.96(-0.60,4.93)	-0.02(-2.77,3.28)
COPD					
Ab2	-0.12(-0.81,0.60)	0.07(-0.84,1.04)	0.68(-0.30,1.72)	0.64(-0.37,1.71)	-0.54(-1.48,0.47)
Ab3	-0.52(-2.95,2.35)	-1.80(-3.97,0.73)	1.28(-1.22,4.19)	1.59(-0.91,4.47)	-0.28(-2.60,2.42)
Ed14	0.68(0.10,1.29)	-0.10(-0.67,0.48)	0.55(-0.04,1.15)	0.36(-0.22,0.96)	0.29(-0.29,0.89)
Ed24	-5.85(-9.84,0.47)	-0.08(-5.44,7.83)	-3.53(-7.70,2.57)	-1.32(-6.32,6.10)	0.31(-4.78,7.58)
Ed25	-1.16(-2.75,0.63)	0.60(-1.23,2.64)	-2.37(-4.05,-0.46)	2.90(0.80,5.25)	1.28(-0.63,3.41)
EdStL	2.54(-1.02,6.92)	1.81(-1.56,5.93)	0.00(-3.31,4.14)	2.68(-0.96,7.16)	-1.39(-4.16,2.00)
Gl20	0.08(-0.19,0.36)	0.18(-0.10,0.48)	0.09(-0.19,0.38)	-0.17(-0.46,0.12)	0.32(0.02,0.61)
Gl51	-0.09(-0.40,0.23)	-0.15(-0.46,0.18)	-0.07(-0.40,0.27)	0.01(-0.31,0.33)	0.17(-0.16,0.50)
Gl95	-0.40(-0.90,0.12)	-0.40(-0.89,0.11)	0.17(-0.30,0.66)	0.20(-0.26,0.67)	-0.07(-0.54,0.41)
GlCen	1.30(-4.04,8.90)	-2.46(-6.05,2.35)	0.97(-2.96,6.03)	-1.29(-4.17,2.26)	0.35(-3.48,5.29)
Ischaemic Heart Disease					
Ab2	0.00(-0.35,0.36)	0.13(-0.31,0.58)	0.01(-0.46,0.49)	0.35(-0.12,0.84)	0.32(-0.15,0.80)
Ab3	-0.68(-2.04,0.81)	-0.29(-1.63,1.17)	0.56(-0.78,2.00)	-0.90(-2.16,0.46)	-0.77(-2.04,0.60)
Ed14	-0.11(-0.40,0.18)	-0.33(-0.62,-0.04)	0.22(-0.07,0.52)	0.16(-0.13,0.46)	-0.30(-0.59,-0.01)
Ed24	-5.63(-8.20,-2.29)	-2.38(-5.37,1.41)	-2.49(-5.31,1.03)	1.36(-2.15,5.71)	0.58(-2.28,4.01)
Ed25	-1.12(-2.12,-0.06)	-1.20(-2.21,-0.11)	-0.23(-1.40,1.03)	0.75(-0.42,2.01)	-0.73(-1.81,0.43)
EdStL	-1.21(-3.17,1.05)	1.14(-1.09,3.68)	-1.63(-3.72,0.79)	0.00(-2.22,2.57)	-0.95(-2.94,1.34)
Gl20	-0.06(-0.22,0.10)	-0.02(-0.19,0.15)	0.08(-0.09,0.25)	0.00(-0.16,0.17)	-0.03(-0.20,0.14)
Gl51	0.06(-0.12,0.24)	-0.15(-0.33,0.03)	0.05(-0.14,0.24)	-0.08(-0.26,0.10)	0.04(-0.15,0.23)
Gl95	-0.08(-0.37,0.20)	0.03(-0.25,0.31)	0.05(-0.22,0.34)	0.04(-0.24,0.32)	0.06(-0.20,0.33)
GlCen	-4.42(-7.01,1.13)	0.50(-2.54,4.19)	-0.69(-3.05,2.07)	1.02(-1.14,3.49)	-0.79(-3.14,1.97)

a. Percentage Relative Risk (95% Confidence Interval) per 10µgm⁻³ increase of pollutant on any single day within the lag period.
Mon = Pollution Monitor (Ab. Aberdeen Ed. Edinburgh Gl. Glasgow)

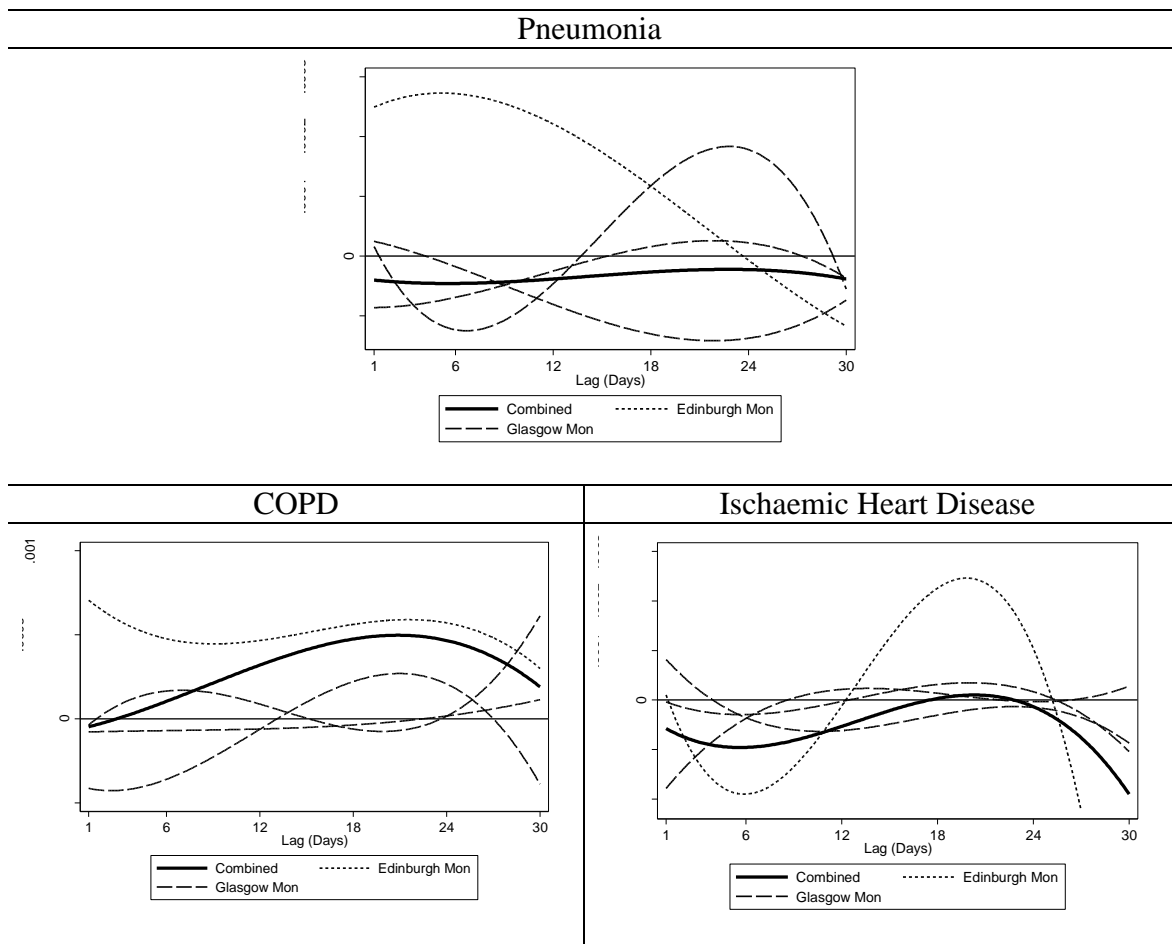
The variation not only in the direction of the effect, but the magnitude, and the shape across the lag period can be seen more clearly in the plots of the cubic polynomial function for each monitor used in the study. In addition, to each monitor individually, Figure 8.13 also reports the cubic function associated with the combined result from the main analysis for each cause of death (thick solid black line). The wide variation in the results makes it difficult to distinguish the main analyses from the horizontal x-axis causing the change in risk over the lag period to be undetectable.

Figure 8.13 – Monitor specific cubic distributed lag plots for increases in black smoke on risk of cause specific mortality (Pneumonia, COPD, and IHD) using all study monitors.



To get remove any bias due to a small sample size, the four monitors with the largest sample size as identified from Table 4.3 as Edinburgh Mon 14, and Glasgow 20, 51, 95 were plotted on their own in Figure 8.14.

Figure 8.14 - Monitor specific cubic distributed lag plots for increases in black smoke on risk of cause specific mortality (Pneumonia, COPD, and IHD) using four largest study monitors only.



Again the results associated with the main study analysis are reported as a solid black line; with the change in scale they are more visible here than in Figure 8.13. Firstly, Edinburgh monitor 14 results for pneumonia risk match the previously reported results in Chapter 5. Indicating the results have remained consistent despite any differences in the two analyses. Comparing just the four largest monitors indicates no common pattern even when restricted to the three monitors located in the same city for the same time-period.

9 DISCUSSION

9.1 Introduction

In March 2014 the World Health Organization estimated that 3.7 million deaths a year are linked to outdoor air pollution, 580,000 of which are located within Western Europe.¹ As recently as February 2016, The Royal College of Physicians went to great lengths to highlight the burden of air pollution on ill-health to the public, industry, and policymakers.¹⁸ Specific causes of death such as pneumonia, ischaemic heart disease, and chronic obstructive pulmonary disease are strongly linked to air pollution and are important considerations in public health. This along with their high prevalence and short induction period between initial contact, first symptoms, and final outcome makes them ideal candidates when modelling the influence of air pollution and temperature.

The work presented here focused on improving the accuracy of any relationship between monitored data related to pollution and temperature exposure and records based mortality. This involved applying techniques to reclaim dropped or missing independently monitored exposure data, stratifying the study participants to reduce interference between subject and exposure using subject matched hospital admission data, and analysis techniques that account for characteristics of the data that may distort the exposure-mortality relationship. Hopefully, improving on the already available information regarding change in acute risk will provide further insight into the underlying pathogenesis of the condition, or possible prevention and preparation measures that could be employed by the health service and policymakers overseeing public health. This might include:

- Hospitals can better predict health care demands due to episodes of extreme pollution and temperature.
- Policymakers could more accurately predict the burden of ill-health due to air pollution and hence take steps to mitigate the risk.
- Doctors and patients at increased risk will be better informed about the need for individual precautions.
- Researchers analysing short-term time dependent factors will be better informed regarding some of the potential sources of bias and misclassification.

As by way of an example, similar studies reporting the effect of extreme temperature have led to Public Health England producing both a ‘cold weather plan’ and a ‘heatwave’ plan that identifies the procedure for alerting, monitoring, and preventing the negative effects of extreme temperature on both the public and NHS.^{340,341} These along with other government initiatives such as ‘The Warm Front Scheme’, equivalent Scottish ‘Home Insulation Scheme’, or the winter fuel payments have helped those more vulnerable in the population to weather the effects of the temperature. Similar official schemes or guidance could therefore be implemented for management of extreme pollution days.

The primary aim of this large scale observational epidemiology study was to describe and compare the delayed effect of air pollution on acute mortality risk from specific causes of death, namely pneumonia, chronic obstructive pulmonary disease, and ischaemic heart disease (see Chapter 2.13 for full objectives). This was approached in a number of ways the results of which will be discussed.

9.2 Strengths

Increased sample size due to length of the study period

To achieve the study objectives independently monitored pollution and temperature exposure data were matched to mortality and hospital admission data. To maximise the sample sizes all subjects based in major cities in Scotland between January 1980 and December 2011 with reliable exposure data were included in the study upon death. Though dependent on exposure measurements being available, a study length of 30 years has resulted in one of the largest pollution and temperature studies not investigating long-term influence but rather acute effects on mortality.

The 30 years covered was limited by the availability of pollution data, with only BS and SO₂ covering the majority of the time period. Including pollution (BS & SO₂) from 1980 until the mid-2000s does not necessarily limit these results to a historical interest only. Given one aim was to compare the effect of pollution on subjects in the community versus within hospital, as long as the linear exposure-response assumption holds the higher outdoor concentrations of black smoke in the 1980's and early 1990's were advantageous for testing this hypothesis. Higher pollution concentrations meant the incidence of pollution related mortality would have been higher making any difference in risk associated with indoor and outdoor exposure easier to detect. The dominant source of black smoke during both the study period and present day were from road vehicles as most smoke control procedures to reduce combustion of coal for domestic heating and industrial energy would already have been implemented by the start of the study period. However, the chemical composition of fine black particles may have altered somewhat during the study period as a result of technological changes in vehicle engine and emissions control systems. It is not possible to directly characterise the extent of such changes as it is not possible to selectively collect black particles from non-black particles during atmospheric sampling for chemical analyses.

With short-term pollution effects expected to be relatively small, large sample sizes are important in order to maximise the power and reduce the chance of a false negative effect occurring. Currently pollutants such as NO₂, PM₁₀, and PM_{2.5} have been

recorded in limited areas for a period of 5 to 10 years (within this study period). Though larger than most studies currently available, it may be several years before larger sample sizes are available allowing more confidence in the results, reducing the confidence intervals seen here of those pollutants with smaller samples (i.e. PM_{2.5}). Even with a larger sample size the number of days containing high pollution concentrations was still few. The results of this study indicated that outliers had only a minor effect on the final result, if true then additional higher concentration days should have little influence on the effect estimates other than improving power. Even without a standard power calculation the greater sample size has allowed for the inclusion of more complex techniques modelling both pollution and temperature simultaneously.

Maximising the sample through the use of any cause of death fields

The sample size was maximised further through the use of both the ‘primary’ cause of death field and all ‘secondary’ causes of death fields as designated by the mortality dataset. The ‘primary’ field is the identified ‘underlying’ cause of death i.e. the cause starting the chain of events leading to death, and the secondary fields are all other causes in the chain. This was done primarily to identify all study subjects who had died from the three causes of death. In epidemiological studies where the primary effects are related to short-term changes in an exposure, the underlying cause becomes irrelevant when the induction period is longer than a few weeks or months e.g. most forms of cancer. When the delay is short between exposure, symptoms and outcome due to an already present condition (e.g. COPD) or a short induction period (e.g. pneumonia, and ischaemic heart disease) the immediate cause of death becomes more relevant. By ignoring causes located within the chain, misclassification of subjects cause of death is more likely to be present resulting in causal effects becoming underestimated.^{342,343}

The ‘immediate’ cause of death, i.e. the final cause in the chain of events leading to death was not identified in the mortality dataset and so ‘any’ cause of death field was used. All three causes of death have short risk periods meaning anyone with that cause of death in any field was likely to have had that cause as their ‘immediate’ cause of death, or at least one that was very near the end of the chain. This also meant misclassification was reduced as those with cancer as the underlying cause but

pneumonia as the immediate would still be included in analysis. As noted earlier, the cause of death was commonly determined by the underlying cause of death i.e. the cause that began the chain of events leading to death.^{30,80,163-165}

Only one environmental epidemiology paper in 1990 examined ‘immediate’ and ‘underlying’ causes of death separately, with two papers similarly including associated/contributing causes of death i.e. they used the primary and secondary fields.^{83,171,172} In the paper to model subjects with underlying only, and underlying and associated cause of death combined,¹⁷² circulatory disease effect sizes and standard errors changed very little between underlying only and underlying and associated together whereas respiratory effects and standard errors were smaller when underlying and associated were combined. This was also true in the study here, if you compare the width of the confidence intervals of the pollution effects on the three causes of death in the results for ‘any’ cause of death field (Table 8.1) compared to the ‘primary’ only cause of death field (Table D1 of Appendix D) they are very similar for IHD but reduce for COPD and especially Pneumonia. This is a reflection of the change in respiratory sample size from primary only to any cause of death field where the IHD sample size increased by 27%, COPD increases by 200%, and pneumonia increases by 330%.

Study Design

The case-crossover study design is an attractive method to gauge the effect of air pollution. Time-series models adjust for confounding via inclusion of extra modelling terms that try to optimise the model fit. This can make them complex and highly data specific. The case-crossover method elegantly controls through the design confounding by season, long-term time trends, and day of the week changes. This reduces the required number of terms in the regression model to those relating to pollution and meteorological factors resulting in a simpler regression model.

The increased usage of the case-crossover method meant comparisons have been made with results from the Poisson regression time-series models. Lumley and Levy (2000) showed that under certain circumstances the case-crossover design can give identical

results to the Poisson regression.²²³ Early simulation studies compared the various time-series regression techniques with case-crossover design under differing control strategies (though not the time-stratified model).²⁵⁴ In both the time-series models and the case-crossover designs the underlying model choices had strong influences on the effect estimates. Results of the bidirectional case-crossover analysis gave effect estimates closest to the true estimates though with standard errors larger than the time-series results, possibly indicating the lower power associated with a case-crossover design.²⁵⁴

Navidi went on to show the equivalence formally, in that for the bidirectional case-crossover design the likelihood function defined as the product of each subjects probability of failure at time t , could be equivalent to the Poisson regression. If all subjects have the same time frame in which a possible failure could occur and the same level of exposure then the likelihood function of the bidirectional design is equivalent to a product of independent, identically distributed multinomial trials. The Poisson regression conditions on the total number of failures, the vector of which is a distributed multinomial with a probability of outcome that gives an identical likelihood function to the bidirectional case-crossover. The maximum likelihood estimate of the relative risk would therefore be identical.²⁵³ The identical estimates of relative risk continue in a time-stratified design when the Poisson regression model includes indicator variables equivalent to the stratification. Navidi gives the example of stratifying by all days that fall on the same day of the week. If the Poisson regression also includes an indicator variable of day of the week then the likelihood functions will again be equivalent.²⁵³ Lu and Zeger (2007) later confirmed they can be equivalent when the choice of control strategy in a case-crossover design is matched by the method chosen to control for time varying factors in a Poisson regression model i.e. a step function of time for the time-stratified case-crossover design.²⁵² The reduced power associated with the case-crossover design may be due to over-dispersion in the Poisson variance due to influential outliers not being accounted for in the conditional logistic regression.²⁵²

Even though the case-crossover design contains a slightly reduced power,²²¹ as the effect estimates would be equivalent it was chosen ahead of the time-series modelling. This was reasoned as comparisons of the number of deaths on each day can make it

difficult to perform time-series analysis in small populations. Whereas, the case-crossover design uses data on individual subjects allowing for their use in smaller sample sizes.³⁴⁴ This was thought useful when the analysis was stratified further from already small samples, for example in the analysis investigating the association between PM2.5 on COPD in subjects with ‘all 30’ days of exposure in hospital. This also meant comparison test was possible to measure a significant effect difference of subject specific factors such as location during exposure. The reduced need for additional parameters in the regression model was also thought an advantage, what with the already complex analysis procedure.

9.3 Reducing bias

The study went to great lengths to reduce as much bias as possible occurring between exposure and outcome. This included attempting to reduce any interference between exposure and subject due to location, removing bias caused by influential outliers, and accounting for missing data.

9.3.1 Misclassification of exposure due to hospital admission

Assignment of the exposure measurements from a single fixed site monitor to the entire population has always been considered flawed. Spatial variation in pollution exposure across the city and differing participant activity patterns during the exposure period can contribute to misclassification. Exposure misclassification i.e. assigning a higher or lower pollution measurement than was experienced may lead to bias when estimating the true effect. In situations such as air pollution studies where effect sizes are small; this may increase the likelihood of a false conclusion. In an attempt to improve the pollution-response relationship it was hypothesised that the hospital admission information for each subject could be used to identify the length of time in hospital during exposure. By removing those subjects in hospital for any time during the exposure period the effect estimates calculated for those remaining ‘community based only’ subjects would better reflect the true exposure effect. Later, the length of time in hospital was also thought to be important as those subjects in hospital for longer (i.e. all 30 days) may represent differing underlying characteristics. The analysis was therefore stratified by the subjects’ hospital admission status during the exposure period.

The initial analysis resulted in a paper (full version in Appendix B) and is reported in the results Chapter 5. This analysis classified the subjects into community based subjects only i.e. those who had spent zero days in hospital during the exposure window and compared them with all subjects. The result indicated stronger immediate effects, reporting significant 0.44% and 0.28%, and a non-significant 0.31% increase in the risk of pneumonia mortality within the 1-6, 7-12, and 13-18 day lag periods. Indicating that we may be underestimating the effect estimates when including subjects who were in hospital during the analysis. Similar analyses were carried out for pneumonia, COPD

and IHD in the main thesis dataset where the classification was refined to represent three groups; community based only (zero days in hospital), hospital based only (all 30 days), and occasional hospital attendance (1-29 days).

When comparing the three classification groups the greatest effect estimates were often observed in the 'hospital based only' subjects. This may relate to the sample size as for example the PM_{2.5} analysis for 'hospital based only' pneumonia and COPD subjects have some of the smallest sample size in the study, which often resulted in larger effect sizes and wider confidence intervals. Even though effect sizes are generally smaller than those observed in the 'hospital based subjects' the effect estimates observed for 'community only' subjects are often greater than those reported for 'all subjects', particularly in those pollutants with a larger sample size and when effect estimates represent an increase risk. Indicating some evidence was still present for the conclusion observed in the earlier analysis i.e. that including those subjects in hospital may cause an underestimation of the effect size. For example, the lag 13-18 day 'black smoke' effect on all three cause of death increases from 0.02%, 0.14%, and 0.07% to 0.10%, 0.39%, and 0.13% for pneumonia, COPD, and IHD respectively.

Previous work on indoor and outdoor exposure levels

This study finds a similar increase in BS exposure effects when compared to U.S. studies of TSP/PM₁₀ on all-cause mortality when restricted to deaths located outside of hospital.^{83,84} Black smoke contains finer particle fractions dominated by combustion emissions which are more closely associated with health outcomes than PM₁₀ or PM_{2.5} and so the BS filter darkness measurements are a better marker for harmful combustion-related particles.⁵¹ Increased black smoke measurements are also representative of an increase in the toxicity level. As the measurement of black smoke is based on the percentage of light reflected the darker blacker the colour the more toxic the pollution measured is likely to be. A darker colour relates to a more harmful makeup (e.g. predominantly PAHs or trace metals), whereas a lighter whiter colour may represent less harmful components such as sodium chloride, sulphate, or ammonium.^{36,37} Currently, few published articles comparing indoor and outdoor BS levels exist. Gotschi et al. compared indoor and outdoor BS and PM_{2.5} for 186 homes in Athens, Basel,

Helsinki and Prague. The median indoor-outdoor ratios of BS were slightly less than PM_{2.5} however, Spearman correlation coefficients were larger, possibly due to stronger indoor influences on PM_{2.5}.³⁴⁵ Hoek et al. gave filter darkness regressions slopes (0.63-0.84) between indoor-outdoor concentrations in homes of four European cities.³⁴⁶ Limited information is available comparing indoor and outdoor personal particulate matter exposure but with inconsistent outcomes due to small sample sizes (N≤50).³⁴⁷⁻³⁴⁹ Janssen et al. investigated personal, indoor and outdoor fixed site exposure to PM₁₀ in 37 participants from Amsterdam and PM_{2.5} and BS in 36 and 46 participants from Amsterdam and Helsinki respectively with sampling taken for 24hr periods, bi-weekly, over six months. In both cities the median concentration levels were calculated for the personal monitors, the outdoor fixed monitoring and the indoor fixed monitoring. High correlations were observed between personal and outdoor fixed site monitors indicating that fixed site monitors are a good representation of the day-to-day variation in particulate matter exposure.^{350,351} However, high correlation between personal and outdoor monitors does not imply the same absolute level or change in absolute concentration levels of air pollution just that as one monitoring technique increases the other increases also. The underlying premise - that exposure to airborne pollutants was reduced in hospitalised subjects - is supported by Wang et al. which showed a reduction in indoor concentrations of PM₁₀ and PM_{2.5} in 2 of 4 hospitals in Guangzhou, China.³⁵² Subsequently, Wang et al. and later Morawska et al. further determined that a mechanical ventilation air conditioning system produced the lowest indoor-outdoor PM₁₀ ratios.^{352,353}

Indoor air quality is an important issue for hospitals. Currently the quantity of literature available on the relationship between indoor and outdoor air pollution regarding hospitals is limited. Air pollution within the home is strongly influenced by home cooking or in the developing world open source home heating. Meier et al. attempted in Switzerland to measure and compare indoor exposure to outdoor pollution³⁵⁴ whilst adjusting for indoor pollution source characteristics such as cooking and heating.³⁵⁵ High between site variability between indoor and outdoor pollution indicates that indoor exposure is largely dependent on indoor source characteristics. In an adjusted model, indoor PM₁₀, PM_{2.5}, PM absorbance, and NO₂ concentrations were, respectively 31%, 36%, 66%, and 37% of outdoor levels. Only NO₂ increased when the windows were

‘sometimes’ (49%) and ‘mostly’ open (61%). At 66%, PM absorbance is a measure of the light absorbance of PM_{2.5} and considered a surrogate for BS³⁴⁵ which confirmed previous indications that a large percentage of outdoor BS can infiltrate indoors.^{345,356} Note, unlike hospitals none of these homes had air conditioning. Several studies,^{357,358} such as Chen et al. (2012), have reported strong negative correlations (-0.64) between city specific mortality coefficients and percentage of homes with central air conditioning,³⁵⁹ though individual level rather than city level studies would be required to confirm the association.³⁶⁰ These provide useful information regarding the indoor levels of pollution. However, further observational studies are required to supplement understanding of reduction and fluctuations in indoor hospital air pollution concentrations in terms of distance from combustion sources, changes in ventilation systems and meteorological conditions.³⁶¹

Pneumonia related factors

These findings suggest that a subject’s location is an important consideration when assessing the effect of ambient air pollution on pneumonia mortality. One alternative explanation for the increase in relative risk occurring within the CDP group is that exposure may interact differently with underlying causes of pneumonia that are more prominent within the community i.e. community acquired pneumonia. Pneumonia infection can be caused by a variety of factors with ‘hospital’ and ‘community’ acquired pneumonia commonly caused by different micro-organisms,¹¹⁹ each with differing incubation periods.¹²⁰ This variation may be a contributing confounding factor to the difference in the pollution effect on community based mortality. The lower risk experienced by hospital based subjects may be due to increased access to medical care that may have been able to identify early symptoms and implement preventative measures to delay or stop progression to a critical phase. The bias associated with earlier recognition, and more timely and aggressive therapeutic interventions may be the reason for the apparently negative %RR associated per increase in BS of 10 $\mu\text{g}\text{m}^{-3}$ within the ‘hospital based only’ subject group.

Chronic obstructive pulmonary disease related factors

In contrast in the COPD subjects, the 1-30 day lag effect estimates appear to be similar across the three hospital admission groups. The differences between the three hospital admission classifications occurred within the 30 day lag period. Those who spent '1-29 days' in hospital reported that the strongest risk occurred in the 1-6 day lag, in those with 'all 30 days' it occurred in the 6-15 day lag and for 'zero days' in hospital the 15-24 day lag. This delay in peak risk for COPD may be reflective of the influence of two factors, possibly in combination, the relative frailty of the COPD subject and their access to health care. COPD is a chronic condition and so already represents a frailer subgroup of the population. As COPD progresses it is safe to assume that the number of hospital visits will also increase. Study participants who have spent all 30 days in hospital are likely to be frailer than those who spent less than 30 days, who in turn maybe frailer than those in hospital for none of the 30 days exposed. The increased frailty may be due to a recent exacerbation, a general decrease in COPD related health, or the influence of a comorbidity. Unlike the other subject location groups, those in hospital for all 30 days will also have immediate access to health care that can identify symptoms early and help prevent or delay further progression. Hence the immediate risk seen in those spending some but not all 30 days in hospital followed by the delayed effects in the 30 days in hospital group. The stronger COPD cases who have been affected but do not have quick access to health care may succumb to an exacerbation of their symptoms before health care can be accessed.

Ischaemic heart disease related factors

The main analysis (Chapter 8) indicated a delay in risk for ischaemic heart disease with respect to increases in particulate pollution that contradicts the expected immediate (within days) effect. For 'black smoke' exposure at least, this may be due to hospital health care playing an intermediary role either through reduced exposure in an air conditioned environment or due to the immediate access to health care as previously noted. Community based subjects report a small but increased risk during the first days, whereas the non-community based subjects report a stronger decrease risk. Acute

ischaemic heart disease events usually occur in the form of a myocardial infarction. Maynard et al. (1989) reported an increase in survival from acute myocardial infarction when time to hospital was reduced to less than two hours.³⁶² Other studies have also reported increase survival when treatment has been obtained quickly,^{363,364} even identifying the first hour as a 'golden hour' for thrombolytic treatment and survival.³⁶⁵

In contrast to mortality, significant increases in short-term hospital admissions numbers have been observed with increases pollution. Early studies (pre-1995) into black smoke (also SO₂ and NO₂) have shown a strong association with same day and up to lag 3 increases in ischaemic heart disease hospital admissions, with percentage increase typically around 2.0%-2.9%.^{48,366} Later, in several European cities a 1.1% (0.7%, 1.5%) increase in ischaemic heart disease hospital admissions was observed per 10 $\mu\text{g}\text{m}^{-3}$ lag 0-1 average black smoke increase.²¹⁹ More recently the particulate pollutant metrics particularly PM_{2.5}, and the two gaseous pollutants have also been shown to increase hospital admissions at same day (e.g. PM_{2.5} increase for lag 0 = 0.27%) and following day (PM_{2.5} increase for lag 1 = 0.16%) for ischaemic heart disease.^{154,158,367-369} In contrast to BS, in this study the two size specific particulate matter pollutions (PM₁₀ and PM_{2.5}) behave similarly to each other but also across the three hospital admission groups. In both cases the increased risk only occurs in the 1-29 days in hospital subjects after ≈ 6 days and lasts until ≈ 18 days. The community only and hospital only subjects do not report an increased risk until later in the lag period. It would appear here that location (in or out of hospital) might be having very little influence on the relationship between pollution and IHD mortality.

Even with a significant increase in hospital admission for heart disease it has been estimated worldwide that 75% (66.9% in this study Table 4.1) of all heart disease deaths occur outside of hospital.³⁷⁰ If hospitalisation during increased air pollution exposure is having a strong protective effect then increased awareness, early detection of symptoms and quick admission to hospital may go some way to mitigate the influence exposure is having on cardiovascular disease.

Other potential factors

Though it is unlikely to have impacted on the results of this study, it is worth noting the recent addition to hospitals of positive pressure isolation rooms. Originally intended as an isolation room for infectious patients they are often used for those at risk from getting an infection rather than being a carrier. These rooms are kept at a positive pressure between 8 and 12 pascals,³⁷¹ assuming the (outdoor) air used to maintain the pressure and supply to the occupants is clean, this represents an additional level of protection against airborne material³⁷² in a room more likely to be inhabited by a frailer group. Even though hospital based subjects may represent a frailer subject group, the additional protection and the immediate access to treatment in the ‘golden hour’ could be a deciding factor in increased survival during the first few days.

Outdoor temperature on hospitalised subjects

Currently the UK has no laws regulating indoor temperature at work or public buildings such as a hospital. The NHS does provide recommendations that homes are heated to minimum of 18°C, if aged over 65 or a health condition is present,³⁴¹ and hospitals are kept below 26°C during heat waves.³⁴⁰ However, few studies have investigated the impact of outdoor temperature on indoor temperature, or the resulting risk to ill-health.³⁷³ A small study based in Boston, USA and a secondary study based in a variety of climates around the world, compared indoor temperature to outdoor temperature that was measured at the nearest airport weather station.^{374,375} In both papers, regardless of location indoor temperature held within a range of 19-30°C and the indoor to outdoor temperature patterns were similar in that they held constant during the winter when outdoor temperatures dropped, but matched consistently with outdoor fluctuations during warmer summer months. Though the second study included a variety of locations around the world (e.g. Greenland, Brazil, and Vietnam) the results of both studies reflected the middle-class background of the participants which despite the lack of diversity, might better reflect the pattern experienced in a hospital.

Cold temperatures

The indoor to outdoor temperature patterns are reflected in the cold and warm temperature results of this study comparing community vs hospital based subjects. In all three causes the general pattern in risk at colder temperatures indicates a greater risk for those spending all of their time in hospital. This is particularly true at the short lag periods (<12 days) where for those in the community as outdoor temperature increased by 1°C, the risk of pneumonia, COPD, and IHD significantly decreased (-0.51%, -1.01%, and -0.71%), compared to those in hospital where smaller non-significant differences are present (0.22%, -0.20%, and -0.61%). The difference in effect between community based and hospital based appears to be smaller for IHD than the two respiratory diseases. The reasons for this are not clear though there is some evidence that even a small decrease in indoor temperature may impact on blood pressure, shortness of breath, and self-reported health,^{376,377} which may be important if an underlying health condition is already present. Of the three causes of death COPD reports the largest effect for cold temperature with short-term risk (lag 1-6) for community only (-1.01%), occasional hospital attendance (-0.89%), and hospital only (-0.20%) per increase in cold temperature. The hospital based effect is relatively small considering the possible increased frailty level of the group. In addition to improved access to health care, the reduced risk observed in hospital based subjects may be attributable to a reduced chance of COPD exacerbations due to a constant temperature experience. A study investigating COPD symptoms relating to home indoor temperatures indicated that maintaining a temperature of 21°C in living areas for at least 9 hours a day reduced the symptoms in non-smoking COPD sufferers.³⁷⁸

Higher temperatures

When ambient temperature increased, the effect observed between community and hospital based subjects varied much less with no significant differences occurring and generally similar effect estimates for both subject groups (Table 7.3 - 7.5). The more limited temperature range with fewer days in the extreme heat range mean warmer temperatures may impact less on indoor temperature exposure here. With focus on outdoor temperature, very little work has looked at the impact of indoor high

temperatures. There is some evidence that in economically more developed countries adaption to extreme heat through lifestyle changes and improved temperature management technology is having a protective effect even in the more susceptible elderly.³⁷⁹ Meaning, even if the subject is based in the community during exposure they may be taking measures themselves to avoid exposure to outdoor temperatures, hence the lack of a community-hospital difference.

One slight anomaly is the effect estimates associated with occasional hospital admission i.e. 1-29 days in hospital. Though generally non-significant, warm temperatures appeared to have less impact both protective and harmful than the 'zero' or 'all 30' days in hospital groups. Why this occurs is not clear though activity patterns in someone who has recently been ill and released from hospital may mean less time outdoors putting themselves reducing their risk of increased heat, other pathogens, or over exertion.

The true underlying mechanisms causing differing exposure effect patterns in the hospital admission groups are complex, with the combination of differing levels of frailty, access to health care, exposure between the three groups all playing a part. Even so, it appears clear from the results presented in this study hospital admission status is an important consideration when estimating the true effect of both pollutant and temperature exposure.

9.3.2 Missing data

Pollution effect estimates are traditionally small, particularly when compared to temperature. A small effect size mean large sample sizes are required to maximise the power and reduce the chance of a false negative result. In this study design, a missing exposure on any of the 30 days leading to death can mean the subject is set as missing if the missing exposure day is associated with the 'case' day or in all of the 'control' days. To increase the sample size and maximise the usefulness of all pollution and temperature data, a multiple imputation technique was employed to replace missing air pollution values.

Multiple imputation is a relatively new field of statistics especially in the context of time-series data and is reflected in a limited amount of literature. To gain some insight, a simulation study gauged the effectiveness of multiple imputation at reclaiming lost information whilst also determining the impact of missing data in a standard complete cases analysis. The pollution exposure data indicated two main characteristics of missing data, a summer to winter bias and the occurrence of missing data in 'blocks' or periods of uninterrupted days with missing observations. Other potential missing data characteristics were investigated including long-term time trends, day of the week, or weekend increases in missing data. No clear patterns were observed in the datasets available in this study and so were not included as simulation scenarios. That does not mean these, or other missing data characteristics are not present in air pollution exposure data as there are plausible explanations for them. For example, more recently employed automatic monitoring methods such as the Tapered Element Oscillating Microbalance (TEOM) for particulate matter metrics or the ultra-violet florescent continuous analyser for gaseous pollutants are thought to be more reliable than the previous manually operated monitors. Pollutants such as SO₂ which have been measured since the 1970s will have potentially differing missing data patterns depending on a new or old measurement device. Meaning long term time trends or weekend increases might be present making them important factors to be aware of in any future study of missing data. As long-term trends and day of the week effects are present in the pollution measurements themselves these factors were already included as part of the imputation model.

Each chosen missing data characteristic were applied under increasingly greater severity (e.g. Summer:Winter ratios of 50:50, 60:40, 75:25) along with increasing total percentage of missing data. The results of the complete cases analysis indicated bias when estimating the true effect was present and that it tended towards the null hypothesis increasing the likelihood of underestimating the true effect and reducing the power to see an effect. Bias increased as severity of the missing data characteristics were applied both individually and in combination, though it mostly kept within a predefined acceptable boundary. Bias appeared to improve slightly or hold steady when missing data techniques were applied. Bias in the main effect due to seasonal characteristics, though at an acceptable level, was still present even though season was included in the imputation model. Several attempts were made to improve this with little or an inferior result. The whole simulation study was performed in a real dataset with the missing data simulated. This is unusual as simulation studies often tightly control all aspects of the data. These results are therefore a reminder that results from a standard fully controlled simulation study may not easily translate to an observed empirical dataset.

Even though bias was observed in the multiple imputation analysis to at least match that of the standard complete cases analysis, the bias in the standard error and the temperature covariate were reduced. It was therefore concluded that multiple imputation would be performed in the main analysis and was reported for those located in the community only i.e. 'zero' days in hospital during exposure (Chapter 8.4). The change in shape across the lag period and the effect estimates associated with a $10\mu\text{gm}^{-3}$ daily increase in each pollutant was very similar before and after multiple imputation was applied. This was particularly true for larger sample sizes such as IHD mortality, though even in the respiratory diseases the change was small. Of the two modelling methods inclusion of additional data appeared to influence the distributed lag model to a greater extent than the lag stratified model. This may be due to the cubic shape being applied onto the full lag length unlike the lag stratified model which represents step changes driven by the data at smaller lag intervals. Where increased data at lag 1-6 only affects the effect estimate for lag 1-6 and not lags 7-12, 13-18, etc. in the cubic function additional data at lag 1-6 may also influence the cubic shape further into the 30 day lag

period. As the influence appeared to be small here it is of little consequence, however a more flexible polynomial may be less affected by the additional imputed data.

The similarity between the main study multiple imputed and complete cases effect estimates is reassuring. The agreement with both the complete cases and the results reported by the simulation study indicate any additional underlying missing data characteristics were either not present or accounted for equally by the imputation model. As with the simulation study the confidence intervals did tighten indicating as expected improved precision though the improvement did not result in any changes in significance levels, as very few were close to being borderline this was not unexpected.

Previous methods dealing with missing pollution data

Missing pollution exposure data has largely been ignored in the air pollution literature meaning ‘complete cases’ analysis has been the common analysis method.^{13,212,218,380,381} If the missing values are not considered to be ‘missing completely at random’ (MCAR) or ‘missing at random’ (MAR) given the covariates and a maximum likelihood model. Then performing a ‘complete cases’ analysis is likely to produce biased estimates, and possibly suffer from overfitting if smaller sample sizes are present.³⁸² In cases where the percentage of missing exposure data were small employing a method such as multiple imputation would add extra complexity to the analysis without any appreciable gain in precision or efficiency.³³⁸

Often in studies where multiple monitors have been present within the city of interest, missing data has been reduced through replacement, often by combining the results of available data from other monitors into an average.^{137,148} One method proposed by the APHEA project, was to replace a missing value for a given day by combining the recorded observations from other concurrently active monitors within the city into a mean and weighting by a factor equivalent to the ratio of the annual mean for the station with missing data and the annual mean for the other active monitors on the same day.²⁰³ If observations were missing for all monitors that day then the observation was considered missing.^{135,383-385}

In some cases single imputation methods such as regression models have been employed, though often with little detail given.²⁴⁷ Carder et al in 2008 and later in 2010 employed multiple imputation techniques to account for missing black smoke data from a single monitoring site within each of the cities Edinburgh and Glasgow.^{272,273} In which a univariate regression imputation technique was employed where the covariates represented information from the monitor of interest and alternative monitors running concurrently during the same time-period. Carder et al. ran a sensitivity analysis comparing estimates produced with and without the multiple imputations. In general the coefficients obtained without imputations were slightly smaller than the coefficients when imputation was employed but the confidence intervals overlapped indicating a non-significant change similar to here.²⁷²

In a few cases imputation methods were compared and/or new methods proposed to take account of missing air pollution data. Junninen et al. 2004, in a simulation study compared results of several imputation methods for air quality data.³⁸⁶ The imputation methods included, nearest neighbour, univariate and multivariate linear regression, a hybrid of multiple methods, and multiple imputation in the form of joint modelling. Similar to the simulation study reported in this thesis the air pollution data with simulated missing data were originally from Belfast and Helsinki and contained hourly and not daily average estimates. Prediction models were not clearly defined; however potential covariates included other pollutants, meteorological variables temperature, wind speed, relative humidity, and time in the form of cosine and sine parameters. Results indicated that multiple imputation was the most accurate even under shifting conditions such as varying levels of complexity in the missing data, and short vs long missing data gaps.

Plaia et al. 2006, attempted to propose a new single imputation method and compared it to other imputation techniques: mean replacement, nearest neighbour, and multiple imputation with chained equations.³⁸⁷ Varying missing data patterns were simulated in hourly PM10 data across eight monitoring sites in Palermo, Sicily during 2003. The data contained a multilevel structure with individual observations as the first level and monitor as the second level. Plaia proposed a Site Dependent Effect Method (SDEM)

which accounts for site specific effects such as differences in the week day and hourly levels of each monitor, and then compared with the other single and multiple imputation techniques. The paper indicated that the method proposed showed an improvement in the imputed data compared to the true observed data, even when compared to MI. However, the standard deviations of MI and SDEM were equivalent and the measure of accuracy (root mean square deviation) actually showed a slight improvement in the MI model. Further, it was unclear what predictors were employed in the MI model and what attempts were employed to account for the monitor clustering. The PM10 data were collected hourly from eight monitoring sites running concurrently, this method is ideally suited for these characteristics and may not be suitable when one monitor is replaced by a second over a long time period, as is the case in our data. As would be expected (see Chapter 2.12.8) the MI performed the best when compared to the standard single imputation methods.

Molitor et al. 2006, proposed a Bayesian Markov chain Monte Carlo estimation method to account for measurement error and missing data in NO₂ exposure data when related to lung function in southern Californian children in 2000.³⁸⁸ They also employed frequentist methods such as single and multiple imputation for comparison when modelling NO₂ effects on lung function. No comparison statistic was given, however effect sizes and confidence intervals given for the Bayesian model and MI were similar. Norazian et al. 2008, subsequently indicated that of several single imputation techniques a linear interpolation technique ‘mean of the day before and day after’^{389,390} was considered preferred in PM10 data for Seberand Perai, Penang Malaysia.³⁹¹ Though they briefly discuss the best choice of mean replacement they did not perform MI, report the complete analysis, compare standard errors with the true observed data, or discuss the limitations of the methods they used i.e. the lack of random variation. They also simulated the missing data completely at random, which should cause little bias in the estimates when a complete cases analysis is performed, unlike MAR.

In summary, missing air pollution exposure data has largely been dealt with in one of three ways. The ‘complete cases’ analysis which have been shown in this study to underestimate the true effect estimate, an adjusted mean replacement which may incur

bias if only a few monitors are active or do not accurately represent the monitor with missing data, and early attempts at limited multiple imputation analysis.

Limitations and further work

This study applied the missing data technique ‘multiple imputation’ in a time-series context. Certain characteristics of the data and analysis methods may have contributed to inconsistent results and require further consideration.

The MAR assumption is important but untestable and relies on the judgment of the researcher and so it is possible that MAR may be violated in pollution studies and MNAR be present. This may occur if the concentration of air pollution is so great or so low that the monitor cannot measure the value, or in the case of so great the value has been censored and not reported. It is unclear if this is present in the exposure data as current procedure regarding data collection and reporting is unclear. However, the data itself does indicate that ‘black smoke’ values of zero were recorded which would be extremely unlikely, and certain dates are missing where high pollution would be expected e.g. new year’s eve or bonfire night. The multiple imputation method employed here has attempted to account for these two characteristics and so MAR should be present, but it does indicate possible departures might be present.

In this study, the same imputation model structure was applied to all pollutants and outcomes. The characteristics of the exposure dataset available meant that time dependent variables such as date, year, month, and day of the week were the primary covariates in the model. There was little scope to add further information as all time-dependent, pollution, and temperature information was included as explanatory variables of the imputation model. Though the results are not shown, any attempts in the simulation study to manipulate those covariates only resulted in increased bias. Even so, some factors that were not tried and may need further investigation are:

- The limited availability of suitable monitors for the study meant only Glasgow included concurrent pollution measurements. Can other monitors regardless of activity period be included to improve the imputations?
Inclusion of other pollutants (e.g. SO₂ in a BS model) measured concurrently in the same monitor or different monitors to aid prediction.
- Though none appeared present, is autocorrelation in the pollution & missing values present in the 30 days leading up to and following the imputed day influencing. It was felt the added information on short-term changes in pollution helped predict the missing data. Further study may be required to determine if better strategies are available i.e. are 30 days either side too many, or could a lag stratified approach be a better option?
- Including the outcome variable in the imputation model was important to maintain the relationship between exposure and outcome in the imputed data.²⁹⁸ Here the overall count of the number deaths per day, plus counts of the number per day of deaths for each cause of death were included. Should the outcome be more specific to the cause of death.

In addition, the procedure firstly imputed the same day (lag 0) missing exposure value, then generated the lag stratified and distributed lag covariates. These are known as ‘passive’ variables and it has been recommended that these passive variables should in fact be included in the imputation model and imputed directly rather than be built separately. In order to make sure any short-term patterns in the pollution exposure measurements were kept the passive variables were generated subsequently. This however the most optimum procedure is perhaps an area worth investigating.

There were two levels of clustering within the analysis model that at the time of the study were difficult to account for in the imputation model. Use of several pollution monitors active for varying time periods and in varying locations around the city meant inherent differences in the pollution measurements between monitors. Imputed values would also need to replicate the within monitor and between monitor correlation. Incorporation of the multilevel (clustering) structure into multiple imputation techniques, particularly in observational data, is an ongoing area of research. Current attempts are in their early stages and are not incorporated in main stream statistical programs. The Realcom freeware package was considered but had limitations applicable to this dataset (see Chapter 2.12.16) and was not compatible at the time with the main

statistical package. Instead multiple imputation with chained equations using the cluster option was chosen and as introduced in Chapter 2.12.16 this treats the between monitor observations as independent. Employing pollution measurements in the pre and post days as predictors was thought to provide additional information and improve the representativeness of the imputed values. Even so, as the percentage of missing data increases, then as seen in the simulation study the representativeness is likely to be compromised especially in smaller monitors. As the main analysis employed a two-stage independent meta-analysis where stage one modelled the data for each monitor individually and the second stage combined in a random effects meta-analysis, the multiple imputation with a cluster option was thought adequate.

The analysis was then performed in a time-stratified case-crossover design using a conditional logistic regression clustered around the subject. The imputation model however was run in a time-series format where each row related exposure on a day in the study. The outcome is an important covariate in any imputation model and so in order to maintain the link within the imputed values the number of deaths on each day was included. Imputing in a separate time-series data and not in the case-crossover format was not ideal as within cluster correlation would not be present, however as the exposure data were measured independently without any subject influence it was thought the affect would be minimal. This will require further work to confirm. Currently the literature appears to be limited with most focusing on the analogous matched case-control design. A recent simulation study indicated that multiple imputation with chained equations appeared to produce the least bias in odds ratio estimates.³⁹² However, it relies on including information related to the matching between subjects in the imputation model, a situation not applicable here as even though it is an example of a matched study the subject is matched with themselves and any differences are time-dependent which were already included in the imputation model.

In summary, though the multiple imputation study performed here was able to account for a number of issues and made important progress within environmental epidemiology and time-series data, there are still a number of questions left to answer. Further work is required but time and computational restrictions meant the simulation

study was performed with a limited number of repetitions and used a single (lag 0) main effect estimate rather than covering the 30 day lagged effect. The results though, provide useful information regarding some of the potential complications that need to be addressed in time-series studies based on routine exposure data. With such small effect sizes even large sample size air pollution studies can lack the power to confirm an effect. As multiple imputation becomes more common in wider epidemiology and statistics it will be an attractive method in air pollution studies where reclaiming lost information and maximise sample sizes will be important.

9.3.3 Outliers – High pollution episodes

The early events, such as the London smog of 1952, identified air pollution as a threat to human health in situations where pollution concentrations were extreme. Since then, improvements in pollution reduction technology and the switch from home coal heating have seen the general background level of air pollution gradually decline. The challenge has become to determine the influence of low exposure concentrations yet the occurrence of high pollution episodes still occur, though even they have dropped considerably in the last 10 years.³⁹³ The concern here was that these rarer high pollution days (outliers) may be highly influential on the effect estimates produced in a pollution study.

Pollution exposure data follow a non-negative skewed distribution with a long tail, where the long tail represents extreme exposure measurements. The exposure levels in these tails may potentially have a disproportional influence on the effect estimate. To identify an extreme pollution exposure and to be consistent across pollutants the values in the greatest one percent (i.e. outliers were considered above the 99th percentile) were removed and a complete cases analysis repeated. The results were then compared to the full 100% complete cases analysis. The alternative to using the 99th percentile pollution level was to employ the government standard regulations for the UK or those defined by the Department for Environment Food and Rural Affairs (DEFRA). Table 9.1 compares for each pollutant in the study the current pollution standards as set out by UK legislation, recommendations regarding what constitutes a ‘high’ or ‘very high’

pollution episode day as defined by COMEAP and DEFRA, and the 99th percentile position for the study data. The lack of consistency across pollutants in terms of the averaging period, the lack of a black smoke recommendation, and the thought that government legislation levels may have included some form of negotiation all meant the 99th percentile was considered a better choice. In actuality, the 99th percentile for those that could be compared was very similar to both the UK legislation and the DEFRA guidance.

Table 9.1 – The UK legislated air quality standard,¹⁴ the DEFRA recommended limits,¹⁷ and the 99th percentile position for each pollutant within the study.

Pollutant (μgm^{-3})	Averaging Period	UK Legislation	DEFRA Guidance		Study 99th Percentile
			High	Very High	
Black smoke	24hr Mean	-	-	-	87
Particulate matter 10	24hr Mean	50	>76	>101	62
Particulate matter 2.5	Calendar Year	25	>54	>71	-
	24 Hour Mean	-	-	-	46
Sulphur dioxide	24hr Mean	125	-	-	118
	15 min Mean	-	>532	>1064	-
Nitrogen dioxide	1 hour mean	200	>401	>601	-
	24 Hour Mean	-	-	-	96

Removal of extreme exposure estimates appears to be rare, been done but not explained,^{390,394} or done but only to investigate the exposure-response effect²⁶⁴ and not compared with complete exposure data. In those that have detailed information, some chose a cut off value themselves (e.g. PM10 of $300\mu\text{gm}^{-3}$, or $\text{SO}_2 = 150\mu\text{gm}^{-3}$ BS = $200\mu\text{gm}^{-3}$)^{80,217} or they used government standards e.g. USA Environmental Protection Agency standard of $35\mu\text{gm}^{-3}$.³⁹⁵ Hong et al. (2002) removed outliers in the exposure data that were defined to be 6 standard deviations away from the mean, in this study that would set the cut off for black smoke at $129.9\mu\text{gm}^{-3}$.³⁹⁶ Others have excluded extreme outliers at both ends of the data with the greatest 1% and the smallest 1% being removed; here they investigated the lag 0-4 effect for PM2.5 and a coarse particulate matter (2.5-10 μm) on hospital admissions.³⁹⁷ Observing that the concentration-response across the lag 0-4 became linear rather than the previously slight quadratic curve and the lag 0-4 effect for PM2.5 but the coarse particulate matter increased slightly from 1.94% increase to 2.37% increase.

As with the missing data analysis the outlier analysis reported in Chapter 8.4 again focused on those located in the community i.e. zero days in hospital during exposure. The analysis of ‘all subjects’ combined can be found in the Appendix D. In summary the results tended to remain consistent with the original main analysis, in terms of both the effect size and the shape change over the lag period. The similarity was particularly strong for the particulate metric PM10 and the two gaseous pollutants SO2 and NO2, whereas black smoke and PM2.5 (in the respiratory diseases) did appear to be more susceptible to extreme outliers with some slight differences. However, in all cases the conclusions remained consistent in terms of the direction of the effect. Of the five pollutants covered why ‘black smoke’ and PM2.5 are affected is not clear. In both cases it appears to influence the short-term lags the greatest (lag 1-6 and 7-12), with the already present effect becoming more extreme. When a small sample size is present, it’s difficult to determine if an outlier is an outlier or part of an actual exposure range that is less sampled and the removal of any data may influence the result.

This should have little effect on black smoke estimates with a much larger sample size present. Black smoke has the largest pollutant exposure range followed by SO2. In both BS and SO2 the estimates located in the greatest 1% were measured in the 1980s with very few occurring since 1988. The concentration levels of many of the common pollutants such as those reported here, have gradually been declining,^{398,399} a decline that includes fewer days with high pollution episodes.³⁹³ Those pollutants which have only been recorded in the last few years are less likely to have outliers in an exposure range that impact disproportionately on ill-health compared to the lower pollutants hence very little change. Even so, the results indicated little effect on SO2 whereas black smoke has been. In all three causes of death the black smoke effect decreases in the short-lags (lag 1-6) and increases in the middle lags (lag 13-18). This may indicate that high concentrations overwhelm the system causing ill-health and mortality very quickly, compared to lower concentrations that may take a longer time to have an effect.

In all pollutants regardless of the effect when removed, it is difficult to determine if those in the greatest 1% should be classed as an outlier. With the highly skewed distribution of the data the 99th percentile position was located such that over 50% of the

exposure range (83% for black smoke) is included in the last 1% and the distribution of the values is uniform. The exposure values are evenly spread from the 99th percentile to the maximum with no sudden increase observed meaning that the result observed in the complete cases of 100% of the data is just as likely to be the correct representation just without a large enough sample size in the extreme exposure range.

In summary, this provides an early investigation of the influence of extreme pollution exposure on the lagged-response relationship effect rather than the exposure-response relationship. The results indicated that extreme exposure values had very little effect except maybe when sample sizes are already small or in the case of black smoke.

9.4 Results summary

The following summarises and discusses the results of the study beginning with temperature. Temperature is an important confounder in any pollution study, but also an important contributing factor to ill-health and mortality. The influence of pollution on the three causes of death pneumonia, COPD, and ischaemic heart disease are then discussed. Given the number of potential scenarios an analysis procedure was defined (see Chapter 3.9.3) that still resulted in multiple versions of results. After consideration, the results calculated after accounting for missing data in those with ‘zero’ days in hospital during exposure were chosen for the discussion. It was thought that they represented the most reliable estimate of the outdoor exposure-mortality relationship due the reduced misclassification in exposure and the maximised sample size with very little additional bias when missing observations were reclaimed.

9.4.1 Temperature

Though temperature, particularly the extreme temperatures, is an important confounding factor in pollution studies it is an important factor with its own recognised association with mortality.⁴⁰⁰⁻⁴⁰² The UK experiences a substantial increase in the number of deaths, especially in pneumonia, COPD and ischaemic heart disease mortality during the winter months.⁴⁰³ Therefore the influence of temperature on mortality was investigated thoroughly.

Exploring non-linearity

The analysis began by investigating the non-linear exposure-response relationship across the 30 day lag period using techniques that allowed for a data driven flexible shape unique to each cause of death. The traditional temperature-response relationship - a U, V, or reverse J shape - was replicated (Chapter 7.1). The surface plots confirmed a non-linear temperature-response relationship largely continued across the 30 day lag period. Though non-linear throughout, the shape was not consistent as at various stages

it transformed between the U, V, and reverse J shapes. The optimum temperature (i.e. the lowest observed risk) appeared to be between 10-15°C for COPD and IHD, and between 5-10°C for pneumonia. As expected in an area of the world with mild temperatures, in all three causes of death the cold temperatures showed a greater risk, particularly for COPD and IHD. The effects of cold temperature appeared to be immediate and last for the entire lag period. In contrast, the 'warm' temperature effects were much smaller and also occurred immediately but dissipated very quickly. This is most evident for pneumonia where a sharp immediate increase in risk appears to occur after 15°C that was as strong if not stronger than the cold temperatures. The warm temperature pneumonia risk dissipated quickly within 6 days.

These results similarly match to the relationships observed between temperature and general cardiovascular and respiratory disease as investigated in a number of different climates around the world.^{192,404-406} These studies indicated cold temperatures also report strong immediate effects that last for 2-3 weeks, whereas hot temperatures report an immediate effect which fades quickly within a few days.

Threshold models were created that identified specific temperature zones 'cold', 'mild', and 'warm' for each cause of death. These were generated independently and the most suitable model identified, creating unique 'bespoke' temperature models for pneumonia, COPD, and IHD. In all three causes the cold and mild temperatures connected at 1-2°C and mild to warm temperature zone connected at approximately 15°C, with IHD connecting at 17°C. The delayed lag effects were then modelled for the following 30 days with a ridged step function (lag stratified) and a smoother cubic polynomial function. Of all the analysis reported in this study, the effect estimates for temperature are, as expected, the largest and consistently significant. The following results all refer to those who spent the whole of the exposure period in the community (i.e. zero days in hospital).

Pneumonia & COPD

As with all three causes of death colder temperatures have a strong influence on risk of mortality from Pneumonia and COPD.

The pneumonia associated risk appears to last for three weeks (Chapter 7.4.1). As the average daily temperature increases by 1°C towards the 1°C threshold the percentage relative risk associated with mortality from pneumonia significantly decreases by -0.51% (95% C.I. -0.80%, -0.25%), -0.75% (-1.04%, -0.49%), -0.42% (-0.71%, -0.13%), and -0.34% (-0.63%, -0.05%) for the lags 1-6, 7-12, 13-18, and 19-24 respectively. The risk continues to significantly decrease but at a slower rate until the 15°C, where in the lag 1-6 days only there was a non-significant increase in risk (0.13% (-0.10%, 0.36%).

Of the three causes of death COPD shows the strongest percentage relative risk of all three cause of death (Chapter 7.4.2). With significant decreases in percentage relative risk for lags 1-6, 7-12, and 19-24 that are almost double pneumonia at -1.01% (-1.43%, -0.61%), -0.71% (-1.15%, -0.30%), -0.56% (-1.03%, -0.12%) respectively. At 0.20% (-0.15%, 0.54%) a non-significant increase for warm temperatures above 15°C was also observed at lag 1-6 for COPD mortality.

The number of papers reporting the effects of temperature on pneumonia and COPD over longer lag periods of 30 days is limited. A paper in 2002 by Braga, Zanobetti, and Schwartz modelled the effect of temperature on pneumonia and COPD over a 20 day lag period.⁴⁰⁷ They split into cold and hot cities, with the cold cities better matching Scotland's temperature range. No effect estimates were produced for the within 20 day lag that could be compared, however the surface plot of the distributed lag did indicate small but a similar pattern to those reported in this study. Cold temperature cities showed an increased risk that lasted the majority of the lag before returning to baseline after 20 days. Though a general downward trend in risk was observed across the temperature exposure range, there appears, for warm temperatures (>20°C), to be an immediate increase risk that lasts the first two days at most. No immediate risk of COPD mortality was observed for the colder and hotter temperatures in the cold

temperature cities. In the cold temperatures a very slight gradual increase in risk occurred that appears to have lasted the entire 20 day lag period. In the hot temperatures there appears to be decrease in risk that lasts for the first four days at which risk appears to return to the baseline.

A similar study to the one presented in this thesis, investigated the maximum daily temperature on a 6 week lag effect of pneumonia, COPD (and IHD) in Ireland.⁴⁰⁸ Comparatively effect sizes are much larger as they relate to weekly mean of maximum daily temperature during the winter only. A strong increase in pneumonia risk occurred for cold temperature that was greatest after 3 to 4 weeks that weakened but persisted for pneumonia for the full lag period. In this paper based in Ireland the lagged effect lasted for 3 weeks before dissipating.⁴⁰⁸ As with the results reported for this study the COPD effects were larger than those for pneumonia.

If compared with the more commonly reported results relating to general respiratory outcomes including asthma, influenza, and bronchitis among others. The most recent work using similar techniques to those reported here have been produced for cities with remarkably different climates than Scotland, places such as China, Vietnam and the Philippines. Even though they tend to have temperature ranges between 5°C and 40°C they might still provide a useful contrast. Similar results were found in Hong Kong,⁴⁰⁹ where extreme heat produced an immediate effect that dissipated quickly, and cold temperatures were more gradual and lasted longer into the lag period. In Hue, Viet Nam and Suchou, China, extreme heat did indicate a similar result but no cold temperature effect was observed for the entire lag period.^{410,411}

It should be noted, there appeared to be no evidence of a harvesting effect in both the hot and cold temperatures within this study. The effects tended to increase and return to baseline without the clear rebound period of reduced risk thought to characterise the move forward of susceptible individuals. In the three papers based in Asia, extreme heat did indicate a potential harvesting effect for respiratory mortality but only in those temperatures above 30°C. Even the effects associated with relatively high temperatures between 25°C and 30°C do not appear to indicate a harvesting effect; it may be possible that even as temperature increases risk as it tends towards the extremes, harvesting does

not occur in respiratory mortality until extreme temperatures are reached e.g. above 30°C. Even so, this may indicate that high temperatures predominantly afflict those already with the disease, particularly those with increased severity, hence the immediate but short risk length. In contrast, cold temperatures may also cause the development of the disease either through aiding commencement, or exacerbation of milder cases. Though not enough information was available, it would be interesting to stratify these analyses by a disease severity score such as the GOLD rating in COPD.¹²⁶

Ischaemic Heart Disease

The results presented appear to confirm the link between cold temperatures and ischaemic heart disease (Chapter 7.4.3) thought to contribute to almost half of all cold related mortality.⁴¹² Though a decrease in percentage relative risk occurred for the full 30 day lag period the effect magnitude in cold temperatures dropped considerably after 6 days with a lag 1-6 percentage relative risk of -0.71% (-0.89%,-0.54%) dropping to -0.13% (-0.34%, -0.03%) at lag 7-12 which held consistent for the remaining 30 days. A similar pattern occurred within the mild temperatures but at a smaller magnitude. Interestingly, for warm temperatures no increase effect was observed in the shortest lag 1-6 with a non-significant decrease in risk continuing throughout the exposure range -0.29% (-0.59%, 0.02%).

In both cases, colder and warmer temperatures, it may be the change in temperature from the average experienced rather than an absolute temperature value that makes the difference.^{404,412} The ability to acclimatise along with home heating/cooling technology means humans comfortably live in areas with high or cold average ambient temperatures. It is sudden changes such as heat waves in areas with generally low temperatures or cold snaps in warmer climates that cause the most impact.^{99,404} More recently high temperatures and their influence on mortality have been focused on.^{99,413} Studies that modelled longer lags and reported immediate mortality risk for cardiovascular disease^{414,415} and ischaemic heart disease^{416,417} have tended to be located in countries with more extreme temperatures (>20°C) that occur on a more regular

basis. For example, in the much warmer climate Guo Y et al. (2012) fitted the distributed lag non-linear model to mortality from ischaemic heart disease in six Chinese cities and indicated a strong immediate risk (0-2 days) that dissipated quickly.⁴¹⁶ Smoothed plots of the non-linear change in IHD risk across 27 days were reported for Beijing and Shanghai.⁴¹⁸ A strong increase risk was observed for both extremes of temperature that lasted just a few days. Given the differences in temperature range (Scotland -16 to 25°C, China -5 to 35°C) and fewer days above 20°C in Scotland, the increased risk of IHD mortality may either not be observable or not relative to the change from average temperature experienced but instead an absolute temperature point. In a more similar climate (England and Wales) a significant lag 0-1 increase in risk of all cardiovascular mortality types including IHD was observed for a 1°C increase in maximum temperature above the 93rd percentile at approximately 20-24.7°C.¹⁰³ In Canada, the maximum temperature increase was observed to increase the risk of congestive heart disease mortality lag 0-2 though only once it was over 25°C.

Inspection of the more flexible cubic distributed lag function did indicate a slight increase in risk occurring at lag 1 when temperature increased above 17°C. It may be that lag periods of 6 days were too wide to observe a single day effect not strong enough to dominate. Investigating the same day effect (lag 0) was not suitable for the main temperature analysis as we were not including same day pollution in the analysis and in order to make them comparable they were both modelled under the same conditions. However, a small sensitivity analysis was performed that included lag 0 in the model, which displayed a borderline significant increase in IHD risk of 1.18% (-0.01%, 2.38%) per 1°C increase above 17°C.

Even though hourly temperature measurements were available for Scotland, the exact time of death was not available in the mortality data hence lag modelled in terms of hours since exposure was not possible. Bhaskaran et al. (2012) calculated the odds of hospital admission attendance due to myocardial infarction in the 360 hours since an increase in 1°C above 20°C. Over the first 24hrs (split into 6 hour intervals) the first 6 hours showed an increase in attendance at hospital.⁴¹⁹

Summary

The strongest and consistent effect sizes were seen for temperature, particularly for chronic obstructive pulmonary disease. Cold temperatures had a strong effect on all three causes of death with the two respiratory diseases reporting strong effects that lasted long into the lag period, whereas the effect on ischaemic heart disease was much shorter at less than a week. Given the more temperate climate, brief increased risk was observed for the two respiratory diseases but not for ischaemic heart disease. It may be that effects on ischaemic heart disease last for hours only and so the study was not equipped to see them.

9.4.2 Pollution

The strongest pollutant effects tended to be observed in PM_{2.5} regardless of outcome, and of the three causes of death the strongest effects were observed in COPD regardless of pollutants. COPD mortality risk was observed to be a gradual increase that lasted the full 30 day lag period, whereas pneumonia tended to have an immediate effect that gradually declined after 15 days. In ischaemic heart disease a delay occurred before increased risk of mortality peaked either after 10 days or much later after 18-24 days depending on the pollutant.

The change in risk over an exposure-response and the lagged-response was flexibly modelled in a distributed lag non-linear model. Both the particulate pollutants (black smoke, PM₁₀, and PM_{2.5}) and the gaseous pollutants (sulphur dioxide and nitrogen dioxide) indicated no evidence of non-linearity in exposure-response present for any of the 30 day lag period in all three causes of death. Indicating there was no evidence to reject the assumption that linearity in response across pollution exposure was persistent for the full lag period, an assumption that had previously been based on analysis of the same day or previous day's exposure.²⁶²⁻²⁶⁴ Change in risk over the lag period was generally in the form of a simple quadratic or cubic curve depending on the pollutant and cause of death. The linear-exposure response and the simple change in risk across the lag meant applying a lag stratified model and a distributed lag model to the data were suitable. This meant simpler effect sizes could be obtained, interpreted, and compared in multiple comparison groups and analyses.

The following results all correspond to the multiple imputation analysis in community based subjects only (Results Chapter 8.4) and any effect estimate (percentage relative risk) relate to a 10µgm⁻³ increase on any individual day within the lag period.

9.4.2.1 Pneumonia

Of the two respiratory diseases the lagged risk to Pneumonia was immediate and tended to last a shorter length of time (Results Chapter 8.4.1). Under BS and PM_{2.5} exposure the risk lasted for 18 and 24 days respectively whereas with PM₁₀ a shorter six day initial risk was followed by a second period peaking 20 days later. The strongest percentage relative risk increase (PM_{2.5}) was observed at 0.48%, 0.99%, and 1.60% for lags 1-6, 7-12, and 12-24. Smaller effects were seen for BS (0.10% lag 7-12) and PM₁₀ (0.11% lag 1-6). The size is likely to be due to the much smaller sample sizes as indicated by the still wide confidence intervals. Of the gaseous pollutants only sulphur dioxide indicated percentage relative risks associated with pneumonia that were similar to BS but weaker. In five pollutants there appeared to be little to no evidence of a rebound effect within the 30 day lag period. If harvesting or mortality displacement is occurring then it appears to only be after three weeks and so not seen in this 30 days.

The use of community based only subjects has meant an increased likelihood results relate to Community Acquired Pneumonia (CAP) one of the leading causes of infectious disease in the world.⁴²⁰ Pneumonia, and CAP in particular, has a number of underlying biological mechanisms that may lead to bimodal peaks in risk seen in the particulate matter metrics. Infection may be due to a number of bacterial, viral, or chemical causes each with differing underlying mechanisms and induction periods. Different common micro-organisms may be present such as *Staphylococcus aureus*, Gram-negative enterobacteria, *Streptococcus pneumoniae*.¹¹⁹ Differing underlying pathogens may produce differing risk patterns as for example *Streptococcus pneumoniae* has an incubation period of 1-3 days, whereas *Haemophilus influenzae* and *Mycoplasma pneumoniae* have incubation periods of 2-4, and 6-32 days respectively.¹²⁰ The particulates themselves may cause increased risk of pneumonia by impairing microbial clearance via the mucociliary mechanism,⁴²¹ hindering macrophage phagocytosis,⁴²² or causing intense capillary engorgement and loss of epithelium.⁴²³ The temporal relationship between air pollution and pneumonia mortality analysed here may comprise a period of incubation, chemical insult, and diagnosis before leading to death. In order for pneumonia to occur the bacteria, virus, or chemical particles must reach the alveoli region and induce a reaction from the immune system. If the pathogen is

particularly dominant an intense inflammatory immune response can occur causing additional fluid (plasma and mucus) that interferes with the natural substances (surfactant) designed to aid breathing, resulting in increased breathing rate, coughing, and fever among other symptoms.⁴²⁴ The majority (99%) will recover with an improvement in symptoms after 48 hours, though breathing may be difficult for at least a month. Those at greater risk of pneumonia tend to be the elderly, have a chronic or immunosuppressive condition already present, or lifestyle factors.⁴²⁵ All of these factors may interfere at different rates between the induction, symptoms, and outcome for the differing bacterial, viral, and chemical causes of pneumonia; presenting in the bimodal peak risk observed for pneumonia mortality.

Few studies have reported results specifically for pneumonia. One, that was originally investigating harvesting also reported results for pneumonia (along with COPD and IHD) mortality risk, in which a smoothing technique for PM_{2.5} exposure modelled results across lag lengths of 0, 15, 30, 45, and 60 days.²⁸⁶ The results indicated a similar pattern of risk may be occurring to that which was seen in the PM_{2.5} results i.e. a bimodal risk period. Others have focused on shorter lags (<6 days) for mortality or more commonly hospitalisation.⁴²⁶⁻⁴²⁸ Braga et al (2000) considered lag up to 6 days for PM₁₀ exposure on pneumonia mortality, reporting an immediate increase risk at lag 1 that dissipated quickly by day five.¹⁵⁰ Janssen et al. (2013) reported risk of mortality due to pneumonia associated with 10µgm⁻³ on individual days and cumulative lag 0-6 days.⁵³ The excess risk for lag 0-6 was reported as slightly larger than the most comparable lag 1-6 results reported in this study. Whereas PM₁₀ and PM_{2.5} were reported here to be 0.11% and 0.48%, respectively, Janssen et al reported 5.3% and 4.3% for a cumulative average increase of 10µgm⁻³ over 7 days which would be approximately 0.75% and 0.62% for a single day increase. Looking at the individual day excess risk, both PM₁₀ and PM_{2.5} risk tended towards zero as the lag day tended towards zero, indicating unlike the PM_{2.5} results reported here no immediate increase in risk was observed at lag 0 or 1 rather a gradual increase over the first few days. A six day average lag was looked at by Fischer et al. (2002) who focused on the risk in four age groups (<45,45-64,65-74,>75) in the Netherland.²¹² A percentage relative risk increase between 0.24% and 1.14% (PM₁₀) and -0.20% to 0.91% (BS) was observed for the age groups. Two papers that were then combined in a meta-analysis reported

results for mortality from pneumonia & influenza over lag 0-1.^{43,429} The combined relative risk indicated a statistically significant increase of 0.49% for PM10 increase.

With respect to gaseous pollutants only three papers, Hoek et al (2000), Fischer et al. (2002), and Guaita et al. (2011) appear to have reported SO₂ and NO₂ lag 0-6 day effects for the Netherlands^{212,430} and Madrid³⁹⁴ respectively. The Madrid paper with respect to pneumonia appears to have reported only the lag day (between 0 and 6) that reported a statistically significant relative risk increase for the pollutant and so confidence in any comparison is difficult. A direct comparison with Hoek et al. was also difficult as relative risks for the cumulative lag 0-6 related to a change in the pollution between the 1st and 99th percentile. The Fischer paper however indicated for the four age groups a percentage relative risk of 0.39% (SO₂ >75 only) and 0.36% to 1.89% (NO₂) for the three oldest age groups. No evidence was found to suggest an increase in NO₂ increased risk of Pneumonia mortality in this study, and the effect size associated with SO₂ was over three times that of the one reported here.

The results reported in the handful of studies looking into the pneumonia have, when comparable, tended to be slightly larger. This may be for a number of reasons these studies, unlike here, have tended to be primarily investigating other factors or were done at a time where less sophisticated methods were available reporting results for shorter one day lags. Pneumonia is a leading cause of death, with in a more susceptible population that is growing due to an ageing population and the increase anti-biotic resistant bacteria.^{111,431} Cases of pneumonia presenting at hospitals in the UK have increased by 34% between 1997 and 2004 with the cost to the NHS likely to be considerably more than the £441 million per year estimated in 1993.^{432,433} The results reported here can only be a useful contribution that could help predict and manage increased risk periods of pneumonia during both pollution and extreme temperature episodes.

9.4.2.2 Chronic Obstructive Heart Disease (COPD)

Of the three causes of death investigated here, chronic obstructive pulmonary disease appeared to be influenced the most by both particulate and gaseous pollutants (Results Chapter 8.4.2). The largest percentage increase in relative risk was observed relating to COPD in all pollutants of which PM_{2.5} indicated the largest %RR at 0.61% (-0.93%,2.30%), 1.29% (-0.18%,2.89%), 0.73% (-2.41%,4.57%) increase was observed at lag 1-6, 19-24, and 25-30 days respectively. The risk pattern for COPD tended to be an elongated quadratic curve across the entire 30 day lag period. Only PM_{2.5} (and SO₂ at a much weaker effect) indicated an immediate increased risk, whereas the others showed a gradual increase over the first 12 days before peaking between 18-20 days BS lag 13-18 days = 0.28% (-0.05%, 0.61%) and PM₁₀ lag 19-24 days 0.48% (0.02%,0.95%). The gaseous pollutant SO₂ suggested a similar pattern but with reduced effect sizes to PM_{2.5}. A lag 1-6 %RR of 0.11%, (-0.08%, 0.31%) and second increase in risk at lag 13-18 0.15% (-0.04%, 0.34%). Nitrogen dioxide indicated a small increase risk but not until the end of the lag 19-24 and 25-30 0.09% (-0.42%,0.62%) and 0.09% (-0.14%,0.32%).

Long-term exposure or repeated short-term exacerbations may be causal factors in acquiring COPD, a predominantly chronic non-communicable disease.^{140,141} Unlike pneumonia, an individual short-term exposure to air pollution is unlikely in COPD to cause death in a previously disease free individual but rather exacerbate stress on an already present condition that may be at one of several stages in its progression. Though the underlying mechanisms are unclear,¹³⁸ exacerbations in COPD sufferers triggered by air pollution have been shown through increased symptoms, worsening lung function, and bronchial biopsies.⁴³⁴ Using diary cards and home breathing monitors, in several small studies black smoke, PM₁₀, SO₂, and NO₂ have after one day been observed to negatively affect peak respiratory flow, FVC, FEV, and increased symptoms such as shortness of breath.⁴³⁵⁻⁴³⁸ Recently a larger study also identified an association between low concentrations of PM_{2.5} and NO₂ on a reduced FEV and FEV that lasted for seven days and was significant for the first day (NO₂) or two days (PM_{2.5}).⁴³⁹

Exacerbations lead to the plausible conclusion that already weakened defence mechanisms against viruses, bacteria or other antioxidant defences may be further impaired due to the increases in either particulate or gaseous pollutants.^{440,441} It is thought that air pollution in both particulate and gaseous forms impedes the immune system allowing bacterial and viral agents to cause an exacerbation.¹⁴¹ This may be true in the pollutants observed in this study with a delayed effect i.e. black smoke, PM10, and NO2. However, PM2.5 and SO2 had an immediate effect indicating the pollutant itself maybe causing the irritation. Both SO2 and PM2.5 may cause an immediate irritation to the respiratory mucosa in the upper respiratory tract, or cause inflammation in the lung wall much like the potential mechanism discussed for pneumonia.⁴⁴² Particulates and NO2 are oxidants that generate free radicals causing cellular oxidative stress or stimulating the lung wall which may cause constriction of the smaller airways.⁴³⁹ The complication however, is that even those with stable COPD have a persistent bacterial presence in the lower airways and an impediment to the immune system may allow a more rapid contamination.⁴⁴⁰ The pollutants with a delayed effect or the secondary peak in others (PM2.5 and SO2) may have a differing mechanism. It has been shown that 78% patients with exacerbations entering hospital have viral or bacterial present that results in increased damage to the lung and longer hospital stays.⁴⁴³ The stronger effects observed towards the end of the lag period may be related to this. As discussed with pneumonia many types of bacterial and viral pathogens may be present each with differing delays between colonisation and symptoms dependent on the host susceptibility levels.

The delay of four or five days before a gradual increase in risk that extends over the remaining lag period may be indicative of a pollution influence on a population suffering from a chronic condition with different stages of frailty. Participants already suffering from the condition may be monitored more closely resulting in early symptoms spotted and swift access to healthcare. Those at high risk may already be located in hospital. One study attempted to investigate the influence of air pollution on those already suffering from the condition. Sunyer & Basagana (2001), modelled mortality risk in a bi-directional case-crossover design for those previously attending hospital in Barcelona with COPD, some 10 years earlier. Only PM10 indicated a

significant increase (OR=1.11 (1.00,1.24) per lag 0-2 day cumulative IQR increase of $27\mu\text{gm}^{-3}$, which held when secondary gaseous pollutants (NO₂, O₃, and CO) were also included.⁴⁴⁴

Studies have consistently linked short-term increases air pollution to increase risk of COPD, predominantly with respect to hospitalisation,^{213,428,445} but also to increased risk of mortality, particularly with respect to particulate matter.^{80,150} As stated in Chapter 9.4.1, only Schwartz appears to have reported a lagged effect greater than 6 days with respect to COPD.²⁸⁶ Schwartz indicated displacement of frail COPD subjects was by weeks or months rather than a few days as was observed for pneumonia. This was due to the observed shape of increased COPD mortality risk being similar to that reported in this study i.e. a gradual increase that peaks at approximate 15-20 days before a similarly gradual decrease. Schwartz indicated that this would continue into the second month post exposure. The negative percentage COPD after 45 days was thought to be a consequence of the frail COPD subjects dying almost a month earlier. Janssen et al. (2013) also reported increases (significantly) in risk of mortality due to COPD.⁵³ Janssen's excess risk for lag 0-6 was comparable to the results presented here for PM_{2.5} as the 6.1% for a cumulative average increase of $10\mu\text{gm}^{-3}$ or approximately 0.87% for a single day increase. Individual day excess risk was also observed to be slightly stronger for PM_{2.5} (2.5%, 3.5%, 2.1% lag 0, 1, and 2 days respectively) than PM₁₀ (1.5%, 2.5%, and 2.0%), and were unaffected by the inclusion of a second pollutant. Samoli et al (2014) and the MED-PARTICLES project reported the effects of PM_{2.5} and PM₁₀ for varying lag structures (0-1, 2-5, and 0-5 days) in southern European metropolises.⁴⁴⁶ Even with the difference in climates, an immediate effect (lag 0-1) was observed for PM_{2.5} but not PM₁₀. When secondary gaseous pollutants (SO₂ and NO₂) were included the PM_{2.5} lag 0-5 saw a small drop from 2.53% on its own to 1.95% and 1.67% respectively for a cumulative average increase in PM_{2.5}.

Chronic obstructive pulmonary disease is a leading cause of death with a significant economic cost. COPD has been estimated to cost over £800 million in health care per year.⁴⁴⁷ As with pneumonia the numbers of COPD cases have been increasing with its current status as the fourth leading cause of death is likely to become the third within the next 10 years.⁴⁴⁸ Compared to the other causes of death, the sample size relating to

COPD is comparatively small which may manifest in the larger effect sizes and wider confidence intervals. However, when comparable the results do tend to agree with the previous literature increasing confidence in these results. Though more work is required to identify those within COPD being effected i.e. is it simply the most severe COPD sufferers? These results are some of the first to gauge the length and size of the risk of acute COPD mortality across longer lag periods that will hopefully improve management and prevention in periods of increased pollution.

9.4.2.3 Ischaemic Heart Disease (IHD)

Of the three causes of death, the effect estimates were smallest for ischaemic heart diseases with each pollutant indicating a similar pattern of risk over the 30 days (Results Chapter 8.4.3). In each case a period of negative risk was observed before increased risk of IHD mortality presents near the middle of the lag period. The strongest increase %RR was observed in the two gaseous pollutants SO₂ at lag 13-18 (0.11% (0.02%,0.21%)) and NO₂ at lag 19-24 (0.23% (-0.12%, 0.12%)), and the particulate matter PM_{2.5} at lag 25-30 (0.35% (-0.49%,1.24%)).

Ischaemia, or ischaemic body tissue, occurs when delivery of blood to the tissue is inadequate causing restricted oxygen supply and an increase in toxins that would normally be removed.⁴⁴⁹ Lack of oxygen in the myocardial (cardiac muscle) cells leads to angina and myocardial infarction (heart attack). Though the causal mechanism between exposure and acute IHD has not been confirmed there appears to be two main pathways where air pollution either influences the autonomic nervous system or stimulates the inflammatory and coagulation responses in the blood.

Immediate contact of particulate and gaseous air pollution is thought to cause an immediate ischaemic heart disease response by impairing the *autonomic nervous system* altering the hemodynamic and haemostatic functions that regulate the heart and blood pressure.^{95,450,451} In particular, the air pollution PM_{2.5} has been linked to a decreased heart rate variability which is in turn linked to increase risk of cardiovascular events such as arrhythmias.^{94,452} As heart rate variability decreases due to exposure, it has also been shown that blood pressure can rise possibly due to a combination of low heart rate

and increased vascular constriction.⁴⁵³ Symptoms such as irregular heart rate, irregular cardiac muscle function, or increases in plasma viscosity can have an immediate effect with final outcome occurring within hours. An immediate effect was not observed here with a negative relative risk reported for both the particulates and gaseous pollutants before a small increase in risk.

In addition to influencing the autonomic nervous system alternative biological mechanisms have been proposed that may be more relevant here.^{93,454,455} Inhalation may directly or indirectly through damage sustained to the lung wall release pro-oxidative or pro-inflammatory material into the circulatory system, causing *inflammatory and increased blood coagulation*. As discussed in Pneumonia and COPD, contact between particulate and gaseous pollutants can cause an inflammatory response through oxidative stress in the epithelium cells, resulting in an increase in a number of inflammatory markers.⁹⁶ Inhaled particles may also enter directly into the blood stream and have a direct toxic effect. Elements of inhaled pollution are thought to be pro-oxidative and once in the blood stream interact with the vessel wall, or the blood itself encouraging coagulation and thrombosis. Seaton et al. (1999) reported an association between inhaled PM10 and a reduction in the haemoglobin protein found in red blood cells.⁴⁵⁶ The inhalation of ultra-fine particles increased the adhesion properties the blood vessel walls and the red blood cells making them more likely to combine and increase the likelihood of atherosclerosis i.e. fatty material located on the inner walls. Others have indicated an increase, after exposure to particulates and SO₂, in fibrinogen and C-reactive protein. Fibrinogen is a protein involved in regulating the viscosity and likelihood of clotting in the blood,^{457,458} whereas C-reactive protein is produced when inflammation is present and has been associated with developing atherosclerosis and at the same time plaque instability.⁴⁵⁹

Though some overlap between the mechanisms may be present.⁴⁵⁴ If the breakdown of the lung wall is required to allow the non ultra-fine particulates to gain entry to the system a greater delay in effect between exposure and outcome will occur compared to when autonomic nervous system is impaired. This however does not explain initial decrease in risk observed in the first six days. It may be that 'harvesting' is playing a

role, where a rebound effect is occurring in lags 1-6 due an increased risk at lag 0 that cannot be observed because lag 0 was not modelled in this study.

The results presented here are reported for lag 1 (exposure on the day prior to death) until lag 30. Same day (lag 0) was removed from the analysis for two reasons. Firstly, to reduce the number of subjects being assigned an exposure occurring after death, for example if the subject died at 9am but the same day exposure measurement was determined by the following 15 hours. Secondly, the exact start and finish measurement times each day were unknown making them potentially unsuitable i.e. 9am to 9am daily average rather than 12am to 12am. The implication being that the anything other than midnight to midnight would mean exposure measurements associated with the following or preceding day would be assigned. It was therefore felt lag 0 should be dropped. Part 1 of the sensitivity analysis (Table 8.8) reports same day effects for IHD. A non-significant increase in percentage relative risk was only observed for lag 0 once the random effects meta-analysis was removed indicating a strong monitor influence on the results. Assuming a persistent negative effect at lag 0, one explanation may be that the data were not sensitive enough to observe effects occurring within hours (<24hrs) of exposure on IHD. Peters et al. (2001) and Bhaskaran et al, (2011) looked at the influence of exposure to pollutants stratified over 24 and 72 hours, respectively.^{232,460} Peters et al. reported strong significant effects occurring in the first two hours, with increased risk persisting until 6 hours for PM_{2.5}, whereas Bashkaran et al, reported similar results for PM₁₀ and NO₂ (hrs 1-6), SO₂ (19-24 hrs) before a negative risk was observed for the for the remaining 72 hours. If those individuals most susceptible to pollution die 'harvested' in these first few hours since exposure who might have otherwise died a few days later then there would be a decrease in risk in the following days.

As with the two respiratory diseases, very few have looked at the lagged effect of air pollution on IHD for a period specifically greater than five days. The Schwartz et al. (2000) paper already discussed indicated for a 10 $\mu\text{g}\text{m}^{-3}$ increase of PM_{2.5} a consistent gradual increase in percentage IHD deaths over the following 60 days. No immediate drop off was observed that may represent an immediate harvesting effect, though the windows may have been too wide to observe the phenomenon. Neuberger reported for

Vienna the PM10, PM2.5 and NO2 effects on IHD mortality for lag 0-1, 0-7, and 0-14.¹³⁹ Adjusting for the differing lag lengths, in all three pollutants a small non-significant immediate (lag 0-1) percentage increase in mortality was observed (strongest in PM2.5 = 0.5%). In all but PM2.5 the pollutant increased slightly and then remained constant across the next 14 days. The broad lag length makes identifying fluctuations difficult, but a decrease in 0-7 day lag for PM2.5 may indicate a rebound effect similar to that observed here. Studies that have also reported lagged effects greater than 6 days have focused on general cardiovascular mortality. Zanobetti et al. (2003) and Goodman et al. (2004) both modelled particulate exposure over 40 days for cardiovascular mortality using polynomial distributed lag models. In both cases they reported an immediate increase in risk, followed by a decrease and then small fluctuations around the baseline level of risk. However, the rebound period associated with harvesting did not occur until 7-10 days after exposure, a contradiction to the results reported above. This was unlikely to be due to other cardiovascular diseases as IHD make up the majority of cardiovascular deaths.

9.4.2.4 Comparing the three causes of death

The lag stratified model in the case-crossover design allowed for comparison of the lagged exposure effect estimates produced by the three causes of death. Regardless of direction, the effect magnitude associated with COPD was greater than both pneumonia and IHD for all pollutants except sulphur dioxide. Though the effect estimates were greater in COPD the significant differences were still largely between the two respiratory diseases and IHD. The strongest effect difference between causes of death related to PM_{2.5}, with the 30 day lag, and lags 1-6, 13-18, and 19-24 all reporting greater percentage relative risk. Even so, only lag 1-6 produced a significantly greater effect size than IHD. PM₁₀ results also indicated significantly different effect between COPD and IHD (1-30 and 1-6 day lag), and Pneumonia than IHD for the 1-6 day lag. Though no formal testing has been done elsewhere, strong COPD effect estimates have consistently been reported in the literature for a number of pollutants,^{446,461} often being greater than comparative causes of death such as IHD, and general cardiovascular and cerebrovascular causes.¹³⁹ The increased risk may be representative of the difference between exposure effects on a chronic progressive condition and on exposure effects instigating an acute condition in a relatively healthier population. In subjects with COPD, fluctuations in particulates and gaseous pollutions adversely affect symptoms leading to greater frequency of episodes of COPD exacerbation.⁴³⁶ As COPD is a rarer cause of death (see Table 4.2) the smaller sample sizes means a reduced power to see the larger effect sizes. Even so, sufferers of chronic obstructive pulmonary disease, particularly long time sufferers, may be a more susceptible frail cohort who when exposed are at greater risk and so a greater effect size is produced.^{132,135}

9.5 Limitations - Additional causes of bias

A number of limitations have been touched upon already whilst discussing the strengths or the results of the study. However as a large part of this study has been focussed on attempting to reducing bias in the effect estimates, it should be pointed out that there are still factors present that are likely to cause bias.

9.5.1 Misclassification in the exposure assignment

The study covers a 30 year period between Jan 1980 and Dec 2011 for multiple cities. With just under one million deaths this represents the largest study of short-term exposure effects on mortality currently known. To maximise the sample size, exposure data were required from a number of different pollution monitors. Through circumstance an exposure measurement for each city on each day was not available from a single monitor active at during the time period, rather multiple monitors with differing concurrent and sequential activity periods were employed.

Assigning exposure measurements from a single fixed site pollution monitor to a large mobile population means they are subject to error. It has been suggest that this comes in three sources:⁴⁶²

- 1) Difference between personal exposure and population average exposure

The bias incurred here is considered to be small, as the error is considered to be Berkson (i.e. random) and the subject exposure level when averaged across the population will tend towards the ambient pollution level, assuming there are no indoor sources of pollution.⁴⁶²

- 2) Population average exposure difference compared to the ambient levels

Of the three this is considered to be the greatest source of bias and is a non-random or classical error, where the error in the exposure measurement is independent of the population exposure i.e. assigning a higher or lower pollution measurement than was

experienced.⁴⁶² The primary cause of this in air pollution relates to indoor and outdoor air pollution, where the subject-population located indoors experiencing a different level to the ambient level. One strength of this study was its attempts to account for this source of error through use of the hospital admission data to identify the subjects location. This has been discussed further in Chapter 9.3.1.

3) Difference between the measured estimate and true exposure level

This represents the error incurred through estimating the true exposure level via single or multiple fixed site monitors. Among others these measurements are affected by two main sources of error. The spatial variability in exposure levels across a city and variability between monitors measuring or estimating the exposure level.

Spatial variation in exposure

Spatial variation in air pollution across a city is to be expected. Studies that have investigated the spatial variation by comparing monitors across a city have reported mixed results for PM10, PM2.5 and various gaseous pollutants. In all cases the correlation between monitors has been high, however the variation in exposure measurements has been shown to either vary significantly, even at short distances,⁴⁶³⁻⁴⁶⁷ or be fairly consistent⁴⁶⁸ between 'background' monitors located at various positions across a city. Though, those that have been consistent have tended to use fewer monitors (<4) across the city.⁴⁶⁸

Even with the attempt (Chapters 5 and 7.4) to improve the pollution-mortality relationship by focusing on those based in the community only, it is still difficult to accurately determine a subjects level of exposure. Any exposure misclassification could be reduced if the number of measurement sites could be increased, making it easier to evaluate local variations in pollution levels. As shown in the results Chapter 8.5, increasing the number of monitors may cause further complications. Therefore, atmospheric measurement studies may become very important in epidemiological studies. Studies such as iSPEX use members of the public with phones to measure the particulate content in light in cities across Europe.^{469,470} As the coverage across cities

increases and technology improves incorporating gaseous pollutants also, the results will help improve estimates of spatial variation in pollution levels in for example land use regression modelling.⁴⁷¹ These techniques were not suitable here due to the time frame, with iSPEX and geographic information systems (GIS) modelling currently only available for a single time point or at most 2 weeks. Improvements in computing power and the GIS software mean these may be useful in once daily estimates of pollution across a city can be provided for several years. Even then, it will still be difficult to confirm the subject's location during the exposure period. A problem that will persist until a cheap, practical and unobtrusive method of personal monitoring is available or a reliable method of identifying the subject location during exposure is proposed. Mobile phone tracking has been proposed as one method,^{472,473} but as with the personal monitors generating a large enough sample size difficult especially with the additional ethical and privacy considerations of tracking a person for several days.

Bias in the monitor measurements

It was suggest that the error incurred here could also become close to random error if the average of multiple (unbiased) monitors located in a city is used,⁴⁶² though choice of monitors may strongly influence the results.³⁸⁵ Elsewhere, multiple concurrently active monitors have been averaged for all suitable sites across a city.^{48,78,79,366,474,475} As in this study this is not always possible and so a single monitor with the least missing data is used,^{48,203,219} or if multiple monitors the city is split into sectors and each subject is assigned the pollution estimate from the monitor in the same sector.^{13,272,273}

Employing exposure measurements from a single monitor and not averaged across multiple monitors, leaves susceptibility to between monitor differences in the estimate of the true pollution exposure.³⁸⁵ Even with quality control, variability between monitors is likely to be observed in the measurements taken which may differ randomly or systematically. Errors in the measurements may occur due to many factors and be unnoticed. Pollution monitors are sophisticated instruments that require careful calibration (which are often done in a lab rather than on site), maybe influenced by a number differing external factors, or may be limited by the methods available at the time. Over time the monitoring methods themselves have been replaced with new

techniques, for example the SO₂ manually operated bubblers were replaced with automated UV fluoresce monitors in the 1990s. All of which may contribute to between monitors biases, where two monitors measuring the same exposure give different estimates.

Modelling monitor variability

In this records based study, choosing a single monitor that covers the entire time period and is not affected by a high concentrated source (e.g. hotspot site) or averaging across multiple monitors all running simultaneously are the ideal scenario. This was not possible and so a number of different monitors located in different parts of the city and active at different time periods were employed. As has been shown in Chapter 8.5, monitoring sites produce systematically different effect estimates even between monitors active at the same time period for the same city that often contradicted each other in terms of effect direction. This appears to have had a large effect on the results produced here (Chapter 8.5) and may go some way to account for inconsistent or unexpected results. As the change in risk associated with a change in pollution (slope in the analysis model) is important and not the absolute pollution level (the intercept in the model), hence the differences between absolute level when estimating true exposure is not as important as the difference between within monitor measurement variation. For example, if the true exposure increases by $10\mu\text{gm}^{-3}$ but monitor A estimates an increase of $5\mu\text{gm}^{-3}$ but monitor B estimates an increase of $13\mu\text{gm}^{-3}$ then our estimate of the percentage relative risk could be compromised.

Adequately accounting for between monitor differences in a combined effect estimate is therefore important due to the several factors may compromise the result. The two-step multivariate random effects meta-analysis performed here was cumbersome particularly when also applying missing data techniques in multiple of monitors of varying sizes. The precision of the combined effect may be compromised by error within each monitor and between each monitor. This is especially true when the within monitor sample size is small or the total number of monitors is small.⁴⁷⁶ In the DerSimonian and Laird procedure employed here, the random effects weighting was applied equally regardless

of sample size meaning small-sample monitors were disproportionately amplified. It may be that a fixed effects meta-analysis would be more appropriate as this employs inverse variance weighting which would prioritise the bigger monitors.⁴⁷⁶ However, this would assume no between monitor differences in the effect estimates, a factor that has been shown to cause the fixed effects meta-analysis to produce confidence intervals with a poor coverage compared to the DerSimonian regardless of the number of monitors.⁴⁷⁷ Zeng & Lin (2015) also proposed a re-sampling method that may improve accuracy, this along with other more efficient methods are available,^{477,478} however they are computationally intensive and attempts here found convergence to be extremely unreliable. One method not applied, a Bayesian approach,⁴⁷⁹ may produce more reliable results though these may also be limited by computational intensity and a priori assumptions.⁴⁸⁰

With such small effect sizes and the apparent large variation observed in the within monitor effects, the results produced here, and similar pollution studies, are particularly susceptible. Effect estimates should therefore be interpreted with caution.

Monitor variability in the exposure variable construction

It has been noted that the variation in monitor measurements scale also impacts on model building in the two-stage meta-analysis of the distributed lag non-linear models. In this study, the knot positions across the exposure range for the distributed lag non-linear model were based on the monitor with the smallest exposure range and then applied to all monitors before being combined together in the multivariate meta-analysis. If one monitor has a small exposure range relative to the others forcing the knot positions to be in a fixed limited space then the flexibility is compromised. At the time of analysis this seemed the most appropriate method, since then a relative scale has been proposed.^{481,482} The relative scale positions the knot points at monitor specific percentiles meaning monitors with exposure ranges that do not overlap can be included.⁹⁷ This method means relative comparisons are made i.e. the relative change in risk compared to the normal expected temperature range. This seems logical in temperature where change from the expected temperature is a potential, though not confirmed,⁴⁸³ health concern⁴⁸⁴⁻⁴⁸⁶

In this study, even though the monitor exposure ranges differed, all monitors overlapped at the lower concentrations meaning non-overlapping ranges was not a concern. Also, an effect estimate relating to a fixed change in pollution (i.e. $10\mu\text{gm}^{-3}$) was preferred rather than a percentage relative change, which could relate to differing relative change for each monitor. However, a similar investigation of pollution change from the expected pollution level would be an interesting piece of future work.

Temperature specific monitor limitation

To reduce within monitor clustering temperature measurements from a single monitor for the entire study period was used. The nearest suitable monitors to Glasgow and Inverness were located at Prestwick airport (30 miles) and Aviemore Park (24 miles), respectively. Simple comparisons indicated Prestwick has an average temperature 1°C greater than Glasgow and Aviemore 1°C less than Inverness. As with pollution monitors, the distance from the city will compromise temperature results if the relative change in temperature measured was different from the true relative change in temperature in the city. This may be true as a number of factors are present that may influence temperature differently between the two cities and the monitor sites. Prestwick is located at an airport on the west coast of Scotland that is south-west from Glasgow. The ambient temperature for coastal areas such as Prestwick and Inverness can experience a slower rate of change than inland areas is thought to be due to the large body of water maintaining a consistent temperature. Cities can also experience an urban heat island effect that would not be present at Prestwick or Aviemore park, which is south-east of the northern city of Inverness.⁴⁸⁷

9.5.2 Misclassification in the cause of death assignment

So far this study has focussed on reducing bias caused by misclassification in measurements of the true exposure level. Misclassification can equally be present in the outcome. Misclassification of the cause of death field may cause bias in the results particularly if a cause of death is over or under reported.⁴⁸⁸ Subject level records based mortality data are generally considered complete with respect to containing a death record making them ideal sources of information. However, they may lack completeness and are susceptible to misclassification.^{489,490}

Though unlikely to be an issue in this study, the level of completeness is dependent on the within country data collection procedures with data collected by higher economically developed countries being more complete and reliable.⁴⁹¹ Even so, misclassification of the cause of death can occur at several stages in the process. Final cause of death fields are determined by ICD coding procedures that rely on the information reported in the death certificate, which in turn relies on the reported information within the certificate being correct, clear, and complete. This will depend heavily on the doctor in terms of their enthusiasm, experience, administrative training, medical education, specialism, and access to the subject's medical history or autopsy results. This is more difficult in the elderly and those with multiple morbidities as perceived typical causes may be assumed rather than the true cause being identified.⁴⁹¹ It can also be difficult to determine the level of influence chronic diseases, particularly as the person ages, have had on the process leading to death. A chronic disease may have started many years prior making them a consideration for both starting the chain of events or a contributing factor.⁴⁹²

The ICD coding procedure provides guidance regarding choice of the cause of death and position in the chain; however the guidelines are subject to interpretation⁴⁹³ particularly as the number of cause of death codes increases.⁴⁹² Despite the data processing guidance and training given to coders, further errors can occur during ICD coding. This may be due to poor handwriting on the death certificate, information gaps, incorrect interpretation of guidelines or incorrect determination of the underlying cause of death, and coding entry errors.^{494,495} Even under controlled conditions where the

coders are given the same level of training and the same amount of information the level of agreement can be as little as 56%.⁴⁹⁵ Even if the information recorded and the coding procedure is followed a greater than expectable amount of variation in final identified cause of death may still occur. A study in the Netherlands asked four coders to re-code death certificates using ICD coding procedure.⁴⁹² Even though the overall agreement was high at 78%, (pneumonia (85.5%), COPD (85.2%), and IHD (81.6%)) this still would result in \approx 11,000 deaths on average a year (assuming 50,000 total deaths) in Scotland being ambiguously coded.

The cause of death has been commonly determined by the underlying cause of death i.e. the cause that began the chain of events leading to death. In epidemiological studies where the primary effects relate to short-term changes in exposure, the underlying cause becomes irrelevant when the induction period is longer than a few weeks or months e.g. most forms of cancer. When the delay is short between exposure and symptoms or outcome, such as in an already present condition (COPD) or a short induction period (pneumonia, and ischaemic heart disease) the immediate cause of death also becomes relevant. By ignoring causes located within the chain, misclassification of many subjects cause of death is likely to be present with some causal effects becoming underestimated.^{342,343} Ideally, the immediate cause of death would be easily identifiable and a comparison with underlying cause could then be made. This was not the case in this dataset and resulted in any cause of death fields i.e. primary and all secondary, being used. An attempt was made to acquire a more suitable dataset for England from ONS and the health and social care information centre (HSCIC) who have indicated that they can identify the 'immediate' cause of death separately. As of writing the application is still on going and so was not available for inclusion in this submission. If the data is acquired this may be an area that may become interesting not only in air pollution research but for any observational epidemiological study with acute time-dependent effect estimates. Even then, misclassification of the outcome will still be present and acknowledged in conclusions based on the results of cause-specific mortality studies such as those provided here.

9.5.3 Further confounding – Multiple pollutants

One confounding factor not accounted for in the analyses presented here was the influence of co-pollutants. In this study all models were fitted as a single pollutant model only, in order to reduce the complexity of the monitor clustering. Due to the length of the study the pollutant measurements were sourced from multiple monitors running for different time-periods. To include additional pollutants would have meant increasing the number of monitor clusters accounting for all combinations creating additional complexity with in some cases very small cluster sizes. Using single pollutants models to represent a complex atmosphere comprised of varying particulate and gaseous compositions and interactions is likely to be inadequate. As can be seen by the correlation coefficients reported in Table 4.4, pollutants tend to be correlated in some cases highly correlated meaning any significant single pollutant exposure effect may be due to a highly correlated co-pollutant. However, the inclusion of multi-pollutants in the model is likely to cause collinearity to be present. Previous studies have started with single pollutant model and then if statistically significant gradually added a second pollutant with some attempt to identify clear signs of collinearity.^{153,475,496} Several methods have been proposed regarding multiple pollutant models and collinearity.^{497,498} This has included basic and complex models such as the mean of alternate days or lag lengths,⁴⁹⁹ creating pooled estimates using random effects or Bayesian meta-analysis techniques,^{500,501} or principal component regression.⁵⁰² Results commonly reported that the particulate matter metric (PM10 and PM2.5) effects are robust to the inclusion of most gaseous pollutants (SO₂, CO, O₃) with exception of NO₂ which appears to inhibit the PM10 effect⁵⁰³⁻⁵⁰⁵

Others have attempted to go further and account for potential individual pollutant components that make up a more general pollutant classification, usually PM_{2.5}.^{159,506} PM_{2.5} is thought to be made up of more than 40 chemical components including zinc, sulphate, selenium, nickel, bromine, sodium, nitrate, organic compound and elemental carbon. As these components must be measured hence the use of PM_{2.5}, a more recent addition to pollution monitoring. Sample sizes are still considered modest which is evident by the contrasting conclusions regarding important components; the elemental carbon and organic carbon, or bromine, nitrate, chromium etc.⁵⁰⁷⁻⁵¹⁰ Similar techniques

have been employed as in the multi-pollutant models with the recent addition of cluster analysis to group pollutant profiles, where clusters containing higher elements related to traffic pollution or oil burning caused a 3.7% increase in all-cause mortality.⁵¹¹

The atmospheric mixture of multi-pollutant and multi-components and their interaction with human population is complex in an urban setting and dependent on a number of factors including weather, emission sources, time of day, and lifestyle. Even if collinearity was not present the effects of multi-pollutant exposure may potentially be additive, synergistic, antagonistic, inhibitive, or may mask each other out, all of which could be occurring at once or at different time points.⁵¹² This makes it very difficult to determine the combined health effects, untangle the specific single pollutant effects, and account for any between pollutant interaction effects especially when the pollutants individually have been shown to be statistically significant. Even with these difficulties the multi-pollutant composition of the air means co-pollutants are an important confounder that needs further work. Particularly in terms of the lagged effects such as those reported in this study.

9.5.4 Multiple testing

Even though all comparison tests were performed within the context of the original aims of the study, it is recognised that multiple testing is a concern and its presence here should be clear to anyone interpreting the results. The multiple testing problem occurs when the number of performed hypothesis tests increases such that the likelihood of observing a significant effect through chance alone also increases. A significant result may then occur even if one is not truly present i.e. a false positive result is reported.⁵¹³ This particular problem is largely ignored in epidemiology⁵¹⁴ especially within the air pollution literature where many studies model a number of different scenarios relating to pollutants, outcome, lag length and sensitivity analyses and then report only the results that were a significant “increase” effect.^{43,394} This is misleading and makes comparison between analyses performed within the same study difficult, let alone between studies. This thesis, has reported the p-values associated with the comparison tests performed between causes of death and between hospital admission groups. The

significance level and patterns across the tests will allow the reader to determine if the comparison test should be considered a false positive or a true event. Due to the number of comparison tests performed, a ‘multiple testing’ correction such as a Bonferroni adjustment, was not applied due to concern it may cause an over correction creating a false negative to occur. The report however does display all results performed either in the main report or in the appendix. It can therefore be observed that the proportion all comparison tests with at least a borderline significant result (at the 5% level) was 8.5%. This combines both pollution and temperature results, for temperature only 14.8% were statistically significant compared to 5.4% of pollution tests. Indicating we can have some confidence in the significant temperature results are not false positives especially if the p-value is small. Whereas the 5.4% of significant pollution results does imply that they may be due to random chance alone rather than representing a true result. Given relatively small effect sizes and effect differences, even large pollution studies may struggle to gain the power required to observe a significant effect.

9.5.5 The exposure metric

The study results were reported in terms of a percentage relative risks relating to a single unit increase of $10\mu\text{gm}^{-3}$. This $10\mu\text{gm}^{-3}$ increase was repeated for all pollutants both particulates and gaseous in order to be able to combine effect estimates from different monitors, reduce complexity when reporting the differing pollutant effect estimates, and simplify direct comparisons both within the study and with external studies. However, differing pollutants and pollution monitors vary in the range of measured atmospheric exposure levels. The interquartile ranges (IQR) reported in Table 4.3, indicate that monitors measuring BS reported IQRs between $5\mu\text{gm}^{-3}$ to $14\mu\text{gm}^{-3}$, whereas PM2.5 monitors were lower at approximately $6\mu\text{gm}^{-3}$, and SO2 ranged from $3\mu\text{gm}^{-3}$ to $25\mu\text{gm}^{-3}$. If as here, an effect estimate represents an increase in $10\mu\text{gm}^{-3}$ but the measured range of exposure is less (e.g. interquartile range = 6) then the effect estimate will represent a larger effect than that occurring in the population i.e. the exposure or policy relevant effect will be lower. Hence the larger PM2.5 and smaller SO2 and NO2 effects when compared to BS and PM10 which have exposure measurements with interquartile ranges of approximately $10\mu\text{gm}^{-3}$. Though many studies have used the $10\mu\text{gm}^{-3}$ as the exposure metric others have used a study specific interquartile range,^{11,123,515} or in

some cases the 1st to 99th percentile points.⁴³⁰ The effect estimates can then also be thought of as representing the change in effect occurring between high and low exposure days. This may be an important consideration in environments where the exposure range may be more extreme. Here the variation in exposure was likely to be more consistent and so for the reasons already noted regarding model building and comparisons, it was thought acceptable to report the effect estimates with a consistent $10\mu\text{gm}^{-3}$ exposure unit throughout. Results across those pollutants that were adjusted for change in IQR (see examples provided in Chapter 8.2) were closer in magnitude to each other than when a $10\mu\text{gm}^{-3}$ increase was used. Though the effect estimates associated with PM_{2.5} were smaller for the population experienced range of $6\mu\text{gm}^{-3}$ they still tend to be larger than those produced in the remaining pollutants, confirming PM_{2.5} was still observed to have had a stronger effect in all three causes of death.

9.6 Potential future work

There are a number of possible areas for future research that have been identified by this work, however three stand out.

A large part of this study has been focused on reducing misclassification in both the exposure and the outcome data. Even though the analysis looking at hospital admission during exposure could be classed as an effect modifier in truth it is more concerned with reducing bias due to exposure misclassification. Little analysis has been performed here investigating factors influencing risk in specific subjects. The hospital admission data has a wealth of additional information that may be useful in identify those more susceptible individuals. Disease severity is both a confounding issue and potential modifier, particularly in COPD and pneumonia where a severity rating system such as the GOLD criteria or pneumonia severity index (PSI) may be useful.^{126,516} This was not thought to be available for this dataset however an alternative could be in the Charlson comorbidity index,⁵¹⁷ which is a scoring system used to assess susceptible patients in hospital that is based on the ICD coding. This along with other mortality data and hospital admission information such as co-causes of death, comorbidities during

previous attendances, number of hospital visits, length of previous hospital visits, age, and gender could be investigated.

Analysis of missing data in both pollution exposure studies and time-series data in general is in its early stages. The simulation study performed here was an early attempt at applying multiple imputation techniques in such datasets, and though the results were satisfactory for the main analysis presented here there are still questions that could be investigated. For example, even though season was included in the imputation model some bias was still present when the missing data was applied in seasonal ratios. The main effect parameter analysed in the simulations was same day (lag 0) effect. This meant ‘blocking’ a characteristic of missing data in pollution studies had little affect when randomly assigned. This is unlikely to be true when a lagged exposure effect parameter is included as is common in studies of this type and may be the reason why the effect estimates and confidence intervals didn’t improve quite as much as expected. These among other questions noted in the discussion would be interesting areas of future research in both environmental exposure studies and time-series data in general.

The misclassification in cause of death i.e. the underlying vs immediate cause of death issue is one that has been addressed here by using ‘any’ cause of death field in the mortality dataset. If the application to HSCIC and ONS is ever completed then this data may provide useful information that can help compare the two types of cause of death. This could be an important issue not just for environmental epidemiology but any study investigating mortality using records based information.

9.7 Conclusion

This observational epidemiological study has extensively investigated the delayed effect of particulate and gaseous pollutants on three specific causes of death in the form of pneumonia, chronic obstructive pulmonary disease, and ischaemic heart disease. Sophisticated analysis techniques suggested a non-linear temperature and linear pollution exposure-response relationship was present both immediately and after a delay up to 30 days. Bias due to exposure misclassification was reduced through use of the subject's hospital admissions data, and the study power was maximised by applying analysis of missing data techniques to reclaim lost information. The influences of exposure to temperature and pollution on the three cause of death were investigated over the following 30 days.

Of the three causes of death, the largest risk increase was related to COPD mortality especially under the influence of fine particulate matter. Risk associated with COPD appeared after a short delay and then lasted for the following 30 days, whereas pneumonia tended to indicate an immediate risk that dissipated within two to three weeks or in some cases a secondary peak risk would occur. Ischaemic heart disease reported the weakest effect sizes and contrary to expectations showed a negative immediate risk that gradually rose to a small peak at 15-20 days post exposure. This is could be a limitation of the data to see the same day or within hours effects of pollution thought to occur in ischaemic heart disease.

As is expected temperature effects were stronger than the pollution effects. A non-linear relationship was observed that tended to be U shaped in the immediate period but transformed at various points across the following 30 days. The Scottish climate means days with cold temperature outweigh those with a high temperature, hence cold temperature effects were larger and tended to last the full 30 days for the respiratory diseases and just under 2 weeks for ischaemic heart disease. Even with the limited high temperature days strong but non-significant immediate risks were observed for both respiratory diseases, and a small immediate risk for the ischaemic heart disease.

It is hoped that this information can help prevent and manage ill-health in the population due to increased pollution. Short-term exposure on 'all-causes' of death and general respiratory and cardiovascular disease have been extensively investigated in the literature, however few have investigated in such detail three very common causes of death, that with an ageing more susceptible population may be more common in the years to come.

Reference List

- (1) WHO. Burden of disease from Ambient Air Pollution for 2012. March 1, 2014. Available at: http://www.who.int/phe/health_topics/outdoorair/databases/AAP_BoD_results_March2014.pdf?ua=1. Accessed November 8, 2015.
- (2) Nemery B, Hoet PHM, Nemmar A. The Meuse Valley fog of 1930: an air pollution disaster. *Lancet*. 2001;357(9257):704-708.
- (3) Seinfeld JH. Air pollution: A half century of progress. *Aiche Journal*. 2004;50(6):1096-1108.
- (4) Logan WPD. Fog and Mortality. *Lancet*. 1949;256(JAN8):78.
- (5) Wilkins ET. Air Pollution and the London Fog of December, 1952. *Ama Archives of Industrial Hygiene and Occupational Medicine*. 1954;9(3):247-248.
- (6) Bell ML, Davis DL. Reassessment of the lethal London fog of 1952: Novel indicators of acute and chronic consequences of acute exposure to air pollution. *Environmental Health Perspectives*. 2001;109:389-394.
- (7) Schwartz J. Air-Pollution and Daily Mortality in Birmingham, Alabama. *American Journal of Epidemiology*. 1993;137(10):1136-1147.
- (8) Hastie T, Tibshirani R. *Generalized additive models*. 1st ed. ed. London : Chapman and Hall, 1990.
- (9) Jaakkola JJ. Case-crossover design in air pollution epidemiology. *Eur Respir J Suppl*. 2003;40:81s-85s.
- (10) Kelsall JE, Samet JM, Zeger SL, Xu J. Air pollution and mortality in Philadelphia, 1974-1988. *American Journal of Epidemiology*. 1997;146(9):750-762.
- (11) Forastiere F, Stafoggia M, Picciotto S et al. A case-crossover analysis of out-of-hospital coronary deaths and air pollution in Rome, Italy. *Am J Respir Crit Care Med*. 2005;172(12):1549-1555.
- (12) Schwartz J. The distributed lag between air pollution and daily deaths. *Epidemiology*. 2000;11(3):320-326.
- (13) Carder M, McNamee R, Beverland I et al. The lagged effect of cold temperature and wind chill on cardiorespiratory mortality in Scotland. *Occupational and Environmental Medicine*. 2005;62(10):702-710.
- (14) Legislation EP. Environmental Protection The Air Quality standards regulations 2010. June 11, 2010. Available at: http://www.legislation.gov.uk/ukxi/2010/1001/pdfs/ukxi_20101001_en.pdf. Accessed May 5, 2015.
- (15) EuropeanUnion. Directive 2008/50/EC of the European Parliament and of the Council on ambient air quality and cleaner air for Europe. 2008. Available at: <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008L0050>. Accessed June 5, 2015.
- (16) WHO. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulphur dioxide: Global Update 2005. 2006. Available at:

http://apps.who.int/iris/bitstream/10665/69477/1/WHO_SDE_PHE_OEH_06.02_eng.pdf. Accessed May 6, 2015.

- (17) COMEAP. Review of the UK Air Quality Index. June 22, 2011. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/304633/COMEAP_review_of_the_uk_air_quality_index.pdf. Accessed June 6, 2015.
- (18) Holgate ST, Grigg J, Aguis R et al. Every breath we take - The lifelong impact of air pollution. March, 2016. Available at: <https://www.rcplondon.ac.uk/projects/outputs/every-breath-we-take-lifelong-impact-air-pollution>. Accessed February 28, 2016.
- (19) Kampa M, Castanas E. Human health effects of air pollution. *Environmental Pollution*. 2008;151(2):362-367.
- (20) Semple S. Assessing occupational and environmental exposure. *Occupational Medicine-Oxford*. 2005;55(6):419-424.
- (21) Brauer M, Hoek G, van Vliet P et al. Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems. *Epidemiology*. 2003;14(2):228-239.
- (22) Marshall JD, Nethery E, Brauer M. Within-urban variability in ambient air pollution: Comparison of estimation methods. *Atmospheric Environment*. 2008;42(6):1359-1369.
- (23) Nieuwenhuijsen M, Paustenbach D, Duarte-Davidson R. New developments in exposure assessment: The impact on the practice of health risk assessment and epidemiological studies. *Environment International*. 2006;32(8):996-1009.
- (24) Harrison RM, Thornton CA, Lawrence RG, Mark D, Kinnersley RP, Ayres JG. Personal exposure monitoring of particulate matter nitrogen dioxide, and carbon monoxide, including susceptible groups. *Occupational and Environmental Medicine*. 2002;59(10):671-679.
- (25) Morabia A, Amstislavski PN, Mirer FE et al. Air Pollution and Activity During Transportation by Car, Subway, and Walking. *American Journal of Preventive Medicine*. 2009;37(1):72-77.
- (26) Branis M, Kolomaznikova J. Monitoring of long-term personal exposure to fine particulate matter (PM_{2.5}). *Air Quality Atmosphere and Health*. 2010;3(4):235-243.
- (27) Moller KL, Thygesen LC, Schipperijn J et al. Occupational Exposure to Ultrafine Particles among Airport Employees - Combining Personal Monitoring and Global Positioning System. *Plos One*. 2014;9(9).
- (28) Spix C, Anderson HR, Schwartz J et al. Short-term effects of air pollution on hospital admissions of respiratory diseases in Europe: A quantitative summary of APHEA study results. *Archives of Environmental Health*. 1998;53(1):54-64.
- (29) Villeneuve PJ, Johnson JYM, Pasichnyk D, Lowes J, Kirkland S, Rowe BH. Short-term effects of ambient air pollution on stroke: Who is most vulnerable? *Science of the Total Environment*. 2012;430:193-201.
- (30) Chiusolo M, Cadum E, Stafoggia M et al. Short-Term Effects of Nitrogen Dioxide on Mortality and Susceptibility Factors in 10 Italian Cities: The EpiAir Study. *Environmental Health Perspectives*. 2011;119(9):1233-1238.
- (31) Milojevic A, Wilkinson P, Armstrong B, Bhaskaran K, Smeeth L, Hajat S. Short-term effects of air pollution on a range of cardiovascular events in England and Wales: case-

- crossover analysis of the MINAP database, hospital admissions and mortality. *Heart*. 2014;100(14):1093-1098.
- (32) Wehner B, Birmili W, Gnauk T, Wiedensohler A. Particle number size distributions in a street canyon and their transformation into the urban-air background: measurements and a simple model study. *Atmospheric Environment*. 2002;36(13):2215-2223.
- (33) McAdam K, Steer P, Perrotta K. Using continuous sampling to examine the distribution of traffic related air pollution in proximity to a major road. *Atmospheric Environment*. 2011;45(12):2080-2086.
- (34) Fruin S, Urman R, Lurmann F et al. Spatial variation in particulate matter components over a large urban area. *Atmospheric Environment*. 2014;83:211-219.
- (35) DEFRA. Monitoring Networks: Brief history. February 14, 2011. Available at: <http://uk-air.defra.gov.uk/networks/brief-history>. Accessed August 16, 2015.
- (36) Heal MR, Hibbs LR, Agius RM, Beverland LJ. Total and water-soluble trace metal content of urban background PM10, PM2.5 and black smoke in Edinburgh, UK. *Atmospheric Environment*. 2005;39(8):1417-1430.
- (37) Andreae MO, Gelencser A. Black carbon or brown carbon? The nature of light-absorbing carbonaceous aerosols. *Atmospheric Chemistry and Physics*. 2006;6:3131-3148.
- (38) Holgate ST, Samet JM, Koren HS, Maynard RL. *Air Pollution and Health*. 1st ed. San Diego: Academic Press; 1999.
- (39) Samet JM, Dominici F, Curriero FC, Coursac I, Zeger SL. Fine particulate air pollution and mortality in 20 US Cities, 1987-1994. *New England Journal of Medicine*. 2000;343(24):1742-1749.
- (40) Schwartz J. Short-Term Fluctuations in Air-Pollution and Hospital Admissions of the Elderly for Respiratory-Disease. *Thorax*. 1995;50(5):531-538.
- (41) Sarnat JA, Schwartz J, Catalano PJ, Suh HH. Gaseous pollutants in particulate matter epidemiology: Confounders or surrogates? *Environmental Health Perspectives*. 2001;109(10):1053-1061.
- (42) Fairley D. Daily mortality and air pollution in Santa Clara County, California: 1989-1996. *Environmental Health Perspectives*. 1999;107(8):637-641.
- (43) Wong TW, Tam WS, Yu TS, Wong AHS. Associations between daily mortalities from respiratory and cardiovascular diseases and air pollution in Hong Kong, China. *Occupational and Environmental Medicine*. 2002;59(1):30-35.
- (44) Brimblecombe P. The big smoke
a history of air pollution in London since medieval times. 1987.
- (45) Hoek G, Brunekreef B, Goldbohm S, Fischer P, van den Brandt PA. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. *Lancet*. 2002;360(9341):1203-1209.
- (46) Quincey P. A relationship between Black Smoke Index and Black Carbon concentration. *Atmospheric Environment*. 2007;41(36):7964-7968.
- (47) Loader A. DETR Instruction Manual: UK Smoke and Sulphur Dioxide Network. April 10, 2002. Available at: http://uk-air.defra.gov.uk/library/reports?report_id=718. Accessed July 3, 2015.

- (48) Anderson HR, Bremner SA, Atkinson RW, Harrison RM, Walters S. Particulate matter and daily mortality and hospital admissions in the west midlands conurbation of the United Kingdom: associations with fine and coarse particles, black smoke and sulphate. *Occupational and Environmental Medicine*. 2001;58(8):504-510.
- (49) Mcfarland AR, Ortiz CA, Rodes CE. Wind-Tunnel Evaluation of the British Smoke Shade Sampler. *Atmospheric Environment*. 1982;16(2):325-328.
- (50) Bartzokas A, Kassomenos P, Petrakis M, Celessides C. The effect of meteorological and pollution parameters on the frequency of hospital admissions for cardiovascular and respiratory problems in Athens. *Indoor and Built Environment*. 2004;13(4):271-275.
- (51) WHO. Health Aspects of Air Pollution with Particle Matter, Ozone and Nitrogen Dioxide. January 15, 2003. Available at: www.euro.who.int/document/e79097.pdf. Accessed 2009.
- (52) Brunekreef B, Forsberg B. Epidemiological evidence of effects of coarse airborne particles on health. *European Respiratory Journal*. 2005;26(2):309-318.
- (53) Janssen NAH, Fischer P, Marra M, Ameling C, Cassee FR. Short-term effects of PM_{2.5}, PM₁₀ and PM_{2.5-10} on daily mortality in the Netherlands. *Science of the Total Environment*. 2013;463:20-26.
- (54) Zanobetti A, Schwartz J. The Effect of Fine and Coarse Particulate Air Pollution on Mortality: A National Analysis. *Environmental Health Perspectives*. 2009;117(6):898-903.
- (55) Belleudi V, Faustini A, Stafoggia M et al. Impact of Fine and Ultrafine Particles on Emergency Hospital Admissions for Cardiac and Respiratory Diseases. *Epidemiology*. 2010;21(3):414-423.
- (56) Maynard D, Coull BA, Gryparis A, Schwartz J. Mortality risk associated with short-term exposure to traffic particles and sulfates. *Environmental Health Perspectives*. 2007;115(5):751-755.
- (57) Stieb DM, Judek S, Burnett RT. Meta-analysis of time-series studies of air pollution and mortality: Effects of gases and particles and the influence of cause of death, age, and season. *Journal of the Air & Waste Management Association*. 2002;52(4):470-484.
- (58) Stolzel M, Peters A, Cyrys J et al. Particulate matter in several size classes and daily mortality in Erfurt, Germany. *Epidemiology*. 2004;15(4):S59.
- (59) Wichmann HE, Peters A. Epidemiological evidence of the effects of ultrafine particle exposure. *Philosophical Transactions of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences*. 2000;358(1775):2751-2768.
- (60) Wilson WE, Chow JC, Claiborn C, Wei FS, Engelbrecht J, Watson JG. Monitoring of particulate matter outdoors. *Chemosphere*. 2002;49(9):1009-1043.
- (61) Patashnick H, Rupprecht EG. Continuous Pm-10 Measurements Using the Tapered Element Oscillating Microbalance. *Journal of the Air & Waste Management Association*. 1991;41(8):1079-1083.
- (62) Ariola V, D'Alessandro A, Lucarelli F et al. Elemental characterization of PM₁₀, PM_{2.5} and PM₁ in the town of Genoa (Italy). *Chemosphere*. 2006;62(2):226-232.
- (63) Bigi A, Harrison RM. Analysis of the air pollution climate at a central urban background site. *Atmospheric Environment*. 2010;44(16):2004-2012.

- (64) Chen RJ, Huang W, Wong CM et al. Short-term exposure to sulfur dioxide and daily mortality in 17 Chinese cities: The China air pollution and health effects study (CAPES). *Environmental Research*. 2012;118:101-106.
- (65) Sunyer J, Castellsague J, Saez M, Tobias A, Anto JM. Air pollution and mortality in Barcelona. *Journal of Epidemiology and Community Health*. 1996;50:S76-S80.
- (66) Parrish DD, Holloway JS, Fehsenfeld FC. Routine, Continuous Measurement of Carbon-Monoxide with Parts-Per-Billion Precision. *Environmental Science & Technology*. 1994;28(9):1615-1618.
- (67) Navas MJ, Jimenez AM, Galan G. Air analysis: Determination of nitrogen compounds by chemiluminescence. *Atmospheric Environment*. 1997;31(21):3603-3608.
- (68) Campbell GW, Stedman JR, Stevenson K. A Survey of Nitrogen-Dioxide Concentrations in the United-Kingdom Using Diffusion Tubes, July-December 1991. *Atmospheric Environment*. 1994;28(3):477-486.
- (69) Heal MR, O'Donoghue MA, Cape JN. Overestimation of urban nitrogen dioxide by passive diffusion tubes: a comparative exposure and model study. *Atmospheric Environment*. 1999;33(4):513-524.
- (70) Kirby C, Fox M, Waterhouse J, Drye T. Influence of environmental parameters on the accuracy of nitrogen dioxide passive diffusion tubes for ambient measurement. *Journal of Environmental Monitoring*. 2001;3(1):150-158.
- (71) Fenger J. Urban air quality. *Atmospheric Environment*. 1999;33(29):4877-4900.
- (72) NCAS British Atmospheric Data Centre. British Atmospheric Data Centre (BADC). September 7, 2015. Available at: <http://badc.nerc.ac.uk/home/>.
- (73) NCAS British Atmospheric Data Centre. UK Meteorological Office. MIDAS Land Surface Stations data (1853-current). October 12, 2012. Available at: http://badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_dataent_ukmo-midas. Accessed June 1, 2013.
- (74) Postolache O, Girao PMBS, Pereira JMD, Ramos HG. Dew point and relative-humidity smart measuring system. *Ieee Transactions on Instrumentation and Measurement*. 2006;55(6):2259-2264.
- (75) Ustymczuk A, Giner SA. Relative humidity errors when measuring dry and wet bulb temperatures. *Biosystems Engineering*. 2011;110(2):106-111.
- (76) DEFRA. Site environment types. February 14, 2011. Available at: <http://uk-air.defra.gov.uk/networks/site-types>. Accessed July 6, 2015.
- (77) Namdeo A, Bell MC. Characteristics and health implications of fine and coarse particulates at roadside, urban background and rural sites in UK. *Environment International*. 2005;31(4):565-573.
- (78) Anderson HR, deLeon AP, Bland JM, Bower JS, Strachan DP. Air pollution and daily mortality in London: 1987-92. *British Medical Journal*. 1996;312(7032):665-669.
- (79) Goodman PG, Dockery DW, Clancy L. Cause-specific mortality and the extended effects of particulate pollution and temperature exposure. *Environmental Health Perspectives*. 2004;112(2):179-185.

- (80) Zeka A, Zanobetti A, Schwartz J. Short term effects of particulate matter on cause specific mortality: effects of lags and modification by city characteristics. *Occupational and Environmental Medicine*. 2005;62(10):718-725.
- (81) Office for National Statistics. Mortality Statistics: Deaths registered in 2008. 2009.
- (82) Byrne M. Aerosols exposed. *Chemistry in Britain*. 1998;34(8):23-26.
- (83) Schwartz J. What Are People Dying of on High Air-Pollution Days. *Environmental Research*. 1994;64(1):26-35.
- (84) Schwartz J. Is there harvesting in the association of airborne particles with daily deaths and hospital admissions? *Epidemiology*. 2001;12(1):55-61.
- (85) Janssen NAH, Schwartz J, Zanobetti A, Suh HH. Air conditioning and source-specific particles as modifiers of the effect of PM10 on hospital admissions for heart and lung disease. *Environmental Health Perspectives*. 2002;110(1):43-49.
- (86) Tellez-Rojo MM, Romieu I, Ruiz-Velasco S, Lezana MA, Hernandez-Avila MM. Daily respiratory mortality and PM10 pollution in Mexico City. *European Respiratory Journal*. 2001;18(6):1076.
- (87) Zeka A, Zanobetti A, Schwartz J. Individual-level modifiers of the effects of particulate matter on daily mortality. *American Journal of Epidemiology*. 2006;163(9):849-859.
- (88) Gouveia N, Fletcher T. Respiratory diseases in children and outdoor air pollution in Sao Paulo, Brazil: a time series analysis. *Occupational and Environmental Medicine*. 2000;57(7):477-483.
- (89) Anderson LM, Thurston G, Stolzel M. Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health. *Journal of Medicine and Toxicology*. 2012;8:166-175.
- (90) Jovic-Stosic J, Jovasevic-Stojanovic M. Potential Pathophysiological Mechanisms of Ultrafine Particle Toxic Effects in Humans. *Chemical Industry & Chemical Engineering Quarterly*. 2008;14(1):47-49.
- (91) Martinelli N, Olivieri O, Girelli D. Air particulate matter and cardiovascular disease: A narrative review. *European Journal of Internal Medicine*. 2013;24(4):295-302.
- (92) Brook RD, Rajagopalan S, Pope CA et al. Particulate Matter Air Pollution and Cardiovascular Disease An Update to the Scientific Statement From the American Heart Association. *Circulation*. 2010;121(21):2331-2378.
- (93) Brook RD. Cardiovascular effects of air pollution. *Clinical Science*. 2008;115(5-6):175-187.
- (94) Pieters N, Plusquin M, Cox B, Kicinski M, Vangronsveld J, Nawrot TS. An epidemiological appraisal of the association between heart rate variability and particulate air pollution: a meta-analysis. *Heart*. 2012;98(15):1127-1135.
- (95) Pope CA. Epidemiology of fine particulate air pollution and human health: Biologic mechanisms and who's at risk? *Environmental Health Perspectives*. 2000;108:713-723.
- (96) Brook RD, Franklin B, Cascio W et al. Air pollution and cardiovascular disease - A statement for healthcare professionals from the expert panel on population and prevention science of the American Heart Association. *Circulation*. 2004;109(21):2655-2671.

- (97) Gasparrini A, Guo YM, Hashizume M et al. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *Lancet*. 2015;386(9991):369-375.
- (98) Basu R, Ostro BD. A multicounty analysis identifying the populations vulnerable to mortality associated with high ambient temperature in California. *American Journal of Epidemiology*. 2008;168(6):632-637.
- (99) Basu R. High ambient temperature and mortality: a review of epidemiologic studies from 2001 to 2008. *Environmental Health*. 2009;8.
- (100) Bouchama A, Knochel JP. Medical progress - Heat stroke. *New England Journal of Medicine*. 2002;346(25):1978-1988.
- (101) Huynen MMTE, Martens P, Schram D, Weijenberg MP, Kunst AE. The impact of heat waves and cold spells on mortality rates in the Dutch population. *Environmental Health Perspectives*. 2001;109(5):463-470.
- (102) Cheng XS, Su H. Effects of climatic temperature stress on cardiovascular diseases. *European Journal of Internal Medicine*. 2010;21(3):164-167.
- (103) Gasparrini A, Armstrong B, Kovats S, Wilkinson P. The effect of high temperatures on cause-specific mortality in England and Wales. *Occupational and Environmental Medicine*. 2012;69(1):56-61.
- (104) Michelozzi P, Accetta G, De Sario M et al. High Temperature and Hospitalizations for Cardiovascular and Respiratory Causes in 12 European Cities. *American Journal of Respiratory and Critical Care Medicine*. 2009;179(5):383-389.
- (105) Mourtzoukou EG, Falagas ME. Exposure to cold and respiratory tract infections. *International Journal of Tuberculosis and Lung Disease*. 2007;11(9):938-943.
- (106) Sharovsky R, Cesar LAM, Ramires JAF. Temperature, air pollution, and mortality from myocardial infarction in Sao Paulo, Brazil. *Brazilian Journal of Medical and Biological Research*. 2004;37(11):1651-1657.
- (107) Atsumi A, Ueda K, Irie F et al. Relationship Between Cold Temperature and Cardiovascular Mortality, With Assessment of Effect Modification by Individual Characteristics - Ibaraki Prefectural Health Study -. *Circulation Journal*. 2013;77(7):1854-1861.
- (108) Goldberg MS, Burnett RT, Bailar JC et al. The association between daily mortality and ambient air particle pollution in Montreal, Quebec 2. Cause-specific mortality. *Environmental Research*. 2001;86(1):26-36.
- (109) Zanobetti A, Bind MAC, Schwartz J. Particulate air pollution and survival in a COPD cohort. *Environmental Health*. 2008;7.
- (110) Lozano R, Naghavi M, Foreman K et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2095-2128.
- (111) ONS. Deaths Registered in England and Wales in 2011, by Cause. November 6, 2012.
- (112) Blasi F, Mantero M, Santus P, Tarsia P. Understanding the burden of pneumococcal disease in adults. *Clinical Microbiology and Infection*. 2012;18:7-14.
- (113) File TM. Community-Acquired Pneumonia: Controversies and Questions Preface. *Infectious Disease Clinics of North America*. 2013;27(1):XIII-XXIV.

- (114) Cilloniz C, Ewig S, Polverino E et al. Microbial aetiology of community-acquired pneumonia and its relation to severity. *Thorax*. 2011;66(4):340-346.
- (115) Jennings LC, Anderson TP, Beynon KA et al. Incidence and characteristics of viral community-acquired pneumonia in adults. *Thorax*. 2008;63(1):42-48.
- (116) Almirall J, Bolibar I, Vidal J et al. Epidemiology of community-acquired pneumonia in adults: a population-based study. *European Respiratory Journal*. 2000;15(4):757-763.
- (117) Johansson N, Kalin M, Tiveljung-Lindell A, Giske CG, Hedlund J. Etiology of Community-Acquired Pneumonia: Increased Microbiological Yield with New Diagnostic Methods. *Clinical Infectious Diseases*. 2010;50(2):202-209.
- (118) Lim WS, Macfarlane JT, Boswell TCJ et al. Study of community acquired pneumonia aetiology (SCAPA) in adults admitted to hospital: implications for management guidelines. *Thorax*. 2001;56(4):296-301.
- (119) Wilkinson M, Woodhead M. Pneumonia. *Medicine*. 2004;32(2):129-134.
- (120) Hawker J, MyLibrary. *Communicable disease control handbook [electronic resource]*. 2nd ed. ed. Malden, Mass. ; Oxford : Blackwell Pub., 2005.
- (121) Rothman KJ, Greenland S. *Modern epidemiology*. 2nd ed. ed. Philadelphia, PA : Lippincott-Raven, 1998.
- (122) Schwartz J, Dockery DW. Increased Mortality in Philadelphia Associated with Daily Air-Pollution Concentrations. *American Review of Respiratory Disease*. 1992;145(3):600-604.
- (123) Halonen JI, Lanki T, Yli-Tuomi T, Tiittanen P, Kulmala M, Pekkanen J. Particulate Air Pollution and Acute Cardiorespiratory Hospital Admissions and Mortality Among the Elderly. *Epidemiology*. 2009;20(1):143-153.
- (124) Zanobetti A, Woodhead M. Air Pollution and Pneumonia The "Old Man" Has a New "Friend". *American Journal of Respiratory and Critical Care Medicine*. 2010;181(1):5-6.
- (125) Neupane B, Jerrett M, Burnett RT, Marrie T, Arain A, Loeb M. Long-Term Exposure to Ambient Air Pollution and Risk of Hospitalization with Community-acquired Pneumonia in Older Adults. *American Journal of Respiratory and Critical Care Medicine*. 2010;181(1):47-53.
- (126) Vestbo J, Hurd SS, Agusti AG et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease GOLD Executive Summary. *American Journal of Respiratory and Critical Care Medicine*. 2013;187(4):347-365.
- (127) Hurst JR, Vestbo J, Anzueto A et al. Susceptibility to Exacerbation in Chronic Obstructive Pulmonary Disease. *New England Journal of Medicine*. 2010;363(12):1128-1138.
- (128) Svanes C, Sunyer J, Plana E et al. Early life origins of chronic obstructive pulmonary disease. *Thorax*. 2010;65(1):14-20.
- (129) Pauwels RA, Rabe KF. Burden and clinical features of chronic obstructive pulmonary disease (COPD). *Lancet*. 2004;364(9434):613-620.
- (130) Smith KR. National burden of disease in India from indoor air pollution. *Proceedings of the National Academy of Sciences of the United States of America*. 2000;97(24):13286-13293.

- (131) Vestbo J, Lange P. Natural history of COPD: Focusing on change in FEV1. *Respirology*. 2016;21(1):34-43.
- (132) Sunyer J, Schwartz J, Tobias A, Macfarlane D, Garcia J, Anto JM. Patients with chronic obstructive pulmonary disease are at increased risk of death associated with urban particle air pollution: A case-crossover analysis. *American Journal of Epidemiology*. 2000;151(1):50-56.
- (133) Bateson TF, Schwartz J. Who is sensitive to the effects of particulate air pollution on mortality? A case-crossover analysis of effect modifiers. *Epidemiology*. 2004;15(2):143-149.
- (134) De Leon SF, Thurston GD, Ito K. Contribution of respiratory disease to nonrespiratory mortality associations with air pollution. *American Journal of Respiratory and Critical Care Medicine*. 2003;167(8):1117-1123.
- (135) Forastiere F, Stafoggia M, Berti G et al. Particulate matter and daily mortality - A case-crossover analysis of individual effect modifiers. *Epidemiology*. 2008;19(4):571-580.
- (136) Zhu RX, Chen YH, Wu SW, Deng FR, Liu Y, Yao WZ. The Relationship between Particulate Matter (PM10) and Hospitalizations and Mortality Of Chronic Obstructive Pulmonary Disease: A Meta-Analysis. *Copd-Journal of Chronic Obstructive Pulmonary Disease*. 2013;10(3):307-315.
- (137) Kan HD, Chen BH. A case-crossover analysis of air pollution and daily mortality in Shanghai. *Journal of Occupational Health*. 2003;45(2):119-124.
- (138) Meng X, Wang CC, Cao DC, Wong CM, Kan HD. Short-term effect of ambient air pollution on COPD mortality in four Chinese cities. *Atmospheric Environment*. 2013;77:149-154.
- (139) Neuberger M, Rabczenko D, Moshhammer H. Extended effects of air pollution on cardiopulmonary mortality in Vienna. *Atmospheric Environment*. 2007;41(38):8549-8556.
- (140) Gan WQ, FitzGerald JM, Carlsten C, Sadatsafavi M, Brauer M. Associations of Ambient Air Pollution with Chronic Obstructive Pulmonary Disease Hospitalization and Mortality. *American Journal of Respiratory and Critical Care Medicine*. 2013;187(7):721-727.
- (141) Schikowski T, Mills IC, Anderson HR et al. Ambient air pollution: a cause of COPD? *European Respiratory Journal*. 2014;43(1):250-263.
- (142) Davies CA, Dundas R, Leyland AH. Trends and Inequalities in Cardiovascular Disease Mortality in Scotland, 1974-2009. *Journal of Epidemiology and Community Health*. 2011;65:A11.
- (143) Finegold JA, Asaria P, Francis DP. Mortality from ischaemic heart disease by country, region, and age: Statistics from World Health Organisation and United Nations. *International Journal of Cardiology*. 2013;168(2):934-945.
- (144) Lopez AD, Mathers CD. Measuring the global burden of disease and epidemiological transitions: 2002-2030. *Annals of Tropical Medicine and Parasitology*. 2006;100(5-6):481-499.
- (145) Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. *Plos Medicine*. 2006;3(11).

- (146) Pope CA, Burnett RT, Thurston GD et al. Cardiovascular mortality and long-term exposure to particulate air pollution - Epidemiological evidence of general pathophysiological pathways of disease. *Circulation*. 2004;109(1):71-77.
- (147) Schwartz J. Air pollution and hospital admissions for heart disease in eight US counties. *Epidemiology*. 1999;10(1):17-22.
- (148) Levy D, Sheppard L, Checkoway H et al. A case-crossover analysis of particulate matter air pollution and out-of-hospital primary cardiac arrest. *Epidemiology*. 2001;12(2):193-199.
- (149) Samoli E, Zanobetti A, Schwartz J et al. The temporal pattern of mortality responses to ambient ozone in the APHEA project. *Journal of Epidemiology and Community Health*. 2009;63(12):960-966.
- (150) Braga ALF, Zanobetti A, Schwartz J. The lag structure between particulate air pollution and respiratory and cardiovascular deaths in 10 US cities. *Journal of Occupational and Environmental Medicine*. 2001;43(11):927-933.
- (151) Kim H, Kim Y, Hong YC. The lag-effect pattern in the relationship of particulate air pollution to daily mortality in Seoul, Korea. *International Journal of Biometeorology*. 2003;48(1):25-30.
- (152) Zanobetti A, Schwartz J, Samoli E et al. The temporal pattern of respiratory and heart disease mortality in response to air pollution. *Environmental Health Perspectives*. 2003;111(9):1188-1193.
- (153) Katsouyanni K, Touloumi G, Samoli E et al. Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 European cities within the APHEA2 project. *Epidemiology*. 2001;12(5):521-531.
- (154) Luo CM, Zhu XX, Yao CJ et al. Short-term exposure to particulate air pollution and risk of myocardial infarction: a systematic review and meta-analysis. *Environmental Science and Pollution Research*. 2015;22(19):14651-14662.
- (155) Schwartz J, Dockery DW, Neas LM. Is daily mortality associated specifically with fine particles? *Journal of the Air & Waste Management Association*. 1996;46(10):927-939.
- (156) Ostro B, Broadwin R, Green S, Feng WY, Lipsett M. Fine particulate air pollution and mortality in nine California counties: Results from CALFINE. *Environmental Health Perspectives*. 2006;114(1):29-33.
- (157) Pope CA, Muhlestein JB, May HT, Renlund DG, Anderson JL, Horne BD. Ischemic heart disease events triggered by short-term exposure to fine particulate air pollution. *Circulation*. 2006;114(23):2443-2448.
- (158) Xie WX, Li G, Zhao D et al. Relationship between fine particulate air pollution and ischaemic heart disease morbidity and mortality. *Heart*. 2015;101(4):257-263.
- (159) Kim SY, Dutton SJ, Sheppard L et al. The short-term association of selected components of fine particulate matter and mortality in the Denver Aerosol Sources and Health (DASH) study. *Environmental Health*. 2015;14.
- (160) Hong YC, Lee J, Kim H, Kwon H, Schwartz J, Christiani D. Air pollution on stroke and ischemic heart disease mortality in Seoul (1991-1997). *Epidemiology*. 2000;11(4):S106.
- (161) Neuberger M, Moshhammer H, Rabczenko D. Acute and Subacute Effects of Urban Air Pollution on Cardiopulmonary Emergencies and Mortality: Time Series Studies in

- Austrian Cities. *International Journal of Environmental Research and Public Health*. 2013;10(10):4728-4751.
- (162) Revich B, Shaposhnikov D. The effects of particulate and ozone pollution on mortality in Moscow, Russia. *Air Quality Atmosphere and Health*. 2010;3(2):117-123.
- (163) Neas LM, Schwartz J, Dockery D. A case-crossover analysis of air pollution and mortality in Philadelphia. *Environmental Health Perspectives*. 1999;107(8):629-631.
- (164) Wichmann J, Voyi K. Ambient Air Pollution Exposure and Respiratory, Cardiovascular and Cerebrovascular Mortality in Cape Town, South Africa: 2001-2006. *International Journal of Environmental Research and Public Health*. 2012;9(11):3978-4016.
- (165) Goldberg MS, Burnett RT, Brook J, Bailar JC, Valois MF, Vincent R. Associations between daily cause-specific mortality and concentrations of ground-level ozone in Montreal, Quebec. *American Journal of Epidemiology*. 2001;154(9):817-826.
- (166) Southampton Universities Hospitals NHS Trust. Medical Certificate of Cause of Death: Notes for Doctors. 2010. Available at: <http://www.uhs.nhs.uk/media/suhtideal/doctors/medicalpersonnelinduction/yourinductio nday/medicalcertificateofcauseofdeath-notesfordoctors.pdf>. Accessed October 10, 2015.
- (167) Chief Medical Officer Scotland. Guidance on completion of medical certificates of the cause of death. September 1, 2009. Available at: <http://www.sehd.scot.nhs.uk/cmo/CMO%282009%2910.pdf>. Accessed October 10, 2015.
- (168) ONS. Guidance for doctors completing Medical Certificates of Cause of Death in England and Wales. July 1, 2010. Available at: http://www.gro.gov.uk/images/medcert_july_2010.pdf. Accessed October 6, 2015.
- (169) WHO, WHO. ICD10 - International Statistical Classification of Diseases and Related Health Problems 10th Revision. 2011. Available at: http://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf. Accessed July 20, 2015.
- (170) Shavell RM, Paculdo DR, Kush SJ, Mannino DM, Strauss DJ. Life expectancy and years of life lost in chronic obstructive pulmonary disease: Findings from the NHANES III Follow-up Study. *International Journal of Chronic Obstructive Pulmonary Disease*. 2009;4:137-148.
- (171) Katsouyanni K, Karakatsani A, Messari I et al. Air-Pollution and Cause Specific Mortality in Athens. *Journal of Epidemiology and Community Health*. 1990;44(4):321-324.
- (172) Vaneckova P, Beggs PJ, de Dear RJ, McCracken KWJ. Effect of temperature on mortality during the six warmer months in Sydney, Australia, between 1993 and 2004. *Environmental Research*. 2008;108(3):361-369.
- (173) Hosmer DW, Lemeshow S, May S. Applied survival analysis regression modeling of time-to-event data. 2008;2nd ed.
- (174) Rothman K.J. *Epidemiology: An Introduction*. Oxford University Press Inc.; 2002.
- (175) Nyberg F, Gustavsson P, Jarup L et al. Urban air pollution and lung cancer in Stockholm. *Epidemiology*. 2000;11(5):487-495.

- (176) Abbey DE, Mills PK, Petersen FF, Beeson WL. Long-Term Ambient Concentrations of Total Suspended Particulates and Oxidants As Related to Incidence of Chronic Disease in California 7Th-Day-Adventists. *Environmental Health Perspectives*. 1991;94:43-50.
- (177) Dockery DW, Pope CA, Xu XP et al. An Association Between Air-Pollution and Mortality in 6 United-States Cities. *New England Journal of Medicine*. 1993;329(24):1753-1759.
- (178) Pope CA, Dockery DW. Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*. 2006;56(6):709-742.
- (179) Abbey DE, Nishino N, McDonnell WF et al. Long-term inhalable particles and other air pollutants related to mortality in nonsmokers. *American Journal of Respiratory and Critical Care Medicine*. 1999;159(2):373-382.
- (180) Pope CA, Burnett RT, Thun MJ et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama-Journal of the American Medical Association*. 2002;287(9):1132-1141.
- (181) Laden F, Schwartz J, Speizer FE, Dockery DW. Reduction in fine particulate air pollution and mortality - Extended follow-up of the Harvard six cities study. *American Journal of Respiratory and Critical Care Medicine*. 2006;173(6):667-672.
- (182) Brunekreef B. Design of cohort studies for air pollution health effects. *Journal of Toxicology and Environmental Health-Part A*. 2003;66(16-19):1723-1729.
- (183) Beelen R, Hoek G, van den Brandt PA et al. Long-term effects of traffic-related air pollution on mortality in a Dutch cohort (NLCS-AIR study). *Environmental Health Perspectives*. 2008;116(2):196-202.
- (184) Filleul L, Rondeau V, Vandentorren S et al. Twenty five year mortality and air pollution: results from the French PAARC survey. *Occupational and Environmental Medicine*. 2005;62(7):453-460.
- (185) Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*. 2011;46(3):399-424.
- (186) Austin PC. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*. 2012;31(29):3946-3958.
- (187) Platt RW, Joseph KS, Ananth CV, Grondines J, Abrahamowicz M, Kramer MS. A proportional hazards model with time-dependent covariates and time-varying effects for analysis of fetal and infant death. *American Journal of Epidemiology*. 2004;160(3):199-206.
- (188) D'Agostino RB. Beyond baseline data: the use of time-varying covariates. *Journal of Hypertension*. 2008;26(4):639-640.
- (189) Peters, A., von Klot, S., Berglind, N., Hormann, A., Lowel, H., Nyberg, F., Pekkanen, J., Perucci, C. A., Stafoggia, M., Sunyer, J., Tiittanen, P., and Forastiere, F. Comparison of different methods in analyzing short-term air pollution effects in a cohort study of susceptible individuals. *Epidemiologic Perspectives & Innovations* 3, 10-19. 9-8-2006.

Ref Type: Journal (Full)

- (190) Perrakis K, Gryparis A, Schwartz J et al. Controlling for seasonal patterns and time varying confounders in time-series epidemiological models: a simulation study. *Statistics in Medicine*. 2014;33(28):4904-4918.

- (191) Keatinge WR, Donaldson GC, Cordioli EA et al. Heat related mortality in warm and cold regions of Europe: observational study. *British Medical Journal*. 2000;321(7262):670-673.
- (192) Pattenden S, Nikiforov B, Armstrong BG. Mortality and temperature in Sofia and London. *Journal of Epidemiology and Community Health*. 2003;57(8):628-633.
- (193) Touloumi G, Pocock SJ, Katsouyanni K, Trichopoulos D. Short-Term Effects of Air-Pollution on Daily Mortality in Athens - A Time-Series Analysis. *International Journal of Epidemiology*. 1994;23(5):957-967.
- (194) Schwartz J, Spix C, Touloumi G et al. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology and Community Health*. 1996;50:S3-S11.
- (195) Schwartz J, Dockery DW. Particulate Air-Pollution and Daily Mortality in Steubenville, Ohio. *American Journal of Epidemiology*. 1992;135(1):12-19.
- (196) Schimmel H, Murawski TJ. Relation of Air-Pollution to Mortality. *Journal of Occupational and Environmental Medicine*. 1976;18(5):316-333.
- (197) Schwartz J, Marcus A. Mortality and Air-Pollution in London - A Time-Series Analysis. *American Journal of Epidemiology*. 1990;131(1):185-194.
- (198) Kinney PL, Ozkaynak H. Associations of Daily Mortality and Air-Pollution in Los-Angeles-County. *Environmental Research*. 1991;54(2):99-120.
- (199) Liang KY, Zeger SL. Longitudinal Data-Analysis Using Generalized Linear-Models. *Biometrika*. 1986;73(1):13-22.
- (200) Schwartz J. Particulate Air-Pollution and Daily Mortality in Detroit. *Environmental Research*. 1991;56(2):204-213.
- (201) Dockery DW, Schwartz J, Spengler JD. Air-Pollution and Daily Mortality - Associations with Particulates and Acid Aerosols. *Environmental Research*. 1992;59(2):362-373.
- (202) Schwartz J. Air-Pollution and Daily Mortality - A Review and Meta Analysis. *Environmental Research*. 1994;64(1):36-52.
- (203) Katsouyanni K, Schwartz J, Spix C et al. Short term effects of air pollution on health: A European approach using epidemiologic time series data: The APHEA protocol. *Journal of Epidemiology and Community Health*. 1996;50:S12-S18.
- (204) Katsouyanni K, Touloumi G, Spix C et al. Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: Results from time series data from the APHEA project. *British Medical Journal*. 1997;314(7095):1658-1663.
- (205) Wong TW, Lau TS, Yu TS et al. Air pollution and hospital admissions for respiratory and cardiovascular diseases in Hong Kong. *Occupational and Environmental Medicine*. 1999;56(10):679-683.
- (206) Dominici F, McDermott A, Zeger SL, Samet JM. On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*. 2002;156(3):193-203.
- (207) Cleveland WS, Devlin SJ. Locally Weighted Regression - An Approach to Regression-Analysis by Local Fitting. *Journal of the American Statistical Association*. 1988;83(403):596-610.

- (208) Laden F, Neas LM, Dockery DW, Schwartz J. Association of fine particulate matter from different sources with daily mortality in six US cities. *Environmental Health Perspectives*. 2000;108(10):941-947.
- (209) Baccini M, Biggeri A, Lagazio C, Lertxundi A, Saez M. Parametric and semi-parametric approaches in the analysis of short-term effects of air pollution on health. *Computational Statistics & Data Analysis*. 2007;51(9):4324-4336.
- (210) Touloumi G, Samoli E, Pipikou M, Le Tertre A, Atkinson R, Katsouyanni K. Seasonal confounding in air pollution and health time-series studies: Effect on air pollution effect estimates. *Statistics in Medicine*. 2006;25(24):4164-4178.
- (211) Schwartz J. Nonparametric Smoothing in the Analysis of Air-Pollution and Respiratory Illness. *Canadian Journal of Statistics-Revue Canadienne de Statistique*. 1994;22(4):471-487.
- (212) Fischer P, Hoek G, Brunekreef B, Verhoeff A, van Wijnen J. Air pollution and mortality in the Netherlands: are the elderly more at risk? *European Respiratory Journal*. 2003;21:34S-38S.
- (213) Ko FWS, Tam W, Wong TW et al. Temporal relationship between air pollutants and hospital admissions for chronic obstructive pulmonary disease in Hong Kong. *Thorax*. 2007;62(9):780-785.
- (214) Ramsay TO, Burnett RT, Krewski D. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*. 2003;14(1):18-23.
- (215) Peng RD, Dominici F, Louis TA. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society Series A-Statistics in Society*. 2006;169:179-198.
- (216) Mar TF, Norris GA, Koenig JQ, Larson TV. Associations between air pollution and mortality in Phoenix, 1995-1997. *Environmental Health Perspectives*. 2000;108(4):347-353.
- (217) Samoli E, Schwartz J, Wojtyniak B et al. Investigating regional differences in short-term effects of air pollution on daily mortality in the APHEA project: A sensitivity analysis for controlling long-term trends and seasonality. *Environmental Health Perspectives*. 2001;109(4):349-353.
- (218) Dominici F, Samet JM, Zeger SL. Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society Series A-Statistics in Society*. 2000;163:263-284.
- (219) Le Tertre A, Medina S, Samoli E et al. Short-term effects of particulate air pollution on cardiovascular diseases in eight European cities. *Journal of Epidemiology and Community Health*. 2002;56(10):773-779.
- (220) Moolgavkar SH. Air pollution and daily mortality in three US counties. *Environmental Health Perspectives*. 2000;108(8):777-784.
- (221) Dominici F, Sheppard L, Clyde M. Health effects of air pollution: A statistical review. *International Statistical Review*. 2003;71(2):243-276.
- (222) Maclure M. The Case-Crossover Design - A Method for Studying Transient Effects on the Risk of Acute Events. *American Journal of Epidemiology*. 1991;133(2):144-153.
- (223) Lumley T, Levy D. Bias in the case-crossover design: implications for studies of air pollution. *Environmetrics*. 2000;11(6):689-704.

- (224) Lee JT, Schwartz J. Reanalysis of the effects of air pollution on daily mortality in Seoul, Korea: A case-crossover design. *Environmental Health Perspectives*. 1999;107(8):633-636.
- (225) Schwartz J. The effects of particulate air pollution on daily deaths: a multi-city case crossover analysis. *Occupational and Environmental Medicine*. 2004;61(12):956-961.
- (226) Navidi W. Bidirectional case-crossover designs for exposures with time trends. *Biometrics*. 1998;54(2):596-605.
- (227) Mittleman MA, Maclure M, Tofler GH, Sherwood JB, Goldberg RJ, Muller JE. Triggering of Acute Myocardial-Infarction by Heavy Physical Exertion - Protection Against Triggering by Regular Exertion. *New England Journal of Medicine*. 1993;329(23):1677-1683.
- (228) Dominici F, McDermott A, Hastie TJ. Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association*. 2004;99(468):938-948.
- (229) Mittleman MA, Maclure M, Robins JM. Control Sampling Strategies for Case-Crossover Studies - An Assessment of Relative Efficiency. *American Journal of Epidemiology*. 1995;142(1):91-98.
- (230) Levy D, Lumley T, Sheppard L, Kaufman J, Checkoway H. Referent selection in case-crossover analyses of acute health effects of air pollution. *Epidemiology*. 2001;12(2):186-192.
- (231) Janes H, Sheppard L, Lumley T. Case-crossover analyses of air pollution exposure data - Referent selection strategies and their implications for bias. *Epidemiology*. 2005;16(6):717-726.
- (232) Peters A, Dockery DW, Muller JE, Mittleman MA. Increased particulate air pollution and the triggering of myocardial infarction. *Circulation*. 2001;103(23):2810-2815.
- (233) Lee JT, Shin DC, Chung Y. Air pollution and daily mortality in Seoul and Ulsan, Korea. *Environmental Health Perspectives*. 1999;107(2):149-154.
- (234) Greenland S. Confounding and exposure trends in case crossover and case time-control designs. *Epidemiology*. 1996;7(3):231-239.
- (235) Janes H, Sheppard L, Lumley T. Overlap bias in the case-crossover design, with application to air pollution exposures. *Statistics in Medicine*. 2005;24(2):285-300.
- (236) Bateson TF, Schwartz J. Control for seasonal variation and time trend in case crossover studies of acute effects of environmental exposures. *Epidemiology*. 1999;10(5):539-544.
- (237) Tsai SS, Huang CH, Goggins WB, Wu TN, Yang CY. Relationship between air pollution and daily mortality in a Tropical City: Kaohsiung, Taiwan. *Journal of Toxicology and Environmental Health-Part A*. 2003;66(14):1341-1349.
- (238) Lin M, Chen Y, Burnett RT, Villeneuve PJ, Krewski D. Effect of short-term exposure to gaseous pollution on asthma hospitalisation in children: a bi-directional case-crossover analysis. *Journal of Epidemiology and Community Health*. 2003;57(1):50-55.
- (239) Kwon HJ, Cho SH, Nyberg F, Pershgen G. Effects of ambient air pollution on daily mortality in a cohort of patients with congestive heart failure. *Epidemiology*. 2001;12(4):413-419.

- (240) Yang CY, Chen CJ. Air pollution and hospital admissions for chronic obstructive pulmonary disease in a subtropical city: Taipei, Taiwan. *Journal of Toxicology and Environmental Health-Part A-Current Issues*. 2007;70(13-14):1214-1219.
- (241) Henrotin JB, Besancenot JP, Bejot Y, Giroud M. Short-term effects of ozone air pollution on ischaemic stroke occurrence: a case-crossover analysis from a 10-year population-based study in Dijon, France. *Occupational and Environmental Medicine*. 2007;64(7):439-445.
- (242) Lee JT, Kim H, Schwartz J. Bidirectional case-crossover studies of air pollution: Bias from skewed and incomplete waves. *Environmental Health Perspectives*. 2000;108(12):1107-1111.
- (243) Navidi W, Weinhandl E. Risk set sampling for case-crossover designs. *Epidemiology*. 2002;13(1):100-105.
- (244) Carracedo-Martinez E, Taracido M, Tobias A, Saez M, Figueiras A. Case-Crossover Analysis of Air Pollution Health Effects: A Systematic Review of Methodology and Application. *Environmental Health Perspectives*. 2010;118(8):1173-1182.
- (245) Whitaker HJ, Hocine MN, Farrington CP. On case-crossover methods for environmental time series data. *Environmetrics*. 2007;18(2):157-171.
- (246) Sullivan J, Ishikawa N, Sheppard L, Siscovick D, Checkoway H, Kaufman J. Exposure to ambient fine particulate matter and primary cardiac arrest among persons with and without clinically recognized heart disease. *American Journal of Epidemiology*. 2003;157(6):501-509.
- (247) D'Ippoliti D, Forastiere F, Ancona C et al. Air pollution and myocardial infarction in Rome - A case-crossover analysis. *Epidemiology*. 2003;14(5):528-535.
- (248) Schwartz J. Is the association of airborne particles with daily deaths confounded by gaseous air pollutants? An approach to control by matching. *Environmental Health Perspectives*. 2004;112(5):557-561.
- (249) Barnett AG, Williams GM, Schwartz J et al. The effects of air pollution on hospitalizations for cardiovascular disease in elderly people in Australian and New Zealand cities. *Environmental Health Perspectives*. 2006;114(7):1018-1023.
- (250) Rich DQ, Kim MH, Turner JR et al. Association of ventricular arrhythmias detected by implantable cardioverter defibrillator and ambient air pollutants in the St Louis, Missouri metropolitan area. *Occupational and Environmental Medicine*. 2006;63(9):591-596.
- (251) Zanobetti A, Schwartz J. Air pollution and emergency admissions in Boston, MA. *Journal of Epidemiology and Community Health*. 2006;60(10):890-895.
- (252) Lu Y, Zeger SL. On the equivalence of case-crossover and time series methods in environmental epidemiology. *Biostatistics*. 2007;8(2):337-344.
- (253) Navidi W. Poisson regression and the case-crossover design: Similarities and differences. *Communications in Statistics-Theory and Methods*. 2008;37(2):213-220.
- (254) Fung KY, Krewski D, Chen Y, Burnett R, Cakmak S. Comparison of time series and case-crossover analyses of air pollution and hospital admission data. *International Journal of Epidemiology*. 2003;32(6):1064-1070.
- (255) Lu Y, Symons JM, Geyh AS, Zeger SL. An approach to checking case-crossover analyses based on equivalence with time-series methods. *Epidemiology*. 2008;19(2):169-175.

- (256) Schwartz J. Assessing confounding, effect modification, and thresholds in the association between ambient particles and daily deaths. *Environmental Health Perspectives*. 2000;108(6):563-568.
- (257) Schwartz J, Zanobetti A. Using meta-smoothing to estimate dose-response trends across multiple studies, with application to air pollution and daily death. *Epidemiology*. 2000;11(6):666-672.
- (258) Daniels MJ, Dominici F, Samet JM, Zeger SL. Estimating particulate matter-mortality dose-response curves and threshold levels: An analysis of daily time-series for the 20 largest US cities. *American Journal of Epidemiology*. 2000;152(5):397-406.
- (259) Kim SY, Lee JT, Hong YC, Ahn KJ, Kim H. Determining the threshold effect of ozone on daily mortality: an analysis of ozone and mortality in Seoul, Korea, 1995-1999. *Environmental Research*. 2004;94(2):113-119.
- (260) Schwartz J, Ballester F, Saez M et al. The concentration-response relation between air pollution and daily deaths. *Environmental Health Perspectives*. 2001;109(10):1001-1006.
- (261) Schwartz J, Laden F, Zanobetti A. The concentration-response relation between PM_{2.5} and daily deaths. *Environmental Health Perspectives*. 2002;110(10):1025-1029.
- (262) Samoli E, Touloumi G, Zanobetti A et al. Investigating the dose-response relation between air pollution and total mortality in the APHEA-2 multicity project. *Occupational and Environmental Medicine*. 2003;60(12):977-982.
- (263) Dominici F, Daniels M, Zeger SL, Samet JM. Air pollution and mortality: Estimating regional and national dose-response relationships. *Journal of the American Statistical Association*. 2002;97(457):100-111.
- (264) Samoli E, Analitis A, Touloumi G et al. Estimating the exposure-response relationships between particulate matter and mortality within the APHEA multicity project. *Environmental Health Perspectives*. 2005;113(1):88-95.
- (265) Rossi G, Vigotti MA, Zanobetti A, Repetto F, Gianelle V, Schwartz J. Air pollution and cause-specific mortality in Milan, Italy, 1980-1989. *Archives of Environmental Health*. 1999;54(3):158-164.
- (266) Lee JT, Kim H, Hong YC, Kwon HJ, Schwartz J, Christiani DC. Air pollution and daily mortality in seven major cities of Korea, 1991-1997. *Environmental Research*. 2000;84(3):247-254.
- (267) Villeneuve PJ, Chen L, Rowe BH, Coates F. Outdoor air pollution and emergency department visits for asthma among children and adults: A case-crossover study in northern Alberta, Canada. *Environmental Health*. 2007;6.
- (268) Checkoway H, Pearce N, Kriebel D. *Research methods in occupational epidemiology*. 2nd ed ed. Oxford: Oxford University Press; 2004.
- (269) Thomas DC. *Statistical methods in environmental epidemiology*. 2009.
- (270) Cakmak S, Dales RE, Vidal CB. Air pollution and mortality in Chile: Susceptibility among the elderly. *Environmental Health Perspectives*. 2007;115(4):524-527.
- (271) Armstrong B. Models for the relationship between ambient temperature and daily mortality. *Epidemiology*. 2006;17(6):624-631.

- (272) Carder M, McNamee R, Beverland I et al. Interacting effects of particulate pollution and cold temperature on cardiorespiratory mortality in Scotland. *Occupational and Environmental Medicine*. 2008;65(3):197-204.
- (273) Carder M, McNamee R, Beverland I et al. Does deprivation index modify the acute effect of black smoke on cardiorespiratory mortality? *Occupational and Environmental Medicine*. 2010;67(2):104-110.
- (274) Almon S. The Distributed Lag Between Capital Appropriations and Expenditures. *Econometrica*. 1965;33(1):178-196.
- (275) Zanobetti A, Schwartz J, Dockery DW. Airborne particles are a risk factor for hospital admissions for heart and lung disease. *Environmental Health Perspectives*. 2000;108(11):1071-1077.
- (276) Martins LC, Pereira LAA, Lin CA et al. The effects of air pollution on cardiovascular diseases: lag structures. *Revista de Saude Publica*. 2006;40(4):677-683.
- (277) Leitte AM, Petrescu C, Franck U et al. Respiratory health, effects of ambient air pollution and its modification by air humidity in Drobeta-Turnu Severin, Romania. *Science of the Total Environment*. 2009;407(13):4004-4011.
- (278) Zanobetti A, Schwartz J, Samoli E et al. The temporal pattern of mortality responses to air pollution: A multicity assessment of mortality displacement. *Epidemiology*. 2002;13(1):87-93.
- (279) Zanobetti A, Wand MP, Schwartz J, Ryan LM. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*. 2000;1(3):279-292.
- (280) Koop G, Tole L. An investigation of thresholds in air pollution-mortality effects. *Environmental Modelling & Software*. 2006;21(12):1662-1673.
- (281) Roberts S, Martin MA. A distributed lag approach to fitting non-linear dose-response models in particulate matter air pollution time series investigations. *Environmental Research*. 2007;104(2):193-200.
- (282) Gasparini A, Armstrong B, Kenward MG. Distributed lag non-linear models. *Statistics in Medicine*. 2010;29(21):2224-2234.
- (283) Zeger SL, Dominici F, Samet J. Harvesting-resistant estimates of air pollution effects on mortality. *Epidemiology*. 1999;10(2):171-175.
- (284) Fung K, Krewski D, Burnett R, Dominici F. Testing the harvesting hypothesis by time-domain regression analysis. 1: Baseline analysis. *Journal of Toxicology and Environmental Health-Part A-Current Issues*. 2005;68(13-14):1137-1154.
- (285) Fung K, Krewski D, Burnett R, Ramsay T, Chen Y. Testing the harvesting hypothesis by time-domain regression analysis. II: Covariate effects. *Journal of Toxicology and Environmental Health-Part A-Current Issues*. 2005;68(13-14):1155-1165.
- (286) Schwartz J. Harvesting and long term exposure effects in the relation between air pollution and mortality. *American Journal of Epidemiology*. 2000;151(5):440-448.
- (287) Zanobetti A, Schwartz J. Mortality displacement in the association of ozone with mortality - An analysis of 48 cities in the United States. *American Journal of Respiratory and Critical Care Medicine*. 2008;177(2):184-189.
- (288) Roberts S, Switzer P. Mortality displacement and distributed lag models. *Inhalation Toxicology*. 2004;16(14):879-888.

- (289) Little RJA. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*. 1988;6(3):287-296.
- (290) Horton NJ, Lipsitz SR. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*. 2001;55(3):244-254.
- (291) Katsouyanni K, Zmirou D, Spix C et al. Short-Term Effects of Air-Pollution on Health - A European Approach Using Epidemiologic Time-Series Data - the Apeha Project - Background, Objectives, Design. *European Respiratory Journal*. 1995;8(6):1030-1038.
- (292) Rubin DB. Inference and Missing Data. *Biometrika*. 1976;63(3):581-590.
- (293) Allison PD, SAGE Research MO. *Missing data*. no. 07-136 ed. Thousand Oaks, Calif: Sage Publications; 2002.
- (294) Enders CK. *Applied missing data analysis*. New York ; London : Guilford Press, 2010.
- (295) Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association*. 1996;91(434):473-489.
- (296) Carpenter JR, Kenward MG. *Multiple imputation and its application*. First edition. ed. Chichester, West Sussex : John Wiley & Sons, 2013.
- (297) Little RJA, Rubin DB. *Statistical analysis with missing data*. 2nd ed. ed. Hoboken, N.J. ; [Chichester] : Wiley, 2002.
- (298) Sterne JAC, White IR, Carlin JB et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*. 2009;339.
- (299) He YL. Missing Data Analysis Using Multiple Imputation Getting to the Heart of the Matter. *Circulation-Cardiovascular Quality and Outcomes*. 2010;3(1):98-U145.
- (300) Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley classics library ed. ed. Hoboken, N.J. : John Wiley, 1987.
- (301) Schafer JL. *Analysis of incomplete multivariate data*. 1st ed. ed. London : Chapman & Hall, 1997.
- (302) Heitjan DF, Little RJA. Multiple Imputation for the Fatal Accident Reporting System. *Applied Statistics-Journal of the Royal Statistical Society Series C*. 1991;40(1):13-29.
- (303) Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*. 1996;22(4):425-446.
- (304) Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. *Statistics in Medicine*. 2001;20(9-10):1541-1549.
- (305) van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011;45(3):1-67.
- (306) Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software*. 2011;45(4):1-20.
- (307) van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*. 1999;18(6):681-694.

- (308) Raghunathan T, Lepkowski J, Van Hoewyk J, Solenberger P. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodol.* 2001;27(1):85-95.
- (309) van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research.* 2007;16(3):219-242.
- (310) Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research.* 2011;20(1):40-49.
- (311) Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. *American Statistician.* 2003;57(4):229-232.
- (312) van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation.* 2006;76(12):1049-1064.
- (313) Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research.* 1999;8(1):3-15.
- (314) White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine.* 2011;30(4):377-399.
- (315) Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology.* 2006;59(10):1092-1101.
- (316) Carpenter JR, Goldstein H, Kenward MG. REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *Journal of Statistical Software.* 2011;45(5):1-14.
- (317) Yucel RM. State of the Multiple Imputation Software. *Journal of Statistical Software.* 2011;45(1):1-7.
- (318) NHS National Services Scotland. Information Services Division Scotland (ISD). 2015. Available at: <http://www.isdscotland.org/index.asp>. Accessed July 20, 2015.
- (319) Uk Government. National Records of Scotland. 2015. Available at: <http://www.nrscotland.gov.uk/registration/registering-a-death>. Accessed July 20, 2015.
- (320) WHO. International Classification of Diseases (ICD). 2015. Available at: <http://www.who.int/classifications/icd/en/>. Accessed July 20, 2015.
- (321) Royal Mail. Postcode Address File Statistics 2011. December 12, 2011. Available at: http://www.royalmail.com/sites/default/files/RM_122011_%20PAF%20Stats.pdf. Accessed February 28, 2016.
- (322) ONS. Total number of households by region and country of the UK, 1996 to 2015. February 17, 2016. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/adhocs/005374totalnumberofhouseholdsbyregionandcountryoftheuk1996to2015>. Accessed February 28, 2016.
- (323) Medical Research Council. MRC Data and Tissues Tool Kit. March, 2015. Available at: <http://www.dt-toolkit.ac.uk/home.cfm>. Accessed February 2, 2016.
- (324) Medical Research Council. MRC Data and Tissues Toolkit - Route Maps. March, 2015. Available at: <http://www.dt-toolkit.ac.uk/routemaps.cfm>. Accessed February 2, 2016.

- (325) National Records of Scotland. Mid-2011 and Mid-2012 Population Estimates Scotland. February 27, 2014. Available at: <http://gro-scotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/mid-year-population-estimates/2012>. Accessed September 2, 2015.
- (326) DEFRA. Department for Environment Food and Rural Affairs. 2015. Available at: <http://uk-air.defra.gov.uk/>. Accessed March 20, 15 A.D.
- (327) National Records of Scotland, Nation. Death Time Series Data. August 20, 2015. Available at: <http://www.gro-scotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/deaths/deaths-time-series-data>. Accessed September 5, 2015.
- (328) Harrell FE. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. 2001.
- (329) Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine*. 2006;25(24):4279-4292.
- (330) Stata Statistical Software: Release 13. Version 13. College Station: TX: StataCorp LP; 2013.
- (331) Collett D. *Modelling binary data*. 2nd ed ed. Boca Raton: Chapman & Hall/CRC; 2003.
- (332) Hosmer DW, Lemeshow S, Sturdivant RX, ebrary I. *Applied logistic regression*. 3rd ed ed. Hoboken, N.J: Wiley; 2013.
- (333) Hardin JW, Hilbe J. *Generalized linear models and extensions*. 3rd ed ed. College Station, Tex: Stata; 2012.
- (334) Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*. 2002;55(1):86-94.
- (335) Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal*. 2010;340.
- (336) Kontopantelis E, Reeves D. metaan: Random-effects meta-analysis. *The Stata Journal*. 2010;10(3):395-407.
- (337) Loader A, Sneddon S, Willis P, Stevenson K, Yardley R. Air Pollution in Scotland 2010. May, 2011. Available at: http://www.scottishairquality.co.uk/assets/documents/reports2/309110404_AQ_scot_14_lowres.pdf. Accessed December 23, 2015.
- (338) Gittins M, McNamee R, Carder M, Beverland I, Agius RM. Has the short-term effect of black smoke exposure on pneumonia mortality been underestimated because hospitalisation is ignored: findings from a case-crossover study. *Environmental Health*. 2013;12.
- (339) Menard SW. *Applied logistic regression analysis*. 2nd ed. ed. Thousand Oaks CA ; London : Sage Publications, 2002.
- (340) Public Health England. Heatwave Plan for England. May 22, 2015. Available at: <https://www.gov.uk/government/publications/heatwave-plan-for-england>. Accessed January 22, 2016.
- (341) Public Health England. Cold weather plan (CWP) for England. January 19, 2016. Available at: <https://www.gov.uk/government/publications/cold-weather-plan-cwp-for-england>. Accessed January 21, 2016.

- (342) Hansell AL, Walk JA, Soriano JB. What do chronic obstructive pulmonary disease patients die from? A multiple cause coding analysis. *European Respiratory Journal*. 2003;22(5):809-814.
- (343) Jensen HH, Godtfredsen NS, Lange P, Vestbo J. Potential misclassification of causes of death from COPD. *European Respiratory Journal*. 2006;28(4):781-785.
- (344) Symons JM, Wang L, Guallar E et al. A case-crossover study of fine particulate matter air pollution and onset of congestive heart failure symptom exacerbation leading to hospitalization. *American Journal of Epidemiology*. 2006;164(5):421-433.
- (345) Gotschi T, Oglesby L, Mathys P et al. Comparison of black smoke and PM_{2.5} levels in indoor and outdoor environments of four European cities. *Environmental Science & Technology*. 2002;36(6):1191-1197.
- (346) Hoek G, Kos G, Harrison R et al. Indoor-outdoor relationships of particle number and mass in four European cities. *Atmospheric Environment*. 2008;42(1):156-169.
- (347) Williams R, Creason J, Zweidinger R, Watts R, Sheldon L, Shy C. Indoor, outdoor, and personal exposure monitoring of particulate air pollution: the Baltimore elderly epidemiology-exposure pilot study. *Atmospheric Environment*. 2000;34(24):4193-4204.
- (348) Vallejo M, Lerma C, Infante O, Hermosillo AG, Riojas-Rodriguez H, Cardenas M. Personal exposure to particulate matter less than 2.5 μ m in Mexico City: a pilot study. *Journal of Exposure Analysis and Environmental Epidemiology*. 2004;14(4):323-329.
- (349) Brauer M, Hirtle RD, Hall AC, Yip TR. Monitoring personal fine particle exposure with a particle counter. *Journal of Exposure Analysis and Environmental Epidemiology*. 1999;9(3):228-236.
- (350) Janssen NAH, Hoek G, Brunekreef B, Harssema H, Mensink I, Zuidhof A. Personal sampling of particles in adults: Relation among personal, indoor, and outdoor air concentrations. *American Journal of Epidemiology*. 1998;147(6):537-547.
- (351) Janssen NAH, Lanki T, Hoek G et al. Associations between ambient, personal, and indoor exposure to fine particulate matter constituents in Dutch and Finnish panels of cardiovascular patients. *Occupational and Environmental Medicine*. 2005;62(12):868-877.
- (352) Wang XH, Bi XH, Sheng GY, Fu HM. Hospital indoor PM₁₀/PM_{2.5} and associated trace elements in Guangzhou, China. *Science of the Total Environment*. 2006;366(1):124-135.
- (353) Morawska L, Jamriska M, Guo H, Jayaratne ER, Cao M, Summerville S. Variation in indoor particle number and PM_{2.5} concentrations in a radio station surrounded by busy roads before and after an upgrade of the HVAC system. *Building and Environment*. 2009;44(1):76-84.
- (354) Meier R, Eeftens M, Phuleria HC et al. Differences in indoor versus outdoor concentrations of ultrafine particles, PM_{2.5}, PM_{absorbance} and NO₂ in Swiss homes. *Journal of Exposure Science and Environmental Epidemiology*. 2015;25(5):499-505.
- (355) Meier R, Schindler C, Eeftens M et al. Modeling indoor air pollution of outdoor origin in homes of SAPALDIA subjects in Switzerland. *Environment International*. 2015;82:85-91.
- (356) Polidori A, Arhami M, Sioutas C, Delfino RJ, Allen R. Indoor/outdoor relationships, trends, and carbonaceous content of fine particulate matter in retirement homes of the

- Los Angeles basin. *Journal of the Air & Waste Management Association*. 2007;57(3):366-379.
- (357) Franklin M, Zeka A, Schwartz J. Association between PM2.5 and all-cause and specific-cause mortality in 27 US communities. *Journal of Exposure Science and Environmental Epidemiology*. 2007;17(3):279-287.
- (358) Bell ML, Ebisu K, Peng RD, Dominici F. Adverse Health Effects of Particulate Air Pollution Modification by Air Conditioning. *Epidemiology*. 2009;20(5):682-686.
- (359) Chen C, Zhao B, Weschler CJ. Indoor Exposure to "Outdoor PM10" Assessing Its Influence on the Relationship Between PM10 and Short-term Mortality in US Cities. *Epidemiology*. 2012;23(6):870-878.
- (360) Vedal S. Does Air Conditioning Modify the Health Effects of Exposure to Outdoor Air Pollution? *Epidemiology*. 2009;20(5):687-688.
- (361) Shilton V, Giess P, Mitchell D, Williams C. The relationships between indoor and outdoor respirable particulate matter: Meteorology, chemistry and personal exposure. *Indoor and Built Environment*. 2002;11(5):266-274.
- (362) Maynard C, Althouse R, Olsufka M, Ritchie JL, Davis KB, Kennedy JW. Early Versus Late Hospital Arrival for Acute Myocardial-Infarction in the Western Washington Thrombolytic Therapy Trials. *American Journal of Cardiology*. 1989;63(18):1296-1300.
- (363) Newby LK, Rutsch WR, Califf RM et al. Time from symptom onset to treatment and outcomes after thrombolytic therapy. *Journal of the American College of Cardiology*. 1996;27(7):1646-1655.
- (364) Weaver WD. Time to Thrombolytic Treatment - Factors Affecting Delay and Their Influence on Outcome. *Journal of the American College of Cardiology*. 1995;25(7):S3-S9.
- (365) Boersma E, Maas ACP, Deckers JW, Simoons ML. Early thrombolytic treatment in acute myocardial infarction: Reappraisal of the golden hour. *Lancet*. 1996;348(9030):771-775.
- (366) Atkinson RW, Bremner SA, Anderson HR, Strachan DP, Bland JM, de Leon AP. Short-term associations between emergency hospital admissions for respiratory and cardiovascular disease and outdoor air pollution in London. *Archives of Environmental Health*. 1999;54(6):398-411.
- (367) Dominici F, Peng RD, Bell ML et al. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Jama-Journal of the American Medical Association*. 2006;295(10):1127-1134.
- (368) San Tam WW, Wong TW, Wong AHS. Association between air pollution and daily mortality and hospital admission due to ischaemic heart diseases in Hong Kong. *Atmospheric Environment*. 2015;120:360-368.
- (369) Lee IM, Tsai SS, Ho CK, Chiu HF, Wu TN, Yang CY. Air pollution and hospital admissions for congestive heart failure: Are there potentially sensitive groups? *Environmental Research*. 2008;108(3):348-353.
- (370) Uzoigwe JC, Prum T, Bresnahan E, Garelnabi M. The emerging role of outdoor and indoor air pollution in cardiovascular disease. *N Am J Med Sci*. 2013;5(8):445-453.
- (371) Department of Health. Health Building Note 04-01 Supplement 1: Isolation facilities for infectious patients in acute settings. 2015. Available at:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/148503/HBN_04-01_Supp_1_Final.pdf. Accessed March 21, 2016.

- (372) Sapkota A, Williams D, Buckley TJ. Tollbooth workers and mobile source-related hazardous air pollutants: How protective is the indoor environment? *Environmental Science & Technology*. 2005;39(9):2936-2943.
- (373) Anderson M, Carmichael C, Murray V, Dengel A, Swainson M. Defining indoor heat thresholds for health in the UK. *Perspectives in Public Health*. 2013;133(3):158-164.
- (374) Nguyen JL, Schwartz J, Dockery DW. The relationship between indoor and outdoor temperature, apparent temperature, relative humidity, and absolute humidity. *Indoor Air*. 2014;24(1):103-112.
- (375) Nguyen JL, Dockery DW. Daily indoor-to-outdoor temperature and humidity relationships: a sample across seasons and diverse climatic regions. *International Journal of Biometeorology*. 2016;60(2):221-229.
- (376) Goldberg MS, Wheeler AJ, Burnett RT et al. Physiological and perceived health effects from daily changes in air pollution and weather among persons with heart failure: A panel study. *Journal of Exposure Science and Environmental Epidemiology*. 2015;25(2):187-199.
- (377) Saeki K, Obayashi K, Iwamoto J et al. Stronger association of indoor temperature than outdoor temperature with blood pressure in colder months. *Journal of Hypertension*. 2014;32(8):1582-1589.
- (378) Osman LM, Ayres JG, Garden C, Reglitz K, Lyon J, Douglas JG. Home warmth and health status of COPD patients. *European Journal of Public Health*. 2008;18(4):399-405.
- (379) Loughnan M, Carroll M, Tapper NJ. The relationship between housing and heat wave resilience in older people. *International Journal of Biometeorology*. 2015;59(9):1291-1298.
- (380) Karr C, Lumley T, Shepherd K et al. A case-crossover study of wintertime ambient air pollution and infant bronchiolitis. *Environ Health Perspect*. 2006;114(2):277-281.
- (381) Chen RJ, Zhang YH, Yang CX, Zhao ZH, Xu XH, Kan HD. Acute Effect of Ambient Air Pollution on Stroke Mortality in the China Air Pollution and Health Effects Study. *Stroke*. 2013;44(4):954-960.
- (382) Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research*. 2007;16(3):277-298.
- (383) Samet JM, Dominici F, Zeger SL, Schwartz J, Dockery DW. The National Morbidity, Mortality, and Air Pollution Study. Part I: Methods and methodologic issues. *Res Rep Health Eff Inst*. 2000;(94 Pt 1):5-14.
- (384) Tramuto F, Cusimano R, Cerame G et al. Urban air pollution and emergency room admissions for respiratory symptoms: a case-crossover study in Palermo, Italy. *Environmental Health*. 2011;10.
- (385) Sajani SZ, Hanninen O, Marchesi S, Lauriola P. Comparison of different exposure settings in a case-crossover study on air pollution and daily mortality: counterintuitive results. *Journal of Exposure Science and Environmental Epidemiology*. 2011;21(4):385-394.

- (386) Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*. 2004;38(18):2895-2907.
- (387) Plaia A, Bondi AL. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*. 2006;40(38):7316-7330.
- (388) Molitor J, Molitor NT, Jerrett M et al. Bayesian modeling of air pollution health effects with missing exposure data. *American Journal of Epidemiology*. 2006;164(1):69-76.
- (389) Jimenez E, Linares C, Rodriguez LF, Bleda MJ, Diaz J. Short-term impact of particulate matter (PM_{2.5}) on daily mortality among the over-75 age group in Madrid (Spain). *Science of the Total Environment*. 2009;407(21):5486-5492.
- (390) Mate T, Guaita R, Pichiule M, Linares C, Diaz J. Short-term effect of fine particulate matter (PM_{2.5}) on daily mortality due to diseases of the circulatory system in Madrid (Spain). *Science of the Total Environment*. 2010;408(23):5750-5757.
- (391) Norazian MN, Shukri YA, Azam RN, Al Bakri AMM. Estimation of missing values in air pollution data using single imputation techniques. *Scienceasia*. 2008;34(3):341-345.
- (392) Seaman SR, Keogh RH. Handling Missing Data in Matched Case-Control Studies Using Multiple Imputation. *Biometrics*. 2015;71(4):1150-1159.
- (393) DEFRA. Defra National Statistics Release: Air quality statistics in the UK, 1987 to 2014. April 23, 2015. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/423353/National_Statistic_on_Air_Quality_2014.pdf. Accessed March 1, 2016.
- (394) Guaita R, Pichiule M, Mate T, Linares C, Diaz J. Short-term impact of particulate matter (PM_{2.5}) on respiratory mortality in Madrid. *International Journal of Environmental Health Research*. 2011;21(4):260-274.
- (395) Lee M, Koutrakis P, Coull B, Kloog I, Schwartz J. Acute effect of fine particulate matter on mortality in three Southeastern states from 2007-2011. *Journal of Exposure Science and Environmental Epidemiology*. 2015;26:173-179.
- (396) Hong YC, Lee JT, Kim H, Kwon HJ. Air pollution - A new risk factor in ischemic stroke mortality. *Stroke*. 2002;33(9):2165-2169.
- (397) Qiu H, Yu ITS, Tian LW et al. Effects of Coarse Particulate Matter on Emergency Hospital Admissions for Respiratory Diseases: A Time-Series Analysis in Hong Kong. *Environmental Health Perspectives*. 2012;120(4):572-576.
- (398) Baldasano JM, Valera E, Jimenez P. Air quality data from large cities. *Science of the Total Environment*. 2003;307(1-3):141-165.
- (399) Mayer H. Air pollution in cities. *Atmospheric Environment*. 1999;33(24-25):4029-4037.
- (400) Curriero FC, Heiner KS, Samet JM, Zeger SL, Strug L, Patz JA. Temperature and mortality in 11 cities of the eastern United States. *American Journal of Epidemiology*. 2002;155(1):80-87.
- (401) Doyle R. Deaths from excessive cold and excessive heat. *Scientific American*. 1998;278(2):26.
- (402) Schwartz J. Who is sensitive to extremes of temperature? A case-only analysis. *Epidemiology*. 2005;16(1):67-72.

- (403) Callaly E, Mikulich O, Silke B. Increased winter mortality: The effect of season, temperature and deprivation in the acutely ill medical patient. *European Journal of Internal Medicine*. 2013;24(6):546-551.
- (404) Goldberg MS, Gasparini A, Armstrong B, Valois MF. The short-term influence of temperature on daily mortality in the temperate climate of Montreal, Canada. *Environmental Research*. 2011;111(6):853-860.
- (405) Yu WW, Hu WB, Mengersen K et al. Time course of temperature effects on cardiovascular mortality in Brisbane, Australia. *Heart*. 2011;97(13):1089-1093.
- (406) Ha J, Shin Y, Kim H. Distributed lag effects in the relationship between temperature and mortality in three major cities in South Korea. *Science of the Total Environment*. 2011;409(18):3274-3280.
- (407) Braga ALF, Zanobetti A, Schwartz J. The effect of weather on respiratory and cardiovascular deaths in 12 US cities. *Environmental Health Perspectives*. 2002;110(9):859-863.
- (408) Zeka A, Browne S, McAvoy H, Goodman P. The association of cold weather and all-cause and cause-specific mortality in the island of Ireland between 1984 and 2007. *Environmental Health*. 2014;13.
- (409) Yi W, Chan APC. Effects of temperature on mortality in Hong Kong: a time series analysis. *International Journal of Biometeorology*. 2015;59(7):927-936.
- (410) Dang TN, Seposo XT, Duc NHC et al. Characterizing the relationship between temperature and mortality in tropical and subtropical cities: a distributed lag non-linear model analysis in Hue, Viet Nam, 2009-2013. *Global Health Action*. 2016;9.
- (411) Wang CC, Chen RJ, Kuang XY, Duan XL, Kan HD. Temperature and daily mortality in Suzhou, China: A time series analysis. *Science of the Total Environment*. 2014;466:985-990.
- (412) Keatinge WR, Donaldson GC, Bucher K et al. Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. *Lancet*. 1997;349(9062):1341-1346.
- (413) Gosling SN, Lowe JA, McGregor GR, Pelling M, Malamud BD. Associations between elevated atmospheric temperature and human mortality: a critical review of the literature. *Climatic Change*. 2009;92(3-4):299-341.
- (414) Breitner S, Wolf K, Peters A, Schneider A. Short-term effects of air temperature on cause-specific cardiovascular mortality in Bavaria, Germany. *Heart*. 2014;100(16):1272-1280.
- (415) Yang J, Ou CQ, Ding Y, Zhou YX, Chen PY. Daily temperature and mortality: a study of distributed lag non-linear effect and effect modification in Guangzhou. *Environmental Health*. 2012;11.
- (416) Guo YM, Li SS, Zhang YS et al. Extremely cold and hot temperatures increase the risk of ischaemic heart disease mortality: epidemiological evidence from China. *Heart*. 2013;99(3):195-203.
- (417) Tian ZX, Li SS, Zhang JL, Jaakkola JJK, Guo YM. Ambient temperature and coronary heart disease mortality in Beijing, China: a time series study. *Environmental Health*. 2012;11.

- (418) Wang XY, Li GX, Liu LQ, Westerdahl D, Jin XB, Pan XC. Effects of Extreme Temperatures on Cause-Specific Cardiovascular Mortality in China. *International Journal of Environmental Research and Public Health*. 2015;12(12):16136-16156.
- (419) Bhaskaran K, Armstrong B, Hajat S, Haines A, Wilkinson P, Smeeth L. Heat and risk of myocardial infarction: hourly level case-crossover analysis of MINAP database. *British Medical Journal*. 2012;345.
- (420) Mandell LA, Wunderink RG, Anzueto A et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clin Infect Dis*. 2007;44 Suppl 2:S27-S72.
- (421) Saldiva PHN, King M, Delmonte VLC et al. Respiratory Alterations Due to Urban Air-Pollution - An Experimental-Study in Rats. *Environmental Research*. 1992;57(1):19-33.
- (422) Zhou HW, Kobzik L. Effect of concentrated ambient particles on macrophage phagocytosis and killing of *Streptococcus pneumoniae*. *American Journal of Respiratory Cell and Molecular Biology*. 2007;36(4):460-465.
- (423) Knox EG. Atmospheric pollutants and mortalities in English local authority areas. *Journal of Epidemiology and Community Health*. 2008;62(5):442-447.
- (424) Torres A, Cilloniz C. *Clinical Management of Bacterial Pneumonia*. Springer; 2015.
- (425) Torres A, Peetermans WE, Viegi G, Blasi F. Risk factors for community-acquired pneumonia in adults in Europe: a literature review. *Thorax*. 2013;68(11):1057-1065.
- (426) Barnett AG, Williams GM, Schwartz J et al. Air pollution and child respiratory health - A case-crossover study in Australia and new Zealand. *American Journal of Respiratory and Critical Care Medicine*. 2005;171(11):1272-1278.
- (427) Cheng MF, Tsai SS, Chiu HF, Sung FC, Wu TN, Yang CY. Air pollution and hospital admissions for pneumonia: Are there potentially sensitive groups? *Inhalation Toxicology*. 2009;21(13):1092-1098.
- (428) Medina-Ramon M, Zanobetti A, Schwartz J. The effect of ozone and PM10 on hospital admissions for pneumonia and chronic obstructive pulmonary disease: a national multicity study. *Am J Epidemiol*. 2006;163(6):579-588.
- (429) Lai HK, Tsang H, Wong CM. Meta-analysis of adverse health effects due to air pollution in Chinese populations. *Bmc Public Health*. 2013;13.
- (430) Hoek G, Brunekreef B, Verhoeff A, van Wijnen J, Fischer P. Daily mortality and air pollution in the Netherlands. *Journal of the Air & Waste Management Association*. 2000;50(8):1380-1389.
- (431) Mandell LA, Wunderink RG, Waterer GW. Community-acquired pneumonia. *N Engl J Med*. 2015;372(3):293-294.
- (432) Trotter CL, Stuart JM, George R, Miller E. Increasing hospital admissions for pneumonia, England. *Emerging Infectious Diseases*. 2008;14(5):727-733.
- (433) Guest JF, Morris A. Community-acquired pneumonia: The annual cost to the national health service in the UK. *European Respiratory Journal*. 1997;10(7):1530-1534.
- (434) Scapellato ML, Lotti M. Short-term effects of particulate matter: An inflammatory mechanism? *Critical Reviews in Toxicology*. 2007;37(6):461-487.

- (435) van der Zee SC, Hoek G, Boezen MH, Schouten JP, van Wijnen JH, Brunekreef B. Acute effects of air pollution on respiratory health of 50-70 yr old adults. *European Respiratory Journal*. 2000;15(4):700-709.
- (436) Peacock JL, Anderson HR, Bremner SA et al. Outdoor air pollution and respiratory health in patients with COPD. *Thorax*. 2011;66(7):591-596.
- (437) Lagorio S, Forastiere F, Pistelli R et al. Air pollution and lung function among susceptible adult subjects: A panel study. *Epidemiology*. 2004;15(4):S45-S46.
- (438) Desqueyroux H, Pujet JC, Prosper M, Le Moullec Y, Momas I. Effects of air pollution on adults with chronic obstructive pulmonary disease. *Archives of Environmental Health*. 2002;57(6):554-560.
- (439) Rice MB, Ljungman PL, Wilker EH et al. Short-Term Exposure to Air Pollution and Lung Function in the Framingham Heart Study. *American Journal of Respiratory and Critical Care Medicine*. 2013;188(11):1351-1357.
- (440) Decramer M, Janssens W, Miravittles M. Chronic obstructive pulmonary disease. *Lancet*. 2012;379(9823):1341-1351.
- (441) Ko FWS, Hui DSC. Air pollution and chronic obstructive pulmonary disease. *Respirology*. 2012;17(3):395-401.
- (442) Kelly FJ, Fussell JC. Air pollution and airway disease. *Clinical and Experimental Allergy*. 2011;41(8):1059-1071.
- (443) Papi A, Bellettato CM, Braccioni F et al. Infections and airway inflammation in chronic obstructive pulmonary disease severe exacerbations. *American Journal of Respiratory and Critical Care Medicine*. 2006;173(10):1114-1121.
- (444) Sunyer J, Basagana X. Particles, and not gases, are associated with the risk of death in patients with chronic obstructive pulmonary disease. *International Journal of Epidemiology*. 2001;30(5):1138-1140.
- (445) Moolgavkar SH. Air pollution and hospital admissions for chronic obstructive pulmonary disease in three metropolitan areas in the United States. *Inhalation Toxicology*. 2000;12:75-90.
- (446) Samoli E, Stafoggia M, Rodopoulou S et al. Which specific causes of death are associated with short term exposure to fine and coarse particles in Southern Europe ? Results from the MED-PARTICLES project. *Environment International*. 2014;67:54-61.
- (447) NICE. Chronic obstructive pulmonary disease: Costing report - Implementing NICE guidance. February, 2011. Available at: <https://www.nice.org.uk/guidance/cg101/resources/costing-report-134511805>. Accessed March 20, 2016.
- (448) Viegi G, Maio S, Pistelli F, Baldacci S, Carrozzi L. Epidemiology of chronic obstructive pulmonary disease: Health effects of air pollution. *Respirology*. 2006;11(5):523-532.
- (449) Gaze DC, InTech Open AB. *Ischemic Heart Disease*. S.l.: InTech; 2013.
- (450) Liao DP, Duan YK, Whitsel EA et al. Association of higher levels of ambient criteria pollutants with impaired cardiac autonomic control: A population-based study. *American Journal of Epidemiology*. 2004;159(8):768-777.

- (451) Liao DP, Heiss G, Chinchilli VM et al. Association of criteria pollutants with plasma hemostatic/inflammatory markers: a population-based study. *Journal of Exposure Analysis and Environmental Epidemiology*. 2005;15(4):319-328.
- (452) Park SK, O'Neill MS, Vokonas PS, Sparrow D, Schwartz J. Effects of air pollution on heart rate variability: The VA Normative Aging Study. *Environmental Health Perspectives*. 2005;113(3):304-309.
- (453) Brook RD, Urch B, Dvonch JT et al. Insights Into the Mechanisms and Mediators of the Effects of Air Pollution Exposure on Blood Pressure and Vascular Function in Healthy Humans. *Hypertension*. 2009;54(3):659-667.
- (454) Chuang KJ, Chan CC, Su TC, Lee CT, Tang CS. The effect of urban air pollution on inflammation, oxidative stress, coagulation, and autonomic dysfunction in young adults. *American Journal of Respiratory and Critical Care Medicine*. 2007;176(4):370-376.
- (455) Franchini M, Mannucci PM. Particulate Air Pollution and Cardiovascular Risk: Short-term and Long-term Effects. *Seminars in Thrombosis and Hemostasis*. 2009;35(7):665-670.
- (456) Seaton A, Soutar A, Crawford V et al. Particulate air pollution and the blood. *Thorax*. 1999;54(11):1027-1032.
- (457) Ghio AJ, Hall A, Bassett MA, Cascio WE, Devlin RB. Exposure to concentrated ambient air particles alters hematologic indices in humans. *Inhalation Toxicology*. 2003;15(14):1465-1478.
- (458) Peters A, Doring A, Wichmann HE, Koenig W. Increased plasma viscosity during an air pollution episode: A link to mortality? *Lancet*. 1997;349(9065):1582-1587.
- (459) Peters A, Frohlich M, Doring A et al. Particulate air pollution is associated with an acute phase response in men - Results from the MONICA-Augsburg Study. *European Heart Journal*. 2001;22(14):1198-1204.
- (460) Bhaskaran K, Hajat S, Armstrong B et al. The effects of hourly differences in air pollution on the risk of myocardial infarction: case crossover analysis of the MINAP database. *British Medical Journal*. 2011;343.
- (461) Atkinson RW, Kang S, Anderson HR, Mills IC, Walton HA. Epidemiological time series studies of PM_{2.5} and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax*. 2014;69(7):660-665.
- (462) Zeger SL, Thomas D, Dominici F et al. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental Health Perspectives*. 2000;108(5):419-426.
- (463) Fischer PH, Hoek G, van Reeuwijk H et al. Traffic-related differences in outdoor and indoor concentrations of particles and volatile organic compounds in Amsterdam. *Atmospheric Environment*. 2000;34(22):3713-3722.
- (464) Monn C. Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone. *Atmospheric Environment*. 2001;35(1):1-32.
- (465) Sajani SZ, Scotto F, Lauriola P, Galassi F, Montanari A. Urban air pollution monitoring and correlation properties between fixed-site stations. *Journal of the Air & Waste Management Association*. 2004;54(10):1236-1241.

- (466) Wilson JG, Kingham S, Sturman AP. Intraurban variations of PM10 air pollution in Christchurch, New Zealand: Implications for epidemiological studies. *Science of the Total Environment*. 2006;367(2-3):559-572.
- (467) Grivas G, Chaloulakou A, Samara C, Spyrellis N. Spatial and temporal variation of PM10 mass concentrations within the Greater Area of Athens, Greece. *Water Air and Soil Pollution*. 2004;158(1):357-371.
- (468) Wilson JG, Kingham S, Pearce J, Sturman AP. A review of intraurban variations in particulate air pollution: Implications for epidemiological research. *Atmospheric Environment*. 2005;39(34):6444-6462.
- (469) Leiden University N. iSPEX-EU 2015. October 20, 2015. Available at: <http://ispex-eu.org/>. Accessed January 22, 2016.
- (470) Snik F, Rietjens JHH, Apituley A et al. Mapping atmospheric aerosols with a citizen science network of smartphone spectropolarimeters. *Geophysical Research Letters*. 2014;41(20):7351-7358.
- (471) Hoek G, Beelen R, de Hoogh K et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*. 2008;42(33):7561-7578.
- (472) de Nazelle A, Seto E, Donaire-Gonzalez D et al. Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environmental Pollution*. 2013;176:92-99.
- (473) Liu HY, Skjetne E, Kobernus M. Mobile phone tracking: in support of modelling traffic-related air pollution contribution to individual exposure and its implications for public health impact assessment. *Environmental Health*. 2013;12.
- (474) Poloniecki JD, Atkinson RW, deLeon AP, Anderson HR. Daily time series for cardiovascular hospital admissions and previous day's air pollution in London, UK. *Occupational and Environmental Medicine*. 1997;54(8):535-540.
- (475) Bremner SA, Anderson HR, Atkinson RW et al. Short term associations between outdoor air pollution and mortality in London 1992-4. *Occupational and Environmental Medicine*. 1999;56(4):237-244.
- (476) Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*. 2010;1(2):97-111.
- (477) Zeng D, Lin DY. On random-effects meta-analysis. *Biometrika*. 2015;102(2):281-294.
- (478) Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*. 2010;140(4):961-970.
- (479) Copetti M, Fontana A, Graziano G et al. Advances in Meta-Analysis: Examples from Internal Medicine to Neurology. *Neuroepidemiology*. 2014;42(1):59-67.
- (480) Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A-Statistics in Society*. 2009;172:137-159.
- (481) Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine*. 2012;31(29):3821-3839.

- (482) Gasparrini A, Guo YM, Hashizume M et al. Temporal Variation in Heat-Mortality Associations: A Multicountry Study. *Environmental Health Perspectives*. 2015;123(11):1200-1207.
- (483) Vicedo-Cabrera AM, Forsberg B, Tobias A et al. Associations of Inter- and Intraday Temperature Change With Mortality. *American Journal of Epidemiology*. 2016;183(4):286-293.
- (484) Guo YM, Barnett AG, Yu WW et al. A Large Change in Temperature between Neighbouring Days Increases the Risk of Mortality. *Plos One*. 2011;6(2).
- (485) Lin HL, Zhang YH, Xu YJ et al. Temperature Changes between Neighboring Days and Mortality in Summer: A Distributed Lag Non-Linear Time Series Analysis. *Plos One*. 2013;8(6).
- (486) Cao JY, Cheng YX, Zhao N et al. Diurnal Temperature Range is a Risk Factor for Coronary Heart Disease Death. *Journal of Epidemiology*. 2009;19(6):328-332.
- (487) Oke TR. City Size and Urban Heat Island. *Atmospheric Environment*. 1973;7(8):769-779.
- (488) Hinchliffe SR, Abrams KR, Lambert PC. The impact of under and over-recording of cancer on death certificates in a competing risks analysis: A simulation study. *Cancer Epidemiology*. 2013;37(1):11-19.
- (489) Nashelsky MB, Lawrence CH. Accuracy of cause of death determination without forensic autopsy examination. *American Journal of Forensic Medicine and Pathology*. 2003;24(4):313-319.
- (490) Smith AE, Hutchins GM. Problems with the proper completion and accuracy of the cause of death statement of the death certificate. *Laboratory Investigation*. 1998;78(1):6A.
- (491) Deckert A. The existence of standard-biased mortality ratios due to death certificate misclassification - a simulation study based on a true story. *Bmc Medical Research Methodology*. 2016;16.
- (492) Harteloh P, de Bruin K, Kardaun J. The reliability of cause-of-death coding in The Netherlands. *European Journal of Epidemiology*. 2010;25(8):531-538.
- (493) Lu TH, Shau WY, Shih TP, Lee MC, Chou MC, Lin CK. Factors associated with errors in death certificate completion: A national study in Taiwan. *Journal of Clinical Epidemiology*. 2001;54(3):232-238.
- (494) Surjan G. Questions on validity of International Classification of Diseases-coded diagnoses. *International Journal of Medical Informatics*. 1999;54(2):77-95.
- (495) Winkler V, Ott JJ, Becher H. Reliability of coding causes of death with ICD-10 in Germany. *International Journal of Public Health*. 2010;55(1):43-48.
- (496) Bell ML, Kim JY, Dominici F. Potential confounding of particulate matter on the short-term association between ozone and mortality in multisite time-series studies. *Environmental Health Perspectives*. 2007;115(11):1591-1595.
- (497) Dominici F, Peng RD, Barr CD, Bell ML. Protecting Human Health From Air Pollution Shifting From a Single-pollutant to a Multipollutant Approach. *Epidemiology*. 2010;21(2):187-194.

- (498) Sun ZC, Tao YB, Li S et al. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health*. 2013;12.
- (499) Mahiyuddin WRW, Sahani M, Aripin R, Latif MT, Thach TQ, Wong CM. Short-term effects of daily air pollution on mortality. *Atmospheric Environment*. 2013;65:69-79.
- (500) Dominici F, McDermott A, Daniels M, Zeger SL, Samet JM. Revised analyses of the National Morbidity, Mortality, and Air Pollution Study: Mortality among residents of 90 cities. *Journal of Toxicology and Environmental Health-Part A-Current Issues*. 2005;68(13-14):1071-1092.
- (501) Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the Health Effects of Exposure to Multi-Pollutant Mixture. *Annals of Epidemiology*. 2012;22(2):126-141.
- (502) Sacks JD, Ito K, Wilson WE, Neas LM. Impact of Covariate Models on the Assessment of the Air Pollution-Mortality Association in a Single- and Multipollutant Context. *American Journal of Epidemiology*. 2012;176(7):622-634.
- (503) Stafoggia M, Samoli E, Alessandrini E et al. Short-term Associations between Fine and Coarse Particulate Matter and Hospitalizations in Southern Europe: Results from the MED-PARTICLES Project. *Environmental Health Perspectives*. 2013;121(9):1026-1033.
- (504) Lanzinger S, Schneider A, Breitner S et al. Associations between ultrafine and fine particles and mortality in five central European cities - Results from the UFIREG study. *Environment International*. 2016;88:44-52.
- (505) Meister K, Johansson C, Forsberg B. Estimated Short-Term Effects of Coarse Particles on Daily Mortality in Stockholm, Sweden. *Environmental Health Perspectives*. 2012;120(3):431-436.
- (506) Ito K, Mathes R, Ross Z, Nadas A, Thurston G, Matte T. Fine Particulate Matter Constituents Associated with Cardiovascular Hospitalizations and Mortality in New York City. *Environmental Health Perspectives*. 2011;119(4):467-473.
- (507) Bell ML, Ebisu K, Peng RD, Samet JM, Dominici F. Hospital Admissions and Chemical Composition of Fine Particle Air Pollution. *American Journal of Respiratory and Critical Care Medicine*. 2009;179(12):1115-1120.
- (508) Ostro B, Feng WY, Broadwin R, Green S, Lipsett M. The effects of components of fine particulate air pollution on mortality in California: Results from CALFINE. *Environmental Health Perspectives*. 2007;115(1):13-19.
- (509) Peng RD, Bell ML, Geyh AS et al. Emergency Admissions for Cardiovascular and Respiratory Diseases and the Chemical Composition of Fine Particle Air Pollution. *Environmental Health Perspectives*. 2009;117(6):957-963.
- (510) Zanobetti A, Franklin M, Koutrakis P, Schwartz J. Fine particulate air pollution and its components in association with cause-specific emergency admissions. *Environmental Health*. 2009;8.
- (511) Zanobetti A, Austin E, Coull BA, Schwartz J, Koutrakis P. Health effects of multi-pollutant profiles. *Environment International*. 2014;71:13-19.
- (512) Mauderly JL, Burnett RT, Castillejos M et al. Is the air pollution health research community prepared to support a multipollutant air quality management framework? *Inhalation Toxicology*. 2010;22:1-19.

- (513) Bender R, Lange S. Adjusting for multiple testing - when and how? *Journal of Clinical Epidemiology*. 2001;54(4):343-349.
- (514) Pocock SJ, Collier TJ, Dandreo KJ et al. Issues in the reporting of epidemiological studies: a survey of recent practice. *British Medical Journal*. 2004;329(7471):883-887.
- (515) Lin HL, An QZ, Luo C, Pun VC, Chan CS, Tian LW. Gaseous air pollution and acute myocardial infarction mortality in Hong Kong: A time-stratified case-crossover study. *Atmospheric Environment*. 2013;76:68-73.
- (516) Chalmers JD, Singanayagam A, Akram AR et al. Severity assessment tools for predicting mortality in hospitalised patients with community-acquired pneumonia. Systematic review and meta-analysis. *Thorax*. 2010;65(10):878-883.
- (517) Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *Journal of Clinical Epidemiology*. 2004;57(12):1288-1294.
- (518) Hosmer DW, Lemeshow S, Cook ED. *Applied logistic regression*. 2nd ed ed. New York ; Chichester : John Wiley & Sons; 2000.
- (519) Beverland IJ, Carder M, Cohen GR, Heal MR, Agius RM. Associations between short/medium-term variations in black smoke air pollution and mortality in the Glasgow conurbation, UK. *Environment International*. 2012;50:1-6.
- (520) Sousa D, Justo I, Dominguez A et al. Community-acquired pneumonia in immunocompromised older patients: incidence, causative organisms and outcome. *Clinical Microbiology and Infection*. 2013;19(2):187-192.

APPENDIX

Appendix A Full description of the data collection procedure

Mortality and corresponding hospital admissions data for several Scottish regions including Edinburgh, Glasgow and Aberdeen between 1981 and 2001 were already possessed by The Centre for Occupational and Environmental Health at the beginning of the study in 2010. One aim was update this Scottish dataset to include as much of the last 10 years as was available. Even though the data does contain 'sector' level postcode no information directly identifies a subject. However, information such as date of death, cause of death, age and gender was also present that if combined would allow a subject to be identifiable in some cases. These subjects have died in order to enter the study meaning consent is not possible and it is not possible to negatively influence any subject in the study, a duty of care regarding data protection is still an important concern.

In order to acquire an updated dataset for Scotland, The Information Services Division (ISD) Scotland was approached who act as guardians for Scottish mortality and hospital admissions data and provided the primary dataset. To gain approval from ISD Scotland initially ethical approval from the NHS Research Ethics Committee (REC) was obtained, and with respect to the data protection approval from the Scottish based Privacy Advisory Committee (PAC) was also obtained. The NHS REC form was submitted, the meeting was held, and a letter of approval received in 27th November 2011 with some minor conditions. Subsequently the PAC form was completed and submitted on the 14th December 2011, a revised version was submitted on the 14th January and approval received in May 2012 with the full data arriving in June 2012.

One early aim of the study was to compare 'immediate' cause of death with 'underlying' cause of death. ISD Scotland indicated that it was not able to identify the 'immediate' cause of death once the coding procedure has been implemented. The Office of National Statistics (ONS) indicated that it was able to provide a 'crude' earlier version of the dataset in which it was identified. It was therefore decided to apply to acquire a dataset from ONS relating to three English cities. In addition to approval from NHS REC, approval from The National Information Governance Board now known as the Confidentiality Advisory Group (CAG) was required with regards to section 251 of

the data protection act being waved. This relates to access to data of those not able to consent, in this case those who have previously died. To receive full approval from CAG confirmation is required that the Health and Social Care Information Centre (HSCIC) approve of the request. Once HSCIC approve the request and so long as the University of Manchester has the appropriate Information Governance tool kit and data protection procedures in place CAG will approve the request. The application is then passed to ONS who will require the investigatory team to become ONS approved researchers, gain approval from a micro data release panel. ONS will then send the mortality data to HSCIC to be linked with the Health Episode Statistics who provide us with the data.

After some confusion between the University of Manchester's Research coordinators office and HSCIC regarding a 'Caldicott Guardian' letter required in the CAG approval, the CAG application was submitted in December 2012 and provisional acceptance received in May 2013. During this period a full set of forms were submitted to HSCIC and ONS regarding the data request. Before full approval CAG requested confirmation that HSCIC were willing to supply us with the data and that the University of Manchester possessed the appropriate Information Governance tool kit rating. Due to the time delay and a change in the application forms, a second application was submitted to HSCIC in June 2013 at which point a number of queries were answered. In November 2013 HSCIC delayed the request due to an issue with the mortality data request which was cleared with ONS. Who had also given approval as an ONS approved researcher and access to the mortality data, subsequent to HSCIC linking with the corresponding Hospital Episodes Statistics (HES). As of Jan 2014, HSCIC were requesting full CAG approval which was granted in Feb 2014. At this stage it was indicated that HSCIC approval would take several months and so no contact was received until in June 2014 when HSCIC announced they had stopped approval of all new projects while they perform a comprehensive review of all current data sharing contracts. I believe this was in response to the reports of a leak of personal information to an online mapping agency called Earthware and uploading the HES data to google (cloud) servers located off shore; as reported by the Guardian in March 2014. HSCIC further announced in Nov 2014 that all applications prior to Jan 2014 would be put to the bottom of the pile and new applications would be dealt with first. Contact from

HSCIC was received the following July 2015, where a further set of queries were answered and all the information governance forms and section 251 approval from CAG were renewed as they had expired. It was also reported to us that the ONS approved researcher status was due to expire, at which point we accepted HSCIC offer to confirm to ONS that it be renewed. No further contact was received until December 2015 when, a HSCIC 'case' officer approached us stating that a newly revised application would be required in order to take our application to DAAG (HSCICs internal Data Access Advisory Group). Note, we were informed at this point that the ONS approved researcher status was not approved (as it wasn't requested by the contact at HSCIC) and so the application would need to be reapplied, a process HSCIC informed us was notoriously slow and may take several months. The first attempt to gain approval from DAAG, was rejected on the basis that the project did not adequately show how it was going to disseminate the information in a health and social care setting. After revising the application, a second attempt was planned in January 2016. Note, Section 251 approval from CAG was about to expire for a second time, and so the relevant request for renewal was made. In February 2016, HSCIC requested the latest information governance status for the University as the status had expired for the second time. At which point they also informed us that their criteria regarding storage and usage of the data had changed and a two page check list arrived. It would appear that these new criteria require a major overhaul of the already highly secure University of Manchester storage system (including complete encryption of the entire data storage disks for the entire University), which had recently been put in place to, in part, appease HSCIC. A current optimistic estimate regarding time till implementation of this new system is 8 months. An attempt is currently being made to use a temporary secure storage system already located within another university department. However, this has not been finalised with HSCIC. Until it has the application is on hold indefinitely.

In short, both the Scottish and English applications were started at roughly the same time. The Scottish application took just over 12 months from to physically get hold of the data. The English data has still not been acquired nearly 6 years later, and at this stage the date of completion is unknown.

Appendix B Published original research paper

Matthew Gittins, Roseanne McNamee, Melanie Carder, Iain Beverland, Raymond Agius 2013 ‘Has the short-term effect of Black Smoke exposure on pneumonia mortality been underestimated because hospitalisation is ignored? A case-crossover study.’ *Environmental Health*. DOI: 10.1186/1476-069X-12-97

Acknowledgements

The paper was written by the author of the thesis, Matthew Gittins. In addition Matthew Gittins, modified the concept, created and manipulated the dataset, performed the statistical analysis, and drafted the paper. Roseanne McNamee, Melanie Carder and Raymond Agius recruited the original dataset, provided advice on the concept, decisions regarding data choices and the statistical analysis. Iain Beverland provided useful advice regarding the discussion. All authors reviewed and provided comments for general improvement of the paper.

Co-authors:

Roseanne McNamee

Centre for Biostatistics, Institute of Population Health University of Manchester

Melanie Carder

Centre for Occupational and Environmental Health, Centre for Epidemiology, Institute of population Health, The University of Manchester, Manchester UK

Iain Beverland

Department of Civil Engineering, University of Strathclyde, Glasgow, UK

Raymond M Agius

Centre for Occupational and Environmental Health, Centre for Epidemiology, Institute of population Health, The University of Manchester, Manchester UK

Title: - Has the short-term effect of Black Smoke exposure on pneumonia mortality been underestimated because hospitalisation is ignored? A case-crossover study.

Abstract

Background

Short-term associations have been demonstrated between air pollution and respiratory mortality including pneumonia. Studies typically estimate exposure based only on place of residence, yet many are in hospital prior to death. This study investigates lag length and tests the hypothesis that the effect of 'black smoke' is greater when restricted to pneumonia deaths in the community – Community deaths from pneumonia.

Methods

A time-stratified case-crossover design using conditional logistic regression estimated the daily percentage increase in risk of pneumonia mortality in relation to 'black smoke' in the preceding 30 days. Cases were pneumonia deaths in Edinburgh 1981-1996. Multiple 'control' periods, were defined using the same weekdays for the same month as the case death. Lag structure was investigated by a stratified lag model with five 6-day periods and by distributed lag models. Hospital admissions data, defined a community death as someone who had not been in hospital in any of the 30 days before death.

Results

Of 14,346 subjects who died from pneumonia, 7,536 were community deaths. Larger estimated increases in risks were seen in the community for all lag periods. Both stratified and distributed lag methods suggested positive effect estimates for 18 days after exposure and negative thereafter; the average percent increase per day across the 18 days was 0.70% (95% C.I. 0.29-1.14) for community subjects and 0.30% (95% C.I. 0.03-0.59) for all subjects.

Conclusions

Studies which fail to account for hospitalisation may underestimate exposure effects as stronger pollution effects on mortality were evident in community based subjects.

Introduction & Background

Many studies worldwide have demonstrated an association between air pollution and all-cause mortality^{39,78}, specifically respiratory mortality^{7,108,109}. In 2011, pneumonia was the 6th leading cause of death in England and Wales for males (10,824) and 4th (14,872) for females¹¹¹. However, only a few studies - with limited findings - have specifically investigated associations between pneumonia related deaths and ambient air pollution. Schwartz and Dockery, indicated an increase in pneumonia mortality of 11% (95% C.I -3%,27%) per 100 μgm^{-3} increase of Total Suspended Particles (TSP)¹⁹⁵. Halonen et al. demonstrated a percentage increase in pneumonia mortality in Finland of 3.16% (95% C.I -2.64%,9.32%) per increase in interquartile range of a 5 day Coarse Particle Matter ($\text{PM}_{10-2.5}$) mean¹²³. Zanobetti et al. proposed that air pollution may be a predisposing factor to community acquired pneumonia (CAP) and that subjects with CAP rather than hospital acquired pneumonia, may be more susceptible to the effects of air pollution¹²⁴. Studies such as Neupane et al. have indicated a relationship between long-term exposure to air pollution and emergency visits to hospital with community acquired pneumonia¹²⁵. However, so far no study has investigated the effect of pollution on deaths from community acquired pneumonia only.

Much of the evidence for the association between air pollution and general mortality has focused on exposure in a short time period – less than 40 days prior to death^{39,79}. This focus on short to medium exposure may be appropriate for pneumonia which is generally an acute condition, though often associated with chronic underlying lung disease. Frequently, a deceased individual's exposure is inferred from information regarding their place of residence with little or no attempt to take account of subject's actual location, circumstances or activities. Exposure is typically assumed to be the same for all subjects living within a given distance from a single pollution monitor or an average of multiple monitors within the area^{78,79}. More recent studies, trying to improve exposure estimates, have taken into account traffic density and other geographical information regarding the subjects neighbourhood or city^{56,80}. The presumption still remained that the deceased was in the geographical location of residence during the exposure period; in fact it is common for people to die in non-residential locations

(65.3% in a NHS hospital/Hospice in England and Wales ⁸¹). If the hospital is located close to the place of residence, one might reasonably suppose that a patients' exposure to outdoor pollution might be reduced when confined indoors ⁸². Epidemiological observations have shown that deaths associated with air pollution, specifically TSP and Particle Matter with a diameter less than 10µm (PM₁₀), are disproportionately increased outside of hospital ^{83,84}. In addition, Jansen et al. 2002, found that the health effects of PM₁₀ on cardiovascular disease and COPD in 14 U.S. cities decreased significantly as the proportion of homes with air conditioning increased ⁸⁵. Previous attempts at comparing risks in and out of hospital, such as Téllez-Rojo et al. 2001 and Zeka et al. 2006, have shown significant increased risk of death from respiratory or cardiovascular causes, in some cases up to a threefold increase. These studies have primarily used location at time of death without confirming location during exposure ^{86,87}. Failure to take account of hospitalisation during exposure could lead to further effect underestimation if a substantial fraction of the population experience reduced exposure in air-conditioned hospitals. A large proportion of subjects hospitalised during exposure might explain why some observational epidemiology studies based on routinely collected data may struggle to replicate previously demonstrated associations between pollution and pneumonia caused mortality ⁸⁸.

Pneumonia occurs usually as a result of bacterial or viral infection. Often progressing rapidly within 24 hours, it presents symptoms such as coughing, chest pain, shortness of breath, and fever. Pneumonia can generally be diagnosed reliably through medical consultation and a chest radiography ¹²⁰. The relatively quick onset of the disease, short diagnosis period, and the time varying nature of air pollution exposure satisfies the conditions for employing the case-crossover design ¹²¹. This design offers protection against possible subject level confounders without the need for complex modelling of mortality levels over time.

This study investigated the effect of 'black smoke' (BS) over 30 days prior to pneumonia mortality using a time-stratified case-crossover design ⁹. We tested the primary hypothesis that estimated association would be greater in subjects who spent the exposure period in the 'community' (i.e. not in hospital) compared to those who spent some or all of the period in hospital. Members of the former group will be defined

as subjects with a community death from pneumonia (CDPs) which should be distinguished from CAP; CAP refers to a clinical category based on the source of pneumonia, CDP are a subgroup of CAP. Concurrently, hypothesis generating analyses explored how exposure affects mortality from pneumonia across the lag period. Subsequently, analyses based on the lag periods which showed effects on mortality were repeated for subgroups defined by gender and age. It is conceivable that differences in lifestyle between these groups could influence exposure or that there is increased susceptibility in older age, leading to differential observed effects of BS on CDPs.

Methods

Deaths due to pneumonia (ICD-9 codes 480-487 pre 2000) between January 1981 and December 1996 from Edinburgh, Scotland formed the cases. Separately, information was provided on all admissions to hospital caused by respiratory, cardiovascular, lung cancer, diabetes, and digestive related causes for the same time period and location. The Information Services Division of NHS National Services in Scotland provided both datasets, they included; date of death, age, gender, admission dates, primary and secondary cause of admission and death, and if the patient died in hospital. The two files – one of deaths and one of hospital admissions – were linked to determine if the subject had been in hospital during the 30 days prior to death. Community acquired pneumonia refers to those subjects who did not acquire pneumonia from a hospital. Subjects with CAP may subsequently enter and later die in hospital. These subjects will not be included in our data as they will have had at least one day in hospital altering their exposure and increasing their chances of a secondary hospital acquired pneumonia infection. A death was considered to be a community death from pneumonia if the hospital admission data showed that the subject had not been in hospital for any of the 30 days prior to death. A CDP subject is therefore a special subgroup of the CAP deaths.

Daily black smoke air pollution data were obtained from one centrally located background monitoring site and hourly ambient air temperature (between 7am-11pm) was provided by the Scottish Meteorological Office. This was used to give daytime mean temperature and mean pollution levels for the area. For each date of death, or case day, ‘black smoke’ daily results for the month prior were averaged firstly across 1-30 days and then separately for 1-6, 7-12, 13-18, 19-24, and 25-30 days. These formed the exposure variables for the cases. Same day exposure (lag 0) was not included as pneumonia has a minimum 24hr incubation period,¹²⁰ and same day exposure potentially includes exposure after death. Temperature displays temporal associations with pollution levels, and has previously been shown in this population to have an approximately double linear relationship with mortality, with a knot at 11°C¹³. Hence, two continuous temperature variables were calculated as, “high” ($t-11$ if $\geq 11^\circ\text{C}$, 0 otherwise) and “low” ($t-11$ if $\leq 11^\circ\text{C}$, 0 otherwise), where t is the daytime mean temperature. Average temperature across lags 1–30 days for both “low” and “high”

variables was included in all models. Including pollution and temperature exposure lags (up to 30 days prior) reduces the chances of underestimating the exposure effect ⁷⁹. Further information regarding the data source and variable manipulations can be found elsewhere ¹³.

The 30 days prior to the date of death from pneumonia defined the case exposure period. In case-crossover designs, reduced bias allows for a time-stratified design to select matched control periods ²²³. Control days were defined as all equivalent days of the week within the same month as the case day, in order to account for any weather, seasonal, or day of the week confounding ²³¹. For example, if the event occurred on the second Monday of May then the control days became all other Mondays of May. The control exposure period was then 30 days prior to the control day, and temperature and pollution variables were formed from these periods.

Statistical analysis

Conditional logistic regression ⁵¹⁸, initially compared average exposure between the case and controls over 30 days prior to death (Model 1). To investigate exposure during the 30 days, a lag-stratified model fitted five exposure variables each representing a 6 day lag period (Model 2). To avoid problems comparing estimates based on different period lengths (e.g. 30 days and 6 days) ^{13,519}, all estimated coefficient were divided by the number of days on which the mean was based. Hence, BS results are expressed as percentage increase in risk associated with an increase of $10\mu\text{gm}^{-3}$ $((e^{B*10}-1)*100$, where B = model coefficient) on any individual day within the lag period. Similarly, both continuous temperature variables are also expressed as a percentage increase in risk corresponding to a *decrease in 1°C of temperature* for any individual day during lag period $((e^{-B}-1)*100)$. The association between BS and All pneumonia deaths (AP) was estimated before restricting to the subgroup (CDP) deemed to have the greatest potential exposure. To test if a significant difference in exposure effects occurred, an interaction term was included to compare CDP and non-CDP for each BS lag term, and the Log-likelihood-ratio was used to test the difference.

Within each lag period the average daily value across the lag period was calculated so long as a minimum of four out of every six days contained a pollution estimate. The

analysis was then performed with any subject with complete data for the case day and at least one control day. This resulted in 4.5% and 4.3% (AP and CDP, respectively) of subjects dropped due to either missing pollution data in the case day or all control days.

The effect of a change in exposure on an individual day might be expected to vary across subsequent days and eventually fall to zero¹². An *estimated* effect might even become negative afterwards if the ‘high risk’ pool of subjects is depleted without sufficient replenishment, causing a mortality rate lower than the underlying rate⁸⁴. In addition to lag-stratified, a quadratic distributed lag¹² estimated the lag time, L, before which the estimated effects are positive. For simplicity in further analysis, average exposure across the period (0,L) days was modelled for each gender and two age groups (≤ 80 and >80). Analysis was performed using STATA version 11³³⁰.

Results

A total of 15,784 people had pneumonia as primary or secondary cause of death in Edinburgh between 1981 and 1996. However, missing pollution data (9.8% of days during the time period) for either a subjects' case day or all of the subjects' control days meant only 14,346 cases with 47,431 control days were eligible for the analysis. Table 1 gives descriptive statistics for the daily average BS air pollution, air temperature and demographic characteristics for both all pneumonia (AP) and CDP (52.5% of the AP) subject groups. Further, summary statistics indicated an interquartile range approximately $10 \mu\text{gm}^{-3}$ for each lag period.

Table 1 – Descriptive statistics of exposure data and study subjects split by All Pneumonia and Community Deaths from Pneumonia.

	Mean	S.D	Median	IQR	Min	Max
Daily Ave Air Temp($^{\circ}\text{C}$)	9.4	5.1	9.4	8	-12.7	24.48
Daily Ave BS(μgm^{-3})	12.7	13.3	9	10	1	194
Lag 1-6 dys Ave BS(μgm^{-3})	12.9	10.8	9.3	9.5	1	95.2
Lag 7-12 dys Ave BS(μgm^{-3})	13	11.1	9.5	9.6	1	95.2
Lag 13-18 dys Ave BS(μgm^{-3})	13.1	11.3	9.5	9.7	1	95.2
Lag 19-24 dys Ave BS(μgm^{-3})	13.1	11.2	9.5	9.5	1	95.2
Lag 25-30 dys Ave BS(μgm^{-3})	13	11.2	9.5	9.3	1	95.2
Lag 1-30 dys Ave BS(μgm^{-3})	13.2	9.2	10.1	9.3	2.7	73.2
Age (CDP)	79.13	12.6	82	12	0	108
Age (Non-CDP)	79.21	11.92	81	13	0	108
Age (AP)	79.15	12.2	81	13	0	108
	Gender		Age Grouped			
Categories	Male	Female	<80	≥ 80	Total	
CDP only Subjects (%)	3409(45.2)	4127(54.8)	3064(40.7)	4,472(59.3)	7536(52.5)	
Non-CDP only Subjects (%)	3166(48.2)	3644(46.9)	3109(50.4)	3701(45.3)	6810(47.5)	
All Pneumonia Subjects (%)	6575(45.8)	7771(54.2)	6173(43.0)	8,173(57.0)	14346(100)	

BS = Black Smoke

The percentage change in relative risk (%RR), with 95% confidence intervals, are given in Table 2 for an increase in BS of $10 \mu\text{gm}^{-3}$ or a *temperature decrease* of 1°C on any individual day. MODEL 1 and 2 refers to the model with BS and temperature averaged over 1-30 days and BS split into five smaller lag periods, respectively. Correlation coefficients indicate strong correlation between average exposures in adjacent lag

periods (≈ 0.7) however the corresponding variance inflation factors (VIF = 2.01 to 2.78) did not indicate the presence of collinearity between the five lag periods in model 2³³⁹. To easily compare effect sizes between lag periods the percentage change in relative risk corresponds to the effect of a change in BS or temperature on any individual day within the associated lag period. The differences in %RR between AP and CDP along with corresponding significance levels are also given.

MODEL 1 considers the effects of exposure on each of the 30 days to be equal. An increase of $10 \mu\text{g m}^{-3}$ black smoke on any of the 30 days, showed a small rise in AP relative risk increasing to 0.19% in the CDP group, resulting from a %RR difference of -0.18% between CDP and non-CDP subjects (CDP-non-CDP %RR). When the 30 days is split into 5 lag periods (MODEL 2), the magnitude of the effect is always larger in the CDP subjects, of whom the largest changes in %RR are seen in the 1-6, 7-12, and 13-18 day lags. This 18 day period prior to death appeared to be the high risk period, as an increase %RR is observed in 1-6, 7-12, and 13-18 day lags whereas a decrease is observed in the 19-24 and 25-30 day lags. Figure 1 plots the change in log rate ratio associated with the 30 day lag period for both AP and CDP as modelled using the quadratic lag distribution model. As suggested in Table 2, the CDP group showed larger effects with a more rapid decline crossing zero at approximately 21 days, almost 2 days earlier than the more gradual AP decline in risk.

Table 2 –Percent change in risk for lagged black smoke air pollution and pneumonia mortality: repeated for AP(All Pneumonia), CDP (Community Death from Pneumonia) & Non-CDP.

	Lag (days)	AP		CDP only		Non - CDP only		CDP - AP Diff	CDP-Non CDP Diff	P-val
		% RR Change	95% C.I	% RR Change	95% C.I	% RR Change	95% C.I			
MODEL 1 Black Smoke	1-30	0.08%	-0.17%,0.35%	0.19%	-0.16%,0.58%	0.01%	-0.34%,0.40%	0.11%	0.18%	0.285
Air Temp "Low"	1-30	0.20%	0.11%,0.29%	0.22%	0.09%,0.35%	0.16%	0.03%,0.30%	0.02%	0.06%	0.502
Air Temp "High"	1-30	-0.05%	-0.20%,0.10%	-0.20%	-0.40%,0.00%	0.11%	-0.11%,0.34%	0.15%	0.31%	0.059
MODEL 2 Black Smoke	1-6	0.12%	-0.37%,0.62%	0.56%	-0.14%,1.29%	-0.31%	-0.99%,0.41%	0.44%	0.87%	0.022
	7-12	0.05%	-0.42%,0.53%	0.32%	-0.33%,1.00%	-0.25%	-0.92%,0.45%	0.28%	0.57%	0.023
	13-18	0.40%	-0.08%,0.90%	0.71%	0.03%,1.42%	0.14%	-0.56%,0.86%	0.31%	0.57%	0.163
	19-24	-0.09%	-0.55%,0.38%	-0.16%	-0.79%,0.50%	0.05%	-0.63%,0.75%	0.06%	0.21%	0.272
	25-30	-0.11%	-0.57%,0.36%	-0.39%	-1.01%,0.25%	0.34%	-0.35%,1.06%	0.28%	0.73%	0.008
Air Temp "Low"	1-30	0.19%	0.10%,0.29%	0.19%	0.06%,0.32%	0.18%	0.05%,0.32%	0.00%	0.01%	0.487
Air Temp "High"	1-30	-0.05%	-0.19%,0.10%	-0.20%	-0.39%,0.01%	0.09%	-0.12%,0.33%	0.15%	0.29%	0.058

%RR Change - percentage change in Relative Risk, associated with an increase of $10\mu\text{gm}^{-3}$ BS or a decrease of 1°C , on any individual day within the lag period, with corresponding 95% Confidence Interval (95% C.I.)

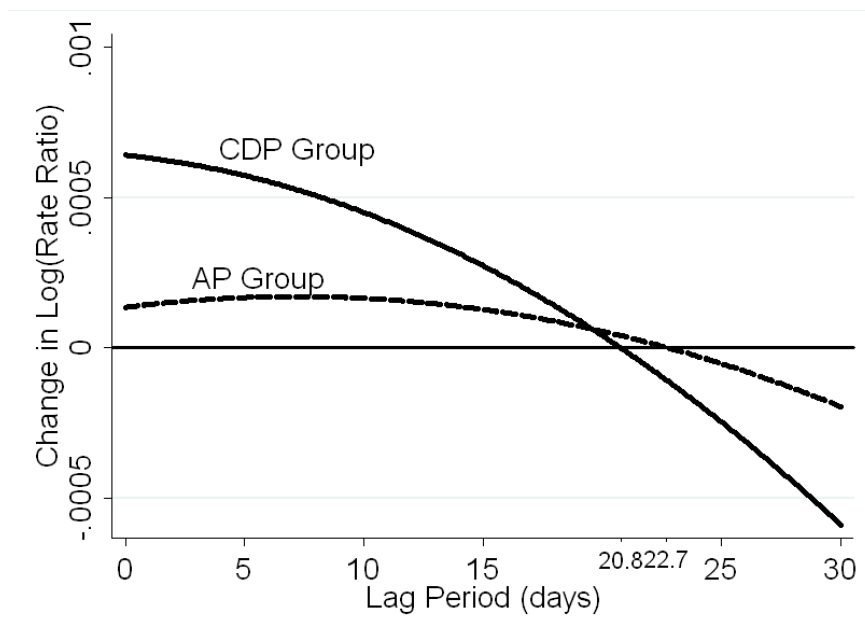
Model 1 - One 30 day lag, Model 2 - The 30 days split into 5 lags of 6 days each fitted simultaneously

|CDP| – |AP| Diff - The difference in the magnitude of the effect size between AP and CDP (CDP-AP)

|CDP-Non CDP| Diff - The difference in the effect size between CDP and Non-CDP (CDP – Non CDP)

Low Temperature effects changed very little, a 1°C decrease corresponded to an increase in relative risk approximately the same for both AP and CDP. In comparison, a 1°C decrease in high temperature shows a small decrease in risk in AP that increases in magnitude in the CDP group.

Figure 1 – The quadratic lag distribution model for subjects with all pneumonia and community deaths from pneumonia.



A secondary analysis concentrated on the minimum lag period where increased risk was observed (18 days) and split the data into the subgroups; gender (male/female), and age (<80/≥80). In all but low temperature for males & age<80 an increase in %RR was seen in the CDP group. The CDP subjects indicated a stronger %RR in males (0.83% - 95% C.I. 0.21%,1.51%) compared to females (0.61% - 95% C.I. 0.08%,1.19%). Subjects aged 80 or above showed larger relative risk in AP (0.37% - 95% C.I. 0.00%,0.77% to 0.23% - 95% C.I. -0.17%,0.66% in <80) but the difference disappeared when restricted to CDP subjects (0.71% - 95% C.I. 0.18%,1.30% and 0.71% - 95% C.I. 0.10%,1.38%, respectively). Analysis was repeated for subjects aged less than 65 to investigate different age distributions of pneumonia subtypes. Even though the effect sizes showed slightly larger differences in magnitude compared to the less than 80 group they were highly imprecise due to substantially smaller sample size, hence they have not been

reported here. Comparisons using the log-likelihood test were performed between the gender and age groups; in all cases the difference was non-significant at the 5% level. Further information can be found in the supplementary Table S1.

Table S1 – Subgroup analysis of 18 day exposure split by subject characteristics; split by AP, CDP & Non-CDP

		AP			CDP only		Non-CDP Only		CDP – AP Diff	CDP-Non CDP Diff	p-val
		Lag (days)	% RR Change	95% C.I	% RR Change	95% C.I	% RR Change	95% C.I			
Overall	Black Smoke	1-18	0.30%	0.03%,0.59%	0.70%	0.29%,1.14%	-0.10%	-0.46%,0.30%	0.40%	0.80%	0.001
	Air Temp “Low”	1-18	0.14%	0.03%,0.25%	0.14%	0.00%,0.29%	0.13%	-0.03%,0.29%	0.00%	0.01%	0.141
	Air Temp “High”	1-18	-0.08%	-0.26%,0.11%	-0.20%	-0.45%,0.06%	0.05%	-0.22%,0.33%	0.12%	0.25%	0.221
Male	Black Smoke	1-18	0.32%	-0.09%,0.75%	0.83%	0.21%,1.51%	-0.14%	-0.66%,0.45%	0.51%	0.97%	0.015
	Air Temp “Low”	1-18	0.18%	0.02%,0.34%	0.16%	-0.05%,0.38%	0.18%	-0.05%,0.42%	0.02%	0.02%	0.442
	Air Temp “High”	1-18	-0.15%	-0.41%,0.13%	-0.40%	-0.76%,-0.01%	0.12%	-0.28%,0.54%	0.25%	0.52%	0.067
Female	Black Smoke	1-18	0.30%	-0.07%,0.69%	0.61%	0.08%,1.19%	-0.07%	-0.56%,0.47%	0.32%	0.68%	0.029
	Air Temp “Low”	1-18	0.11%	-0.03%,0.25%	0.12%	-0.07%,0.32%	0.08%	-0.13%,0.29%	0.02%	0.04%	0.188
	Air Temp “High”	1-18	-0.02%	-0.27%,0.24%	-0.04%	-0.38%,0.33%	-0.01%	-0.37%,0.37%	0.02%	0.03%	0.992
Age <80	Black Smoke	1-18	0.23%	-0.17%,0.66%	0.71%	0.10%,1.38%	-0.19%	-0.71%,0.38%	0.47%	0.90%	0.058
	Air Temp “Low”	1-18	0.11%	-0.06%,0.27%	0.03%	-0.19%,0.26%	0.17%	-0.06%,0.41%	0.08%	0.14%	0.929
	Air Temp “High”	1-18	-0.05%	-0.32%,0.24%	-0.30%	-0.67%,0.11%	0.19%	-0.20%,0.61%	0.25%	0.49%	0.065
Age ≥80	Black Smoke	1-18	0.37%	0.00%,0.77%	0.71%	0.18%,1.30%	-0.02%	-0.52%,0.53%	0.34%	0.73%	0.009
	Air Temp “Low”	1-18	0.16%	0.03%,0.31%	0.22%	0.03%,0.41%	0.09%	-0.11%,0.31%	0.05%	0.13%	0.005
	Air Temp “High”	1-18	-0.10%	-0.34%,0.16%	-0.13%	-0.45%,0.22%	-0.09%	-0.45%,0.30%	0.03%	0.04%	0.963

%RR Change - percentage change in Relative Risk, associated with an increase of $10\mu\text{g}\text{m}^{-3}$ BS or a decrease of 1°C , on any individual day within the lag period, with corresponding 95% Confidence Interval (95% C.I.)

Model 1 - One 30 day lag, Model 2 - The 30 days split into 5 lags of 6 days each fitted simultaneously

|CDP| – |AP| Diff - The difference in the magnitude of the effect size between AP and CDP (CDP-AP)

|CDP-Non CDP| Diff - The difference in the effect size between CDP and Non-CDP

Discussion

These findings suggest that a subject's location is an important consideration when assessing the effect of ambient air pollution on pneumonia mortality. In this study, the percentage relative risk of air pollution was significantly higher in CDP compared to AP by 0.44% (1-6 day lag), 0.28% (7-12 day lag) with a further important but non-significant increase of 0.31% in the 13-18 day lag. The 18 days prior to death indicated a lag period of increased risk, with the largest relative risk (CDP = 0.71% (0.03%,1.42%)) in the 13-18 day lag period. Prior to the 18th day a decrease in relative risk occurred, with the magnitude increasing in the lag period furthest from the event. The quadratic distributed lag plot confirmed an increased risk period of 20-22 days prior to the event. The decrease effect on mortality may be due to a small pool of people susceptible to dying from pollution effects⁸⁴. Initial changes in pollution cause the pool to diminish more rapidly than it can be replenished, creating a net deficit in the number of susceptible subjects; this leaves a larger proportion of stronger subjects to experience the longer lag periods and hence, a reduction in deaths.

In all lag periods, irrespective of the direction, the magnitude of the effect is greater in the CDP group. Hence, significant associations between BS and pneumonia mortality may have been overlooked in previous studies that have not taken into account location during exposure. Removing those who may have had a misleading exposure (due to hospitalisation) may have given a larger observed effect. Increased risk may be experienced in community based subjects due to reduced accesses to medical care that may have been able to prevent early symptoms from progressing to a critical phase. The bias associated with earlier recognition, and more timely and aggressive therapeutic intervention may be the reason for the apparently negative %RR associated with an increase in BS of $10\mu\text{g m}^{-3}$ within the "non-CDP" only subject group.

This study finds a similar increase in BS exposure effects when compared to U.S. studies of TSP/PM₁₀ on all-cause mortality when restricted to deaths located outside of hospital^{83,84}. Black smoke contains finer particle fractions dominated by combustion emissions which are more closely associated with health outcomes than PM₁₀ or PM_{2.5}, and so BS filter darkness measurements are a better marker for harmful combustion-related particles⁵¹. Currently, few published articles comparing indoor and outdoor BS

levels exist. Gotschi et al. compared indoor and outdoor BS and PM_{2.5} for 186 homes in Athens, Basel, Helsinki and Prague. The median indoor-outdoor ratios of BS were slightly less than PM_{2.5}, however, Spearman correlation coefficients were larger, possibly due to stronger indoor influences on PM_{2.5}³⁴⁵. Hoek et al. gave filter darkness regressions slopes (0.63-0.84) between indoor-outdoor concentrations in homes of four European cities³⁴⁶. Limited information with inconsistent outcomes, often due to small sample sizes (N≤50), is available comparing indoor and outdoor personal particulate matter exposure³⁴⁷⁻³⁴⁹. Janssen N.A.H et al. investigated personal, indoor and outdoor fixed site exposure to PM₁₀ in 37 participants from Amsterdam and PM_{2.5} and BS in 36 and 46 participants from Amsterdam and Helsinki respectively. Sampling was taken for 24hr periods, bi-weekly, over six months. In both cases, median concentrations were found in the personal monitors, followed by outdoor monitoring and then indoor. High correlations were produced between personal and outdoor fixed site monitors indicating that fixed site monitors are a good representation of the day-to-day variation in particulate matter exposure^{350,351}. However, high correlation does not imply the same absolute levels. The underlying premise; that exposure to airborne pollutants is reduced in hospitalised subjects; is supported by Wang et al. which showed a reduction in indoor concentrations of PM₁₀ and PM_{2.5} in 2 of 4 hospitals in Guangzhou, China³⁵². Subsequently, Wang et al. and later Morawska et al. further determined that a mechanical ventilation air conditioning system produced the lowest indoor-outdoor PM₁₀ ratios^{352,353}. Indoor air quality is an important issue for hospitals. However, currently the quantity of literature available on the relationship between indoor and outdoor air pollution, particularly regarding hospitals, is limited. Further observational studies are required to supplement understanding of the reduction and fluctuations in indoor hospital air pollution concentrations, in terms of distance from combustion sources, changes in ventilation systems and meteorological conditions³⁶¹.

One possible alternative explanation for the increase in relative risk occurring within the CDP group is that exposure may have a differing interaction with certain types of pneumonia that are specifically associated with the community. Pneumonia infection can be caused by a variety of micro-organisms. Hospital acquired pneumonia is primarily caused by *Staphylococcus aureus* or Gram-negative enterobacteria, and CAP is most commonly *Streptococcus pneumoniae* (35% of CAP cases)¹¹⁹. *Streptococcus*

pneumonia has an incubation period of 1-3 days, shorter than other pathogens such as *Haemophilus influenzae* and *Mycoplasma pneumoniae* with incubation periods of 2-4, and 6-32 days respectively ¹²⁰. Variation in source and incubation period may be a contributing confounding factor to the difference in BS effect on CDP mortality. Limited information from ICD coding constructed from the death certificate, allowed the pneumonia deaths to be classified into; bronchopneumonia (81%), pneumococcal & streptococcal pneumonia (5%), organism unspecified (13%), and all others (1%), of which 67%, 52%, 47%, and 35%, respectively were CDP subjects. Change in exposure effects on differing underlying causes of pneumonia may be a possible explanation for higher relative risk in CDP deaths. If type of pneumonia was the only explanation for higher RR, then we might expect the RRs for BS within the categories of pneumonia to be the same. In fact, the same pattern of a higher RR for CDP compared to hospital deaths was found for; bronchopneumonia, and pneumococcal & streptococcal types (although not for; organism unspecified and all other types which had much smaller sample sizes reducing the power available to determine the true effect).

However, pollution itself may be the causal factor. Particulate pollution may increase the risk of contracting pneumonia in a number of ways; by impairing microbial clearance via the mucociliary mechanism ⁴²¹, hindering macrophage phagocytosis ⁴²², or causing intense capillary engorgement and loss of epithelium ⁴²³. These effects might require an indeterminate dose (product of exposure concentration, respiratory minute volume and time) to materialise before manifesting in an increased susceptibility to pneumonia mortality. The temporal relationship between air pollution and pneumonia death analysed here may therefore comprise a period of chemical insult before, as well as a diagnosis to death interval enveloping the ‘incubation period’ as classically defined. These varying and relatively indeterminate periods may thus explain why %RR is reduced in the 7-12 day lag compared to 1-6 and 13-18 day lags (Table 2).

A comparison of the BS lag periods within the two models indicated the possible presence of mortality displacement within the data. Mortality displacement, also known as harvesting, is the accelerated progression of a frail sub-population to death followed by a delay in its replenishment. This is thought to be illustrated by an increase in the death rate from its baseline for a certain period after exposure, followed by a period

when the death rate seems to be below expected⁸⁴. In Table 2 the lags post 18 days seems to suggest a negative relative risk but in fact this may be due to the shrinking of the at risk population. When the overall effect across the 30 days is estimated (Table 1), the positive and negative estimates balance out to some extent. These results are therefore consistent with the mortality displacement phenomenon. Pneumonia mortality may be more susceptible to ‘harvesting’ as pneumonia is prevalent in the elderly^{124,520} and is often the final cause in the chain of causes leading to death, implying a high incidence in the frail sub-population compared to the general population. If mortality displacement is present then it would have been easy to miss any risk period if the model only included one term representing average black smoke across 30 days.

The stronger risk experienced for the 18 days prior to death in males within the CDP group may be due to a more outdoors lifestyle in males causing an increased interaction with exposure. Younger subjects may also be expected to experience higher exposure to outdoor pollution concentrations. Yet we found no age group difference in risk within the CDP group possibly due to relatively few subjects aged 65 or less (9.2%), or due to elderly patients being allowed to die at home rather than in hospital. One could further argue that it is difficult to accurately determine the exposure level for these subjects as the area contains only one exposure measurement site. Any exposure misclassification could be reduced if the number of measurement sites could be increased, making it easier to evaluate local variations in pollution levels. Even then, it is difficult to determine a subject location during the entire exposure period, especially when multiple control exposure periods are used. Other than to explore the relationship between BS and mortality; restricting the data to an 18 day lag period was not a part of our original aims and so we do recognise that we lose some validity in our p-values. However, as the 18 day lag period showed significant increases in risk we felt it was worth investigating further.

The study time period (1981-1996) does not necessarily limit these results to a historical interest only given that primary aim was to compare the effect of pollution on subjects in the community versus within hospital. In fact, the higher outdoor concentrations of black smoke in the 1980's and early 1990's were advantageous for testing this hypothesis given that higher pollution concentrations meant that the incidence of

pollution related pneumonia mortality would have been higher and this would make the difference in risk , if any, associated with indoor and outdoor exposure easier to detect. The dominant source of black smoke during both the study period and present day were from road vehicles as most smoke control procedures to reduce combustion of coal for domestic heating and industrial energy would already have been implemented by the start of the study period. However, the chemical composition of fine black particles may have altered somewhat since the study period as a result of technological changes in vehicle engine and emissions control systems. It is not possible to directly characterise the extent of such changes as it is not possible to selectively collect black particles from non-black particles during atmospheric sampling for chemical analyses.

Conclusions

In conclusion, evidence suggests that a subject's location is an important factor in relation to their likelihood of pneumonia mortality due to particulate pollution exposure. Including subjects who may have a lower exposure may increase bias in your results and as shown here underestimate the true effect of exposure on pneumonia deaths. The risk to mortality in all subjects, and in particular within the CDP group, tends to last a minimum of 18 days and peaks at the 13-18 day lag. This confirms that air pollution effects do exist beyond short term exposure periods such as 1-3 days, making it is important to investigate extended exposure periods of at least two to three weeks prior to death.

Table 2 and Table S2 repeated with p-value comparison replaced with z-test rather than interaction variable.

Table 2 –Percent change in risk for lagged black smoke air pollution and pneumonia mortality: repeated for AP(All Pneumonia), CDP (Community Death from Pneumonia) & Non-CDP.

	Lag (days)	AP		CDP only		Non - CDP only		CDP - AP Diff	CDP-Non CDP Diff	P-val
		% RR Change	95% C.I	% RR Change	95% C.I	% RR Change	95% C.I			
MODEL 1 Black Smoke	1-30	0.08%	-0.17%,0.35%	0.19%	-0.16%,0.58%	0.01%	-0.34%,0.40%	0.11%	0.18%	0.496
Air Temp "Low"	1-30	0.20%	0.11%,0.29%	0.22%	0.09%,0.35%	0.16%	0.03%,0.30%	0.02%	0.06%	0.546
Air Temp "High"	1-30	-0.05%	-0.20%,0.10%	-0.20%	-0.40%,0.00%	0.11%	-0.11%,0.34%	0.15%	0.31%	0.047
MODEL 2 Black Smoke	1-6	0.12%	-0.37%,0.62%	0.56%	-0.14%,1.29%	-0.31%	-0.99%,0.41%	0.44%	0.87%	0.090
	7-12	0.05%	-0.42%,0.53%	0.32%	-0.33%,1.00%	-0.25%	-0.92%,0.45%	0.28%	0.57%	0.244
	13-18	0.40%	-0.08%,0.90%	0.71%	0.03%,1.42%	0.14%	-0.56%,0.86%	0.31%	0.57%	0.256
	19-24	-0.09%	-0.55%,0.38%	-0.16%	-0.79%,0.50%	0.05%	-0.63%,0.75%	0.06%	0.21%	0.673
	25-30	-0.11%	-0.57%,0.36%	-0.39%	-1.01%,0.25%	0.34%	-0.35%,1.06%	0.28%	0.73%	0.126
Air Temp "Low"	1-30	0.19%	0.10%,0.29%	0.19%	0.06%,0.32%	0.18%	0.05%,0.32%	0.00%	0.01%	0.965
Air Temp "High"	1-30	-0.05%	-0.19%,0.10%	-0.20%	-0.39%,0.01%	0.09%	-0.12%,0.33%	0.15%	0.29%	0.064

%RR Change - percentage change in Relative Risk, associated with an increase of $10\mu\text{g}\text{m}^{-3}$ BS or a decrease of 1°C , on any individual day within the lag period, with corresponding 95% Confidence Interval (95% C.I.)

Model 1 - One 30 day lag, Model 2 - The 30 days split into 5 lags of 6 days each fitted simultaneously

|CDP| - |AP| Diff - The difference in the magnitude of the effect size between AP and CDP (CDP-AP)

|CDP-Non CDP| Diff - The difference in the effect size between CDP and Non-CDP (CDP - Non CDP)

Table S1 – Subgroup analysis of 18 day exposure split by subject characteristics; split by AP, CDP & Non-CDP

		AP			CDP only		Non-CDP Only		CDP – AP Diff	CDP-Non CDP Diff	p-val
	Lag (days)	% RR Change	95% C.I	% RR Change	95% C.I	% RR Change	95% C.I				
Overall	Black Smoke	1-18	0.30%	0.03%,0.59%	0.70%	0.29%,1.14%	-0.10%	-0.46%,0.30%	0.40%	0.80%	0.006
	Air Temp “Low”	1-18	0.14%	0.03%,0.25%	0.14%	0.00%,0.29%	0.13%	-0.03%,0.29%	0.00%	0.01%	0.895
	Air Temp “High”	1-18	-0.08%	-0.26%,0.11%	-0.20%	-0.45%,0.06%	0.05%	-0.22%,0.33%	0.12%	0.25%	0.207
Male	Black Smoke	1-18	0.32%	-0.09%,0.75%	0.83%	0.21%,1.51%	-0.14%	-0.66%,0.45%	0.51%	0.97%	0.027
	Air Temp “Low”	1-18	0.18%	0.02%,0.34%	0.16%	-0.05%,0.38%	0.18%	-0.05%,0.42%	0.02%	0.02%	0.893
	Air Temp “High”	1-18	-0.15%	-0.41%,0.13%	-0.40%	-0.76%,-0.01%	0.12%	-0.28%,0.54%	0.25%	0.52%	0.076
Female	Black Smoke	1-18	0.30%	-0.07%,0.69%	0.61%	0.08%,1.19%	-0.07%	-0.56%,0.47%	0.32%	0.68%	0.080
	Air Temp “Low”	1-18	0.11%	-0.03%,0.25%	0.12%	-0.07%,0.32%	0.08%	-0.13%,0.29%	0.02%	0.04%	0.754
	Air Temp “High”	1-18	-0.02%	-0.27%,0.24%	-0.04%	-0.38%,0.33%	-0.01%	-0.37%,0.37%	0.02%	0.03%	0.925
Age <80	Black Smoke	1-18	0.23%	-0.17%,0.66%	0.71%	0.10%,1.38%	-0.19%	-0.71%,0.38%	0.47%	0.90%	0.037
	Air Temp “Low”	1-18	0.11%	-0.06%,0.27%	0.03%	-0.19%,0.26%	0.17%	-0.06%,0.41%	0.08%	0.14%	0.384
	Air Temp “High”	1-18	-0.05%	-0.32%,0.24%	-0.30%	-0.67%,0.11%	0.19%	-0.20%,0.61%	0.25%	0.49%	0.095
Age ≥80	Black Smoke	1-18	0.37%	0.00%,0.77%	0.71%	0.18%,1.30%	-0.02%	-0.52%,0.53%	0.34%	0.73%	0.064
	Air Temp “Low”	1-18	0.16%	0.03%,0.31%	0.22%	0.03%,0.41%	0.09%	-0.11%,0.31%	0.05%	0.13%	0.377
	Air Temp “High”	1-18	-0.10%	-0.34%,0.16%	-0.13%	-0.45%,0.22%	-0.09%	-0.45%,0.30%	0.03%	0.04%	0.890

%RR Change - percentage change in Relative Risk, associated with an increase of 10µg⁻³ BS or a decrease of 1°C, on any individual day within the lag period, with corresponding 95% Confidence Interval (95% C.I.)

Model 1 - One 30 day lag, Model 2 - The 30 days split into 5 lags of 6 days each fitted simultaneously

|CDP| – |AP| Diff - The difference in the magnitude of the effect size between AP and CDP (CDP-AP)

|CDP-Non CDP| Diff - The difference in the effect size between CDP and Non-CDP

Appendix C Background details regarding the exposure data

Table C1 –Background statistics from Information Services Division Scotland, the number of cause specific deaths for years 1991-2011 in the four cities.

Primary COD	Pneumonia		Chronic lower respiratory diseases ^b		IHD		All cause	
	N	Row%	N	Row%	N	Row%	N	Col %
1991	3,785	6.20	2,572	4.21	16,866	27.63	61,041	4.80
1992	3,729	6.12	2,559	4.20	16,536	27.14	60,937	4.79
1993	4,495	7.02	2,990	4.67	16,925	26.43	64,049	5.03
1994	3,757	6.33	2,499	4.21	15,234	25.68	59,328	4.66
1995	4,021	6.65	2,784	4.60	14,977	24.76	60,500	4.76
1996	4,155	6.85	2,684	4.43	14,647	24.15	60,654	4.77
1997	4,028	6.77	2,764	4.65	14,013	23.55	59,494	4.68
1998	4,064	6.87	2,831	4.79	13,419	22.68	59,164	4.65
1999 ^a	4,526	7.51	3,137	5.20	13,337	22.12	60,281	4.74
2000	2,312	4.00	3,009	5.21	12,412	21.47	57,799	4.54
2001	2,370	4.13	2,988	5.21	11,914	20.76	57,382	4.51
2002	2,466	4.24	3,024	5.20	11,692	20.12	58,103	4.57
2003	2,859	4.89	3,170	5.42	11,441	19.57	58,472	4.60
2004	2,399	4.27	2,907	5.17	10,778	19.18	56,187	4.42
2005	2,483	4.45	3,027	5.43	10,331	18.53	55,747	4.38
2006	2,513	4.56	3,001	5.45	9,532	17.30	55,093	4.33
2007	2,444	4.37	3,104	5.54	9,343	16.69	55,986	4.40
2008	2,453	4.40	3,037	5.45	8,841	15.87	55,700	4.38
2009	2,348	4.36	2,986	5.54	8,274	15.36	53,856	4.23
2010	2,324	4.31	2,807	5.20	8,138	15.08	53,967	4.24
2011	1,948	3.63	3,062	5.71	7,636	14.23	53,661	4.22
Total	65,479	5.38	60,942	5.01	256,286	21.05	1,217,401	

^a –ICD10 replace ICD9 cause of death coding. ^b – ISD report is not COPD specific rather all-encompassing Chronic lower respiratory diseases.

Table C2a – Background Pollution Monitor Information for Scotland’s most populated cities (Aberdeen, Dundee, Edinburgh, Glasgow, Inverness)

Site	Freq	Postcode	Pollutants	Start (mth/yr)	End (mth/yr)	Comp%
Aberdeen 1	D	AB24	BS & SO2	Apr-61	Mar-82	97
Aberdeen 2	D	AB11	BS & SO2	Apr-61	Mar-92	88
Aberdeen 3	D	AB25	BS & SO2	Apr-91	May-05	68
Aberdeen	H	AB24	SO ₂	Jan-01	Sep-07	96.3
	H		CO	Sep-99	Sep-07	95
	H		NO,NO2	Sep-99	Jun-12	93.1
	H		O ₃	Aug-03	Jun-12	96.2
	H		PM ₁₀	Sep-99	Jun-12	93
	H		PM2.5	Feb-09	Jun-12	83-95
	H		NO,NO2	Jan-08	Jun-12	84-92
Aber Union St	H	AB10	NO,NO2	Jan-08	Jun-12	84-92
Dundee 12	D	DD2	BS,SO2	Mar-72	Mar-82	84.3
Dundee 13	D	DD2	BS,SO2	Mar-72	Mar-82	63.6
Dundee 14	D	DD3	BS,SO2	Mar-72	Mar-82	73
Dundee 15	D	DD4	BS,SO2	Mar-72	Mar-82	80.7
Dundee 16	D	DD5	BS,SO2	Mar-72	Mar-82	82.9
Edinburgh 10	D	EH11	BS, SO2	Apr-62	Mar-82	91.7
Edinburgh 12	D	EH5	BS, SO2	Apr-62	Mar-92	88.2
Edinburgh 14	D	EH9	BS, SO2	Apr-66	Mar-97	93.6
Edinburgh 15	D	EH6	BS, SO2	Apr-66	Mar-82	95.5
Edinburgh 16	D	EH15	BS, SO2	Apr-66	Mar-82	96.9
Edinburgh 17	D	EH17	BS, SO2	Apr-66	Mar-82	97.7
Edinburgh 19	D	EH14	BS, SO2	Apr-66	Mar-82	61.9
Edinburgh 20	D	EH7	BS, SO2	Apr-66	Mar-82	92.5
Edinburgh 23	D	EH10	BS, SO2	Apr-79	Mar-82	75.4
Edinburgh 24	D	EH4	BS, SO2	Apr-91	Mar-98	88.2
Edinburgh 25	D	EH1	BS, SO2	Apr-97	Dec-05	47.6
Edinburgh St Leon	D	EH8	BS	Oct-06	Dec-09	86
	H		SO ₂ , O ₃ , NO, NO ₂ , PM ₁₀	Nov-03	Jun-12	89-97.2
	H		PM _{2.5}	Oct-08	Jun-12	94
	H		CO	Nov-03	Jun-12	96.2
Edinburgh Centre	H	EH2	SO ₂ , O ₃ , NO, NO ₂ , PM ₁₀	Oct-92	Oct-03	89.1-92.3

Freq – Frequency (D=Daily, H=Hourly) measurements taken.
Comp% - Percentage of measurements completed.

Table C2b – Background Pollution Monitor Information for Scotland’s most populated cities (Aberdeen, Dundee, Edinburgh, Glasgow, Inverness)

Site	Freq	Postcode	Pollutants	Start (mth/yr)	End (mth/yr)	Comp%
Glasgow 20	D	G1	BS, SO2	Apr-61	Dec-05	81.8
Glasgow 42	D	G20	BS, SO2	Apr-61	Apr-89	90.9
Glasgow 44	D	G42	BS, SO2	Apr-61	Mar-80	92.7
Glasgow 47	D	G32	BS, SO2	Apr-61	Mar-88	94
Glasgow 51	D	G52	BS, SO2	Apr-61	Dec-05	91.2
Glasgow 52	D	G5	BS, SO2	Apr-61	Apr-89	93.1
Glasgow 60	D	G33	BS, SO2	Apr-65	Apr-84	91.4
Glasgow 61	D	G45	BS, SO2	Apr-67	Mar-82	88.2
Glasgow 62	D	G42	BS, SO2	Apr-67	Mar-82	87.1
Glasgow 66	D	G15	BS, SO2	Mar-68	Mar-92	87.6
Glasgow 67	D	G13	BS, SO2	Apr-70	Mar-82	85.9
Glasgow 68	D	G21	BS, SO2	Apr-70	Mar-88	92.5
Glasgow 69	D	G11	BS, SO2	Mar-72	Nov-03	64.6
Glasgow 72	D	G3	BS, SO2	Apr-74	Mar-83	80.7
Glasgow 73	D	G73	BS, SO2	Apr-75	Dec-05	90.8
Glasgow 74	D	G41	BS, SO2	Apr-75	Mar-82	79.6
Glasgow 79	D	G71	BS, SO2	Mar-76	Mar-82	83.9
Glasgow 80	D	G32	BS, SO2	Mar-76	Mar-88	87
Glasgow 86	D	G22	BS, SO2	Apr-78	Apr-89	94.8
Glasgow 87	D	G31	BS, SO2	Apr-78	Mar-88	85
Glasgow 90	D	G32	BS, SO2	Apr-79	Apr-85	94.4
Glasgow 91	D	G40	BS, SO2	Apr-79	Mar-88	83.9
Glasgow 92	D	G31	BS, SO2	Apr-79	Apr-89	92
Glasgow 93	D	G42	BS, SO2	Apr-79	Mar-88	83.5
Glasgow 95	D	G69	BS, SO2	Apr-80	Dec-05	89.6
Glasgow 96	D	G3	BS, SO2	Mar-83	Mar-94	94.1
Glasgow 97	D	G33	BS, SO2	Mar-83	Mar-86	70
Glasgow 98	D	G21	BS, SO2	Mar-88	Dec-05	84.4

Freq – Frequency (D=Daily, H=Hourly) measurements taken.

Comp% - Percentage of measurements completed.

Table C2c – Background Pollution Monitor Information for Scotland’s most populated cities (Aberdeen, Dundee, Edinburgh, Glasgow, Inverness)

Site	Freq	Postcode	Pollutants	Start (mth/yr)	End (mth/yr)	Comp%
Glasgow Centre	D	G2	BS, SO2	Oct-06	Nov-08	83.1
	H		SO2,CO,NO,NO2,PM10	Jul-96	Jun-12	85.6-93.5
	D		PM ₁₀	Oct-00	May-04	67.8
	D		PM _{2.5}	Oct-00	Oct-07	70.1
	H		PM _{2.5}	Dec-08	Jun-12	68.3-96.3
	D		PM _{2.5}	Oct-00	May-04	67.8
	D		PM _{2.5}	Oct-00	Oct-07	70.1
Glas Centre	H		O ₃	Jul-96	Jun-12	97.6
Glas City Chamb	H	G2	CO	Jul-89	Sep-07	94.1
	H		NO,NO2	Jan-87	Mar-11	96.2
Glas Kerb(Rdside)	H	G2	CO	Mar-97	Sep-07	96.4
	H		NO,NO2,PM10	Mar-97	Jun-12	87.5-96.3
	H		PM2.5	May-09	Jun-12	63.5-89.7
Inverness	H	IV3	CO	Jul-01	Sep-07	97.3
	D		PM ₁₀	Jul-01	Dec-11	87.4
	D		PM _{2.5}	Jun-08	Dec-11	88.8
	H		NO,NO2	Jul-01	Jun-12	96.7

Freq – Frequency (D=Daily, H=Hourly) measurements taken.
Comp% - Percentage of measurements completed.

Table C3 - Number of hourly temperature measurements recorded at the four potential monitor sites for Glasgow and Inverness

Time	Glasgow Monitor ID				Inverness Monitor ID			
	972	977	978	1006	110	113	115	116
00:00	0	0	2,090	11,469	917	10,754	0	3,654
01:00	0	0	2,102	11,357	918	10,756	0	3,619
02:00	0	0	2,104	11,248	918	10,753	0	3,619
03:00	0	0	2,099	11,140	918	10,754	0	3,647
04:00	0	0	2,097	11,171	918	10,752	0	4,935
05:00	0	0	2,100	11,329	918	10,754	0	6,195
06:00	0	0	2,101	11,522	917	10,752	0	8,743
07:00	0	0	2,103	11,638	949	10,753	0	9,930
08:00	0	0	2,105	11,641	917	10,754	0	10,202
09:00	6,779	3,656	9,388	10,585	11,361	10,750	11,524	11,466
10:00	0	139	2,105	11,531	918	10,746	0	10,434
11:00	0	0	2,102	11,612	919	10,749	0	10,445
12:00	0	0	2,102	11,628	924	10,747	0	11,472
13:00	0	0	2,093	11,632	926	10,751	0	10,483
14:00	0	0	2,101	11,696	927	10,757	0	10,472
15:00	0	0	2,101	11,691	926	10,755	0	11,404
16:00	0	0	2,097	11,671	924	10,754	0	10,435
17:00	0	0	2,102	11,687	923	10,756	0	10,416
18:00	0	0	2,103	11,664	921	10,755	0	11,204
19:00	0	0	2,101	11,629	919	10,757	0	10,108
20:00	0	0	2,102	11,617	919	10,756	0	9,850
21:00	0	0	2,103	11,571	918	10,756	0	8,162
22:00	0	0	2,100	11,540	918	10,757	0	5,155
23:00	0	0	2,105	11,508	918	10,757	0	3,731
Total	6,779	3,795	57,706	275,777	32,551	258,085	11,524	199,781

Appendix D Supplementary analysis results

Table D1 - Cause of death specific percentage relative risk (%RR) associated with an 10µgm⁻³ increase in pollutants, with corresponding comparison test of three causes of death.

Lag Period/ Pollutant	Primary Cause of Death - %RR (95%.C.I.) ^a					
	Pneumonia	COPD	P-val vPneu	IHD	P-val vPneu	P-val vCOPD
Black Smoke (per 10 µgm-3 increase)						
30 Days	0.00(-0.08,0.09)	0.08(-0.08,0.25)	0.424	0.00(-0.08,0.07)	0.942	0.389
1-6 Days	-0.11(-0.31,0.10)	0.07(-0.28,0.43)	0.388	-0.09(-0.31,0.14)	0.897	0.453
7-12 Days	-0.02(-0.23,0.20)	-0.13(-0.63,0.38)	0.683	-0.07(-0.19,0.05)	0.662	0.821
13-18 Days	0.02(-0.19,0.24)	-0.10(-0.34,0.15)	0.460	0.08(-0.08,0.24)	0.679	0.231
19-24 Days	-0.11(-0.31,0.10)	0.17(-0.20,0.54)	0.201	0.06(-0.05,0.17)	0.152	0.592
25-30 Days	-0.08(-0.38,0.23)	0.19(-0.06,0.44)	0.182	-0.03(-0.16,0.11)	0.754	0.134
PM 10 (per 10 µgm-3 increase)						
30 Days	-0.02(-0.28,0.26)	0.20(-0.09,0.52)	0.284	-0.11(-0.21,0.00)	0.548	0.049
1-6 Days	0.26(-0.31,0.84)	0.01(-0.43,0.45)	0.498	-0.43(-0.66,-0.20)	0.026	0.082
7-12 Days	-0.08(-0.53,0.37)	0.04(-0.41,0.49)	0.715	-0.10(-0.34,0.13)	0.934	0.589
13-18 Days	0.18(-0.27,0.63)	0.35(-0.35,1.07)	0.689	-0.06(-0.30,0.17)	0.354	0.277
19-24 Days	-0.15(-0.67,0.40)	0.22(-0.22,0.67)	0.305	-0.21(-0.45,0.02)	0.818	0.088
25-30 Days	0.30(-0.15,0.76)	0.32(-0.38,1.04)	0.967	0.06(-0.26,0.39)	0.407	0.519
PM 2.5 (per 10 µgm-3 increase)						
30 Days	0.61(-0.41,2.01)	0.38(-0.23,1.10)	0.731	0.06(-0.30,0.47)	0.358	0.403
1-6 Days	0.08(-1.34,1.64)	1.88(0.07,3.90)	0.142	-0.33(-1.10,0.47)	0.623	0.028
7-12 Days	0.73(-0.96,2.60)	0.04(-2.50,3.04)	0.683	0.70(-0.51,2.00)	0.979	0.673
13-18 Days	-0.53(-1.85,0.90)	-0.14(-1.44,1.26)	0.692	0.14(-0.60,0.91)	0.406	0.719
19-24 Days	0.03(-1.28,1.45)	1.35(0.14,2.65)	0.163	-0.51(-1.44,0.46)	0.520	0.019
25-30 Days	0.54(-0.78,1.97)	-0.46(-1.57,0.73)	0.274	0.30(-1.21,1.97)	0.827	0.443
Sulphur Dioxide (per 10 µgm-3 increase)						
30 Days	0.00(-0.09,0.09)	0.01(-0.07,0.10)	0.860	-0.02(-0.07,0.03)	0.675	0.515
1-6 Days	0.01(-0.22,0.23)	0.15(-0.17,0.48)	0.474	-0.06(-0.20,0.08)	0.621	0.239
7-12 Days	0.01(-0.17,0.19)	-0.09(-0.49,0.33)	0.682	-0.02(-0.11,0.08)	0.801	0.753
13-18 Days	0.03(-0.27,0.34)	0.06(-0.15,0.28)	0.876	0.07(-0.02,0.17)	0.814	0.945
19-24 Days	-0.03(-0.31,0.26)	0.08(-0.14,0.30)	0.553	0.02(-0.13,0.17)	0.784	0.636
25-30 Days	0.01(-0.18,0.19)	0.03(-0.26,0.32)	0.912	-0.09(-0.19,0.01)	0.355	0.456
Nitrogen Dioxide (per 10 µgm-3 increase)						
30 Days	-0.05(-0.23,0.16)	0.19(0.06,0.33)	0.051	-0.04(-0.18,0.10)	0.998	0.016
1-6 Days	-0.27(-0.51,-0.02)	-0.29(-0.74,0.18)	0.940	-0.09(-0.23,0.05)	0.216	0.419
7-12 Days	-0.21(-0.51,0.10)	0.03(-0.56,0.64)	0.491	-0.07(-0.21,0.07)	0.433	0.749
13-18 Days	-0.08(-0.33,0.18)	0.23(-0.22,0.69)	0.254	0.11(-0.03,0.25)	0.224	0.621
19-24 Days	0.27(-0.22,0.79)	0.45(0.01,0.90)	0.603	-0.01(-0.14,0.13)	0.293	0.051
25-30 Days	0.06(-0.19,0.32)	0.24(-0.03,0.51)	0.352	0.03(-0.10,0.17)	0.839	0.181

a. Percentage Relative Risk (95% Confidence Interval) per 10µgm⁻³ increase of pollutant on any single day within the lag period.

vPneu. p-value associated with z-test comparison with Pneumonia

vCOPD. p-value associated with z-test comparison with COPD

Figure D1 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in gaseous pollutants on pneumonia split by hospital admission during exposure (zero, 1-29, and all 30 days).

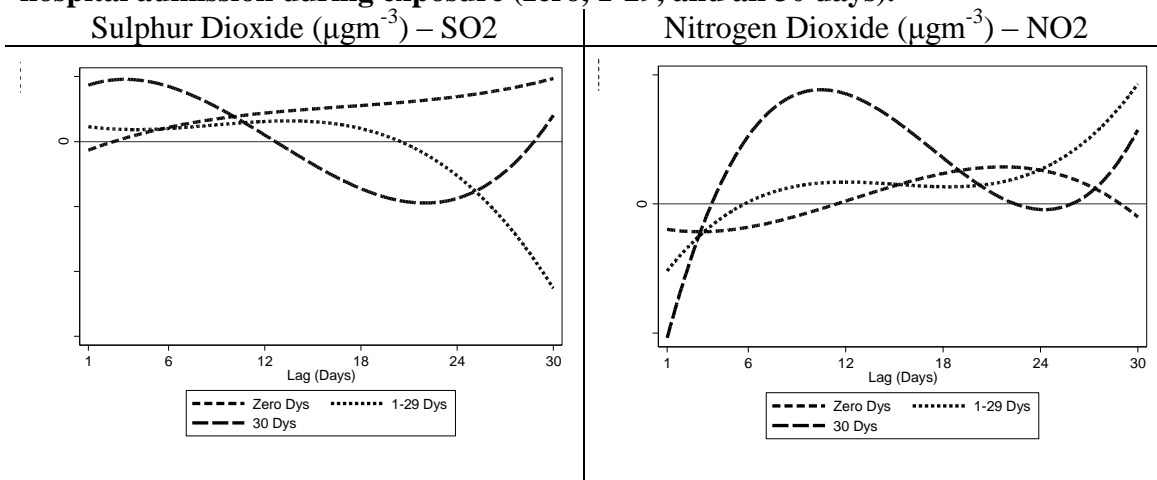


Figure D2 - - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in gaseous pollutants on COPD split by hospital admission during exposure (zero, 1-29, and all 30 days).

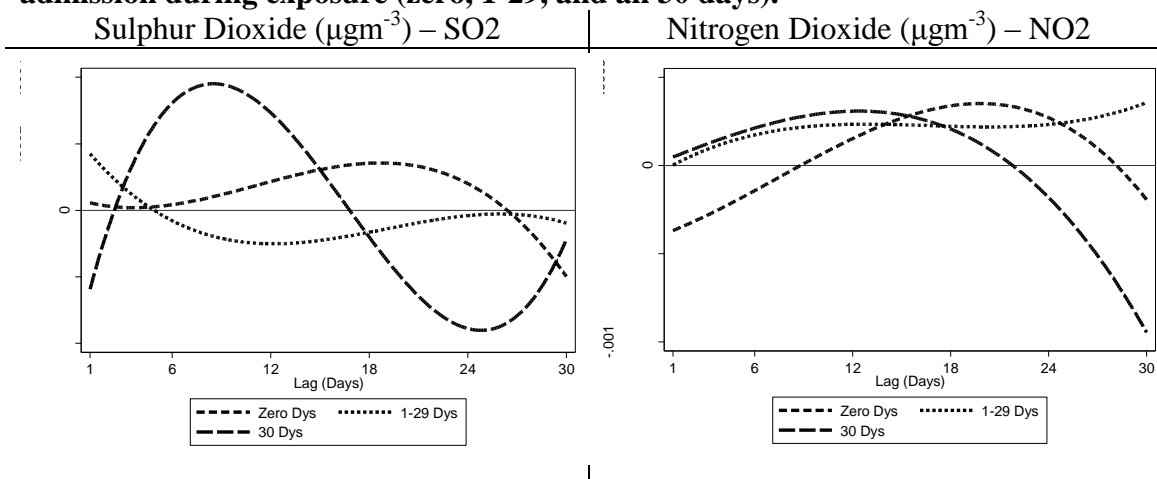
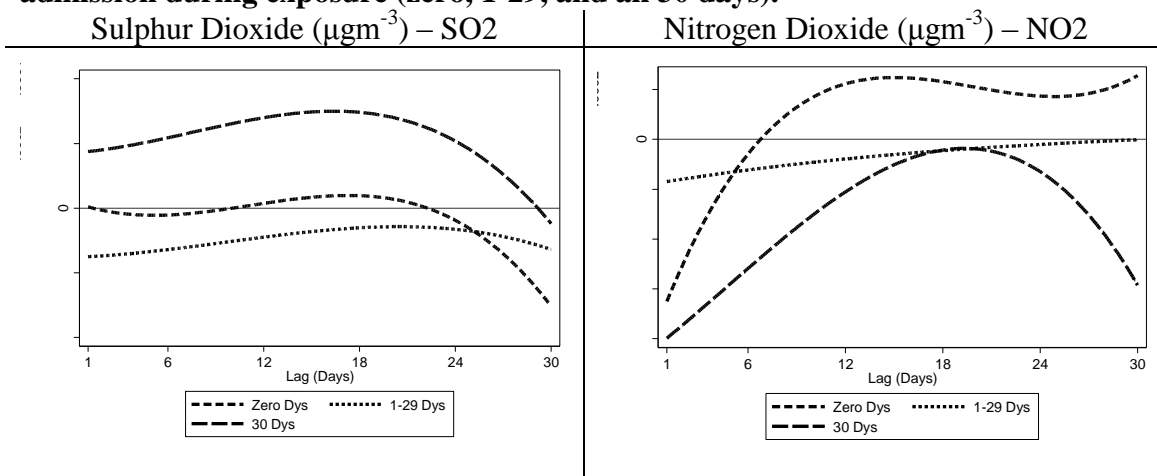


Figure D3 - Plotting the change in mortality risk described by a cubic distributed lag model associated with increase in gaseous pollutants on IHD split by hospital admission during exposure (zero, 1-29, and all 30 days).



Analysis of all subjects repeated for removed outliers and Multiple Imputation

Table D2a – Percentage relative risk per 10µg⁻³ increase of pollutant associated with the lag stratified analysis for pneumonia, results of analysis investigating outliers and missing data in all study subjects.

Lag Period/ Pollutant	Pneumonia %RR (95%.C.I.) ^a - By Hospital Admission Status		
	CC - 100% Exposure	CC - 99% Exposure	MI 100% Exposure
Black Smoke (per 10 µg⁻³ increase)			
30 Days	0.02(-0.05,0.08)	0.04(-0.08,0.17)	0.01(-0.04,0.05)
1-6 Days	-0.09(-0.21,0.04)	-0.18(-0.38,0.01)	0.00(-0.10,0.10)
7-12 Days	-0.01(-0.16,0.13)	-0.17(-0.36,0.03)	0.02(-0.09,0.12)
13-18 Days	0.02(-0.17,0.21)	0.04(-0.30,0.39)	0.03(-0.10,0.15)
19-24 Days	-0.11(-0.24,0.02)	0.01(-0.21,0.22)	-0.06(-0.16,0.04)
25-30 Days	0.03(-0.10,0.15)	-0.07(-0.30,0.17)	0.07(-0.03,0.18)
PM 10 (per 10 µg⁻³ increase)			
30 Days	0.01(-0.09,0.12)	0.00(-0.13,0.14)	-0.02(-0.09,0.06)
1-6 Days	0.11(-0.13,0.34)	0.04(-0.23,0.31)	0.03(-0.18,0.24)
7-12 Days	-0.19(-0.43,0.05)	-0.20(-0.47,0.07)	-0.13(-0.34,0.09)
13-18 Days	0.03(-0.20,0.27)	-0.02(-0.29,0.26)	-0.03(-0.25,0.20)
19-24 Days	-0.06(-0.36,0.25)	-0.06(-0.43,0.32)	0.19(-0.03,0.41)
25-30 Days	0.07(-0.17,0.30)	0.07(-0.20,0.34)	-0.13(-0.36,0.11)
PM 2.5 (per 10 µg⁻³ increase)			
30 Days	0.27(-0.07,0.66)	0.47(-0.01,1.01)	0.17(-0.18,0.57)
1-6 Days	0.36(-0.35,1.11)	0.54(-0.72,1.91)	0.39(-0.35,1.16)
7-12 Days	0.28(-0.58,1.19)	0.35(-0.63,1.40)	0.48(-0.71,1.75)
13-18 Days	-0.19(-1.08,0.74)	-0.12(-1.21,1.05)	-0.27(-1.01,0.51)
19-24 Days	0.32(-0.36,1.03)	0.20(-0.72,1.17)	0.37(-0.33,1.09)
25-30 Days	-0.56(-2.05,1.09)	-0.39(-2.21,1.65)	-0.39(-1.83,1.19)
Sulphur Dioxide (per 10 µg⁻³ increase)			
30 Days	0.03(-0.01,0.08)	0.05(-0.01,0.10)	0.04(-0.01,0.08)
1-6 Days	0.03(-0.08,0.14)	0.05(-0.09,0.18)	0.04(-0.07,0.14)
7-12 Days	-0.03(-0.14,0.08)	-0.04(-0.17,0.10)	-0.02(-0.13,0.08)
13-18 Days	0.10(0.00,0.21)	0.08(-0.05,0.22)	0.09(0.01,0.20)
19-24 Days	0.01(-0.09,0.12)	0.05(-0.09,0.18)	0.00(-0.10,0.10)
25-30 Days	0.08(-0.03,0.19)	0.09(-0.05,0.22)	0.07(-0.03,0.18)
Nitrogen Dioxide (per 10 µg⁻³ increase)			
30 Days	0.02(-0.05,0.08)	0.01(-0.07,0.09)	-0.02(-0.05,0.02)
1-6 Days	-0.24(-0.59,0.12)	-0.26(-0.62,0.10)	-0.24(-0.60,0.12)
7-12 Days	-0.08(-0.23,0.06)	-0.11(-0.27,0.05)	-0.11(-0.24,0.01)
13-18 Days	0.08(-0.11,0.27)	0.07(-0.15,0.29)	0.01(-0.12,0.13)
19-24 Days	-0.03(-0.24,0.18)	0.01(-0.18,0.21)	0.04(-0.08,0.17)
25-30 Days	0.12(-0.02,0.26)	0.11(-0.04,0.27)	0.02(-0.10,0.14)

a. Percentage Relative Risk (95% Confidence Interval) per 10µg⁻³ increase of pollutant on any single day within the lag period.
CC 100% Exposure Data. Complete Cases with 100% exposure data (repeat of main analysis).
CC 99% Exposure Data. Complete Cases with greatest 1% exposure data removed.
MI 100% Exposure Data. Multiple Imputed missing exposure data using complete observed exposure data

Table D2b - Percentage relative risk per 10 $\mu\text{g}\text{m}^{-3}$ increase of pollutant associated with the lag stratified analysis for COPD, results of analysis investigating outliers and missing data in all study subjects.

Lag Period/ Pollutant	COPD %RR (95%.C.I.) ^a - By Hospital Admission Status		
	CC - 100% Exposure	CC - 99% Exposure	MI 100% Exposure
Black Smoke (per 10 $\mu\text{g}\text{m}^{-3}$ increase)			
30 Days	0.08(-0.04,0.21)	0.17(-0.02,0.38)	0.04(-0.02,0.11)
1-6 Days	-0.01(-0.31,0.30)	-0.07(-0.48,0.34)	0.06(-0.19,0.32)
7-12 Days	-0.04(-0.23,0.15)	0.00(-0.29,0.29)	0.02(-0.13,0.18)
13-18 Days	0.14(-0.15,0.43)	0.27(-0.18,0.73)	0.09(-0.06,0.24)
19-24 Days	0.21(-0.12,0.55)	0.41(0.05,0.78)	0.02(-0.12,0.16)
25-30 Days	0.18(0.00,0.37)	0.27(-0.02,0.57)	0.15(-0.02,0.31)
PM 10 (per 10 $\mu\text{g}\text{m}^{-3}$ increase)			
30 Days	0.15(-0.05,0.36)	0.14(-0.13,0.43)	0.04(-0.06,0.15)
1-6 Days	-0.01(-0.32,0.31)	-0.10(-0.46,0.27)	-0.11(-0.39,0.18)
7-12 Days	0.29(-0.18,0.77)	0.36(-0.26,1.01)	0.17(-0.13,0.47)
13-18 Days	0.05(-0.26,0.38)	-0.07(-0.44,0.31)	-0.08(-0.38,0.22)
19-24 Days	0.07(-0.24,0.40)	0.14(-0.24,0.52)	0.36(0.06,0.67)
25-30 Days	0.16(-0.17,0.50)	0.10(-0.31,0.52)	0.12(-0.17,0.41)
PM 2.5 (per 10 $\mu\text{g}\text{m}^{-3}$ increase)			
30 Days	0.47(-0.27,1.38)	0.53(-0.25,1.50)	0.29(-0.16,0.79)
1-6 Days	1.05(0.14,2.01)	1.31(0.11,2.60)	1.04(0.11,2.02)
7-12 Days	0.08(-1.59,1.95)	0.17(-1.39,1.90)	0.10(-1.70,2.11)
13-18 Days	0.32(-0.53,1.20)	0.09(-1.02,1.27)	0.28(-0.59,1.20)
19-24 Days	0.71(-0.13,1.59)	1.16(-0.01,2.42)	0.82(-0.06,1.74)
25-30 Days	-0.24(-1.55,1.19)	-0.17(-1.62,1.43)	-0.02(-1.45,1.55)
Sulphur Dioxide (per 10 $\mu\text{g}\text{m}^{-3}$ increase)			
30 Days	0.00(-0.07,0.06)	0.01(-0.07,0.09)	0.01(-0.06,0.08)
1-6 Days	0.03(-0.19,0.26)	0.02(-0.23,0.27)	0.04(-0.13,0.21)
7-12 Days	-0.02(-0.25,0.21)	-0.03(-0.32,0.26)	-0.06(-0.23,0.12)
13-18 Days	0.09(-0.18,0.37)	0.00(-0.33,0.34)	0.08(-0.10,0.27)
19-24 Days	0.06(-0.09,0.22)	0.20(0.00,0.40)	0.01(-0.14,0.16)
25-30 Days	-0.02(-0.18,0.14)	0.05(-0.17,0.28)	-0.01(-0.16,0.15)
Nitrogen Dioxide (per 10 $\mu\text{g}\text{m}^{-3}$ increase)			
30 Days	0.12(-0.05,0.30)	0.11(-0.08,0.30)	0.06(-0.06,0.18)
1-6 Days	-0.24(-0.64,0.17)	-0.27(-0.65,0.11)	-0.19(-0.48,0.11)
7-12 Days	0.07(-0.14,0.27)	0.03(-0.19,0.25)	0.04(-0.14,0.21)
13-18 Days	0.23(-0.11,0.58)	0.24(-0.09,0.57)	-0.05(-0.24,0.15)
19-24 Days	0.23(-0.18,0.66)	0.26(-0.14,0.66)	0.26(-0.13,0.67)
25-30 Days	0.15(-0.05,0.34)	0.17(-0.05,0.38)	0.05(-0.11,0.22)

a. Percentage Relative Risk (95% Confidence Interval) per 10 $\mu\text{g}\text{m}^{-3}$ increase of pollutant on any single day within the lag period.

CC 100% Exposure Data. Complete Cases with 100% exposure data (repeat of main analysis).

CC 99% Exposure Data. Complete Cases with greatest 1% exposure data removed.

MI 100% Exposure Data. Multiple Imputed missing exposure data using complete observed exposure data

Table D2c - Percentage relative risk per 10µg⁻³ increase of pollutant associated with the lag stratified analysis for IHD, results of analysis investigating outliers and missing data in all study subjects.

Lag Period/ Pollutant	IHD %RR (95%.C.I.) ^a - By Hospital Admission Status		
	CC - 100% Exposure	CC - 99% Exposure	MI 100% Exposure
Black Smoke (per 10 µg⁻³ increase)			
30 Days	-0.02(-0.07,0.04)	-0.07(-0.16,0.03)	-0.01(-0.04,0.03)
1-6 Days	-0.13(-0.34,0.09)	-0.27(-0.57,0.03)	-0.07(-0.20,0.07)
7-12 Days	-0.10(-0.25,0.04)	-0.09(-0.29,0.13)	-0.03(-0.11,0.05)
13-18 Days	0.07(-0.03,0.18)	0.15(-0.04,0.33)	0.07(-0.02,0.16)
19-24 Days	0.02(-0.08,0.12)	0.02(-0.17,0.21)	-0.02(-0.11,0.07)
25-30 Days	-0.03(-0.15,0.09)	-0.03(-0.22,0.17)	0.01(-0.11,0.12)
PM 10 (per 10 µg⁻³ increase)			
30 Days	-0.08(-0.17,0.02)	-0.07(-0.18,0.04)	-0.09(-0.15,-0.03)
1-6 Days	-0.42(-0.61,-0.23)	-0.45(-0.66,-0.23)	-0.33(-0.51,-0.16)
7-12 Days	-0.08(-0.40,0.23)	-0.11(-0.41,0.20)	-0.03(-0.21,0.15)
13-18 Days	0.03(-0.16,0.23)	0.03(-0.19,0.26)	-0.08(-0.26,0.11)
19-24 Days	-0.12(-0.32,0.07)	-0.14(-0.36,0.09)	-0.06(-0.25,0.12)
25-30 Days	0.20(-0.07,0.47)	0.22(-0.14,0.58)	0.06(-0.12,0.24)
PM 2.5 (per 10 µg⁻³ increase)			
30 Days	-0.15(-0.79,0.65)	-0.15(-0.84,0.75)	-0.07(-0.79,0.86)
1-6 Days	-0.16(-0.75,0.46)	-0.37(-1.13,0.43)	-0.21(-0.85,0.46)
7-12 Days	0.38(-0.67,1.51)	0.37(-0.79,1.62)	0.23(-0.82,1.35)
13-18 Days	0.23(-0.35,0.82)	0.22(-0.56,1.03)	0.25(-0.38,0.90)
19-24 Days	-0.27(-0.85,0.32)	-0.16(-0.92,0.64)	-0.24(-0.83,0.37)
25-30 Days	0.38(-0.36,1.15)	0.27(-0.67,1.28)	0.39(-0.29,1.10)
Sulphur Dioxide (per 10 µg⁻³ increase)			
30 Days	-0.03(-0.06,0.01)	-0.02(-0.06,0.03)	-0.03(-0.07,0.01)
1-6 Days	-0.13(-0.28,0.03)	-0.18(-0.38,0.02)	-0.10(-0.24,0.04)
7-12 Days	-0.03(-0.11,0.06)	0.03(-0.08,0.14)	-0.04(-0.12,0.05)
13-18 Days	0.06(-0.03,0.14)	0.05(-0.06,0.16)	0.05(-0.03,0.13)
19-24 Days	0.00(-0.12,0.11)	0.00(-0.14,0.14)	-0.02(-0.13,0.09)
25-30 Days	-0.07(-0.19,0.04)	-0.01(-0.15,0.13)	-0.07(-0.16,0.01)
Nitrogen Dioxide (per 10 µg⁻³ increase)			
30 Days	-0.04(-0.15,0.06)	-0.05(-0.15,0.05)	-0.04(-0.12,0.04)
1-6 Days	-0.27(-0.53,-0.01)	-0.29(-0.52,-0.06)	-0.27(-0.54,0.02)
7-12 Days	-0.10(-0.22,0.02)	-0.13(-0.26,0.00)	-0.05(-0.16,0.05)
13-18 Days	0.06(-0.06,0.19)	0.09(-0.04,0.23)	0.01(-0.09,0.12)
19-24 Days	0.08(-0.14,0.31)	0.10(-0.10,0.31)	0.06(-0.17,0.29)
25-30 Days	0.03(-0.09,0.15)	0.01(-0.11,0.14)	-0.02(-0.11,0.08)

a. Percentage Relative Risk (95% Confidence Interval) per 10µg⁻³ increase of pollutant on any single day within the lag period.
 CC 100% Exposure Data. Complete Cases with 100% exposure data (repeat of main analysis).
 CC 99% Exposure Data. Complete Cases with greatest 1% exposure data removed.
 MI 100% Exposure Data. Multiple Imputed missing exposure data using complete observed exposure data

Figure D4 – Plotting change in pneumonia mortality risk described by a cubic distributed lag model associated with unit increase in pollutant, repeated in all study subjects for removal of outliers and multiple imputation.

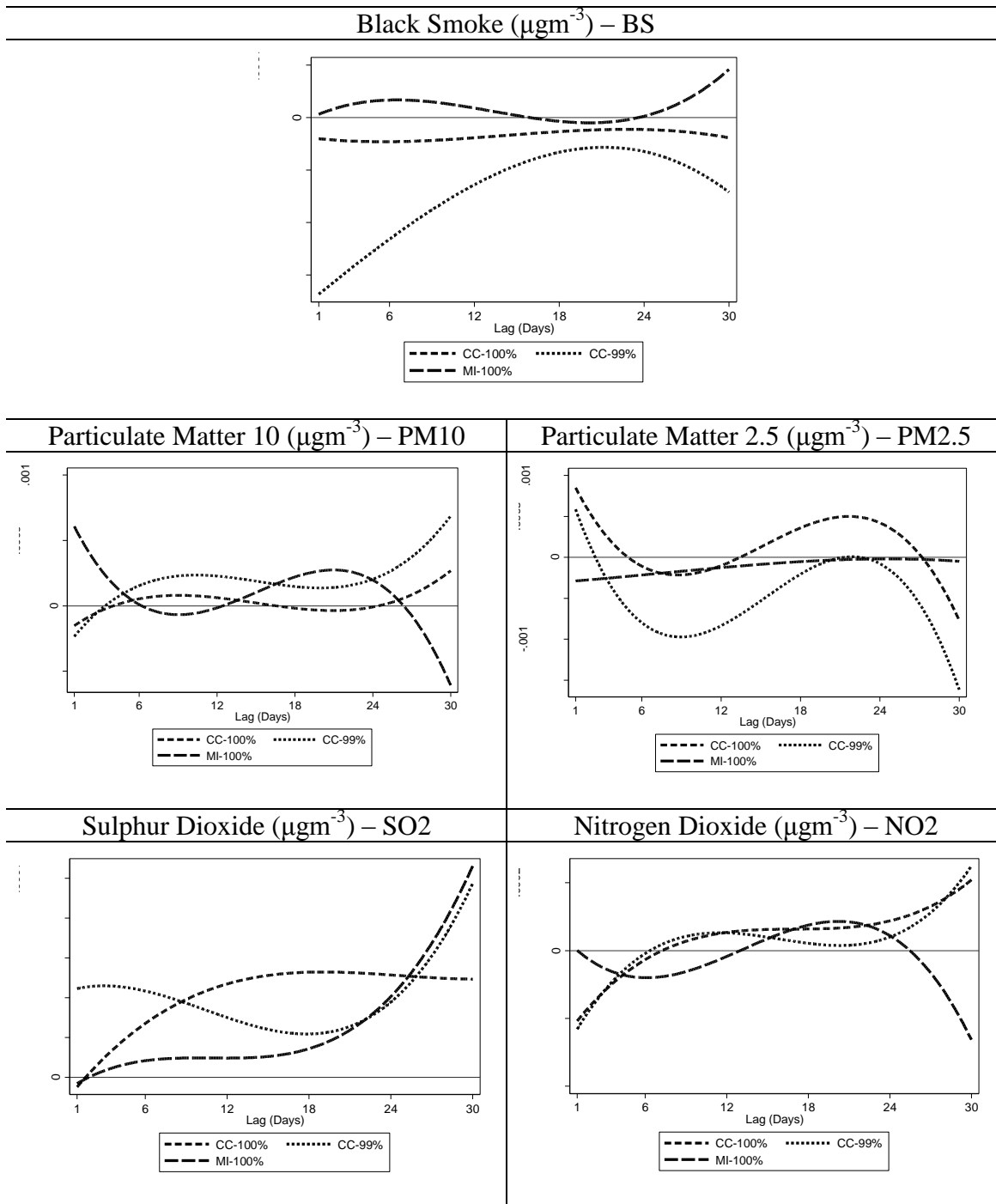


Figure D5 – Plotting change in COPD mortality risk described by a cubic distributed lag model associated with unit increase in pollutant, repeated in all study subjects for removal of outliers and multiple imputation.

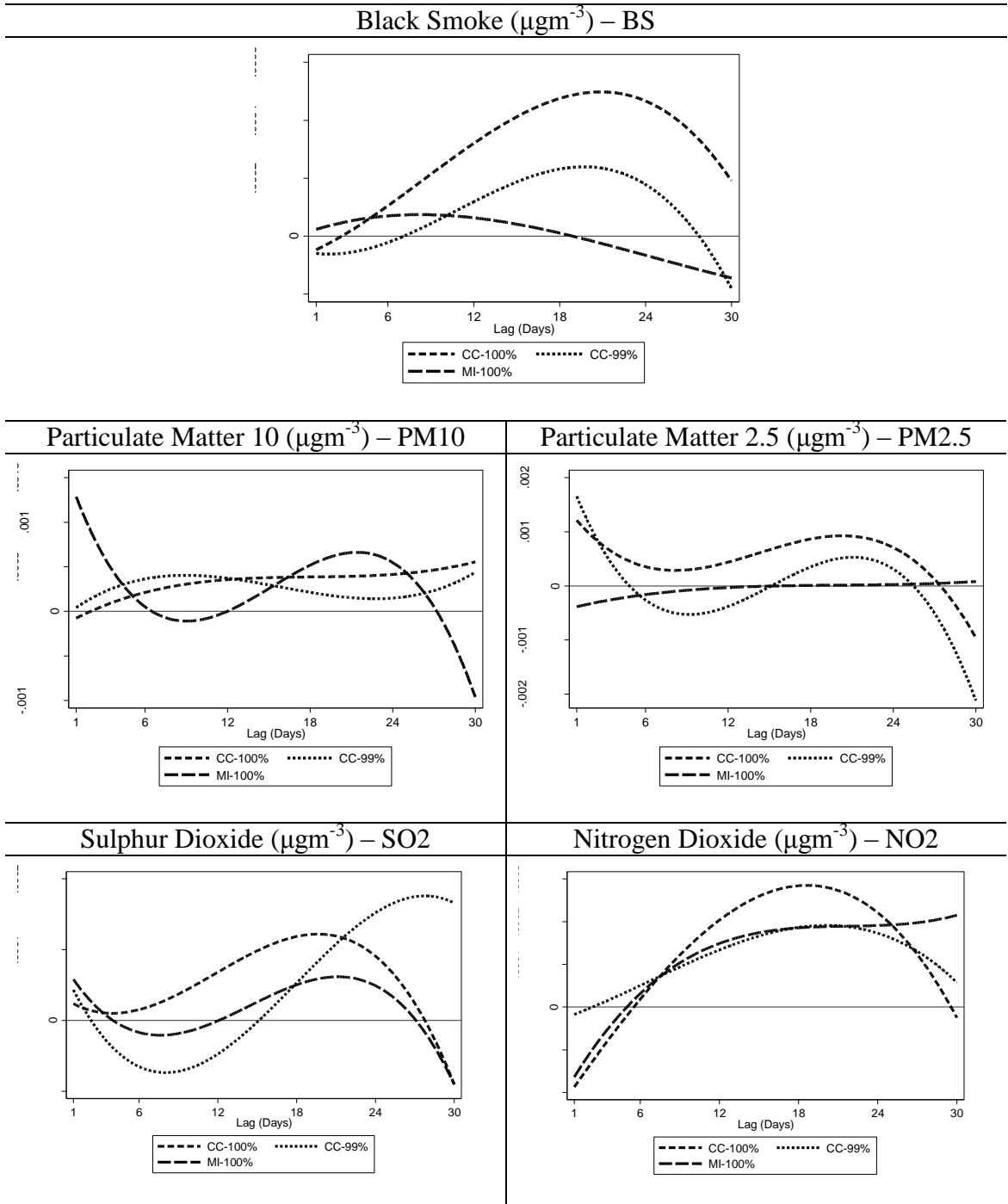
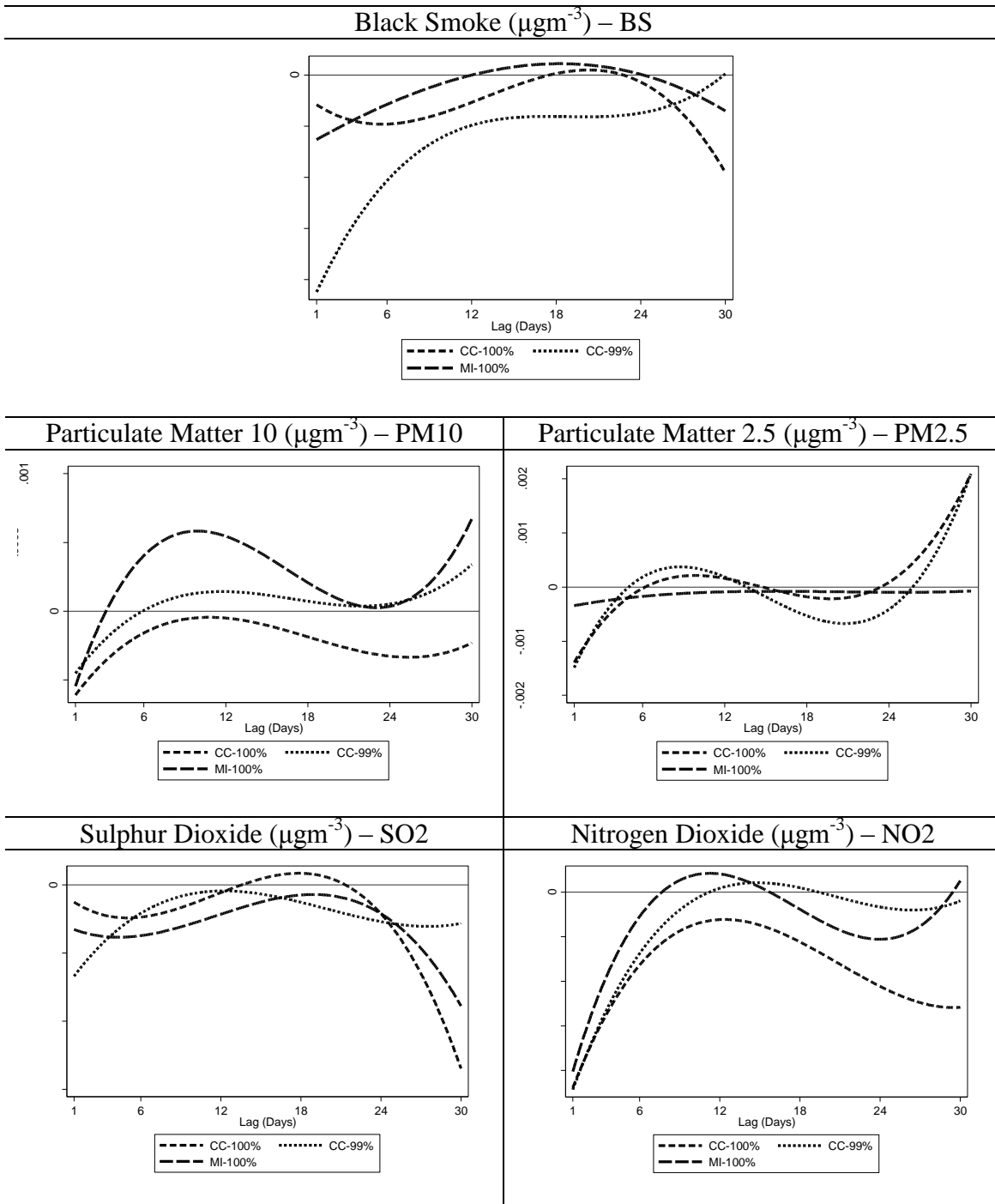


Figure D6 - Plotting change in IHD mortality risk described by a cubic distributed lag model associated with unit increase in pollutant, repeated in all study subjects for removal of outliers and multiple imputation.



Appendix E Example simulation study stata code

Following is an example of the Stata code used to run the simulation study.

```
*****
***MISSING DATA SIMULATIONS WHEN USING PREDICTIVE MEAN MATCHING TO ACCOUNT
***FOR THE NON-NORMALITY OF THE POLLUTION DATA
*****

Note; CC = COMPLETE CLASES ANALYSIS, MI = MULITPLE IMPUTATION, edbs = Pollution variable lag 0,
atempave = Temperature lag 0

Need three files

1) "OUTCOME FILE IN CASE-CROSSOVER FORMAT" - Contains outcome data in the case-crossover format
2) "EXPOSURE FILE CONTAINING COMPLETE POLLUTION AND TEMPERATURE DATA" - Contains complete
   exposure data at lag 0, and 30 days before and 30 days after, temperature lag0 and lag 1-3,
   plus date information year, month, dow, etc
3) "TEMPORARY FILE WITH MISSING DATA PATTERN" - Same as the exposure file except exposure is
   set to 0 and 1 where 1 means set to be a missing value

*****
***MISSING DATA SIMULATION WHEN ARTIFICALLY ADDING A MCAR 5% OF MISSING DATA
*****

use "OUTCOME FILE IN CASE-CROSSOVER FORMAT", clear

***SORT CASE-CROSSOVER FILE AND GENERATE VARIABLES TO STORE RESULTS FOR EACH REPETITION
sort id date
gen n5ro=.
label var n5 "Number and seq of simulations 5%"
gen miss5=.
label var miss5 "Number artificial missing days 5%"
gen ccmomiss5=.
label var ccmomiss5 "Number missing in CC mod 5%"
gen ccmomissg5=.
label var ccmomissg5 "Number subjects missing in CC mod 5%"
gen cpeffect5=.
label var cpeffect5 "Poll Effect size from CC mod 5%"
gen cpcse5=.
label var cpcse5 "Poll S.E. from CC mod 5%"
gen cteffect5=.
label var cteffect5 "Temp Effect size from CC mod 5%"
gen ctse5=.
label var ctse5 "Temp S.E. from CC mod 5%"
gen cmodsamp5=.
label var cmodsamp5 "Total sample in CC mod 5%"
gen mipeffect5=.
label var mipeffect5 "Poll Effect size from MI mod 5%"
gen mipse5=.
label var mipse5 "Poll S.E. from MI mod 5%"
gen miteffect5=.
label var miteffect5 "Temp Effect size from MI mod 5%"
gen mitse5=.
label var mitse5 "Temp S.E. from MI mod 5%"
gen mimodsamp5=.
label var mimodsamp5 "Total sample in MI mod 5%"

save "OUTCOME FILE IN CASE-CROSSOVER FORMAT", replace
**SET RANDOM NUMBER SEED
set seed 1437171
**BEGIN REPETITION LOOP
display "$S_TIME $S_DATE"

quietly forvalues i=1(1)1000 {
**OPEN TEMPORARY FILE CONTAINING MISSING DATA PATTERN
use "TEMPORARY FILE WITH MISSING DATA PATTERN", clear
**RESET MISSING DATA PATTERN TO ZEROS
sort n
foreach var of varlist edbsp* edbsd* {
replace `var'=0
}
**MISSING DATA GENERATION
**RANDOMLY IDENTIFY 5% LAG 0 POLLUTION VARIABLE TO MISSING (5% of 1359 = 68)
**REPLACE MISSING DATA PATTERN FROM ZERO TO ONE IF MISSING
local r = 0
while `r' < 68 {
local d = (1+int((1359-1+1)*runiform()))
replace edbsp=1 if n==`d'
count if edbsp==1
local r = 0 + r(N)
}
```

```

}
sort n
**REPLACE MISSING DATA PATTERN VARIABLE TO ONE FOR THE POLLUTION ON THE 30 DAYS PRIOR AND POST LAG 0
forvalues n=1(1)30 {
replace edbsd'n'p=1 if edbsp[_n-'n']==1 & n!=.
replace edbsp'n'p=1 if edbsp[_n+'n']==1 & n!=.
}
sort n
save "TEMPORARY FILE WITH MISSING DATA PATTERN", replace
**MERGE TEMPORARY MISSING DATA FILE WITH EXPOSURE FILE CONTAINING COMPLETE POLLUTION EXPOSURE
use "EXPOSURE FILE CONTAINING COMPLETE POLLUTION AND TEMPERATURE DATA", clear
sort date
merge 1:1 n using "TEMPORARY FILE WITH MISSING DATA PATTERN"
tab _merge
drop _merge
tab edbsp if edbsp==1
local miss5ro = r(N)
**GENERATE EXPOSURE VARIABLES WITH MISSING DATA CREATED
foreach var of varlist edbs edbsd1 - edbsd30 edbsp1 - edbsp30 {
gen `var'm='var' if `var'p==0
}
**LOG TRANSFORM EXPOSURE VARIABLES AFTER ADDING CONSTANT (SMALLEST VALUE/TWO) TO ACCOUNT FOR ZERO
foreach var of varlist edbs edbsp1 - edbsp30 edbsd1 - edbsd30 {
replace `var'm=0.1 if `var'm==0
replace `var'm=ln(`var'm)
}
*****
**RUN MULTIPLE IMPUTATION PROCESS
*****
mi set wide
**Register variables to be imputed
mi register impute edbsm edbsp1m - edbsp30m edbsd1m - edbsd30m atempave atempavelto3
**Register complete variable
mi register regular date casecontrol dow month y1990 y1991 y1992 y1993 y1994 jan feb mar apr may jun
jul aug sep oct nov dec/*
*/ dowsun downon dowtue dowwed dowthu dowfri dowsat season edbsmon
**Run imputation model with predictive mean matching to create 10 imputation values
mi impute chained (pmm, knn(10)) edbsm edbsp1m edbsp2m edbsp3m edbsp4m edbsp5m edbsp6m edbsp7m
edbsp8m edbsp9m edbsp10m /*
*/ edbsp11m edbsp12m edbsp13m edbsp14m edbsp15m edbsp16m edbsp17m edbsp18m edbsp19m edbsp20m edbsp21m
edbsp22m edbsp23m /*
*/ edbsp24m edbsp25m edbsp26m edbsp27m edbsp28m edbsp29m edbsp30m /*
*/ edbsd1m edbsd2m edbsd3m edbsd4m edbsd5m edbsd6m edbsd7m edbsd8m edbsd9m edbsd10m edbsd11m edbsd12m
edbsd13m edbsd14m /*
*/ edbsd15m edbsd16m edbsd17m edbsd18m edbsd19m edbsd20m edbsd21m edbsd22m edbsd23m edbsd24m edbsd25m
edbsd26m edbsd27m /*
*/ edbsd28m edbsd29m edbsd30m /*
*/ atemp24ave atemp24avelto3 = date casecontrol feb mar apr may jun jul aug sep oct /*
*/ nov dec downon dowtue dowwed dowthu dowfri dowsat y1991 y1992 y1993 y1994, /*
*/ add(10) replace force dots burnin(50)
*****
**MANIPULATE THE IMPUTED VARIABLES INTO FOR READY FOR ANALYSIS
**EXPONENTIAL TRANSFORM THE IMPUTED VARIABLES WITH CONSTANT ADJUSTMENT
foreach var of varlist edbs edbsp1 - edbsp30 edbsd1 - edbsd30 {
replace `var'm=exp(`var'm)
replace `var'm=0 if `var'm<=0.1
}
foreach var of varlist _1_edbsm - _10_edbsm {
replace `var'=exp(`var')
replace `var'=0 if `var'<=0.1
replace `var'=round(`var')
}
**MI UNSET THE DATA
mi unset
**REMOVE EXCESS IMPUTED VALUES VARIABLES
drop edbsp1m_1_ - atemp24avelto3_10_
save "EXPOSURE FILE CONTAINING COMPLETE POLLUTION AND TEMPERATURE DATA", replace
**RE-IMPORT THE LAG 0 IMPUTED VALUE VARIABLES
mi import wide, imputed(edbsm=edbsm_1_ edbsm_2_ edbsm_3_ edbsm_4_ edbsm_5_ edbsm_6_ edbsm_7_ edbsm_8_
edbsm_9_ edbsm_10_)
drop mi_miss - edbsm_10_
sort date
**SAVE EXPOSURE DATA CONTAINING COMPLETE EXPOSURE VARIABLES, MISSING EXPOSURE PATTERN VARIABLES, AND IMPUTED VALUES VARIABLES
save "EXPOSURE FILE CONTAINING COMPLETE POLLUTION AND TEMPERATURE DATA", replace
**OPEN CASE-CROSSOVER OUTCOME FILE
use "OUTCOME FILE IN CASE-CROSSOVER FORMAT", clear
sort id date
replace n5ro='i' if _n=='i'
replace miss5ro='miss5ro' if _n=='i'
save "OUTCOME FILE IN CASE-CROSSOVER FORMAT", replace
**MERGE EXPOSURE FILE WITH MI VARIABLES WITH OUTCOME FILE
sort date
merge m:1 date using "EXPOSURE FILE CONTAINING COMPLETE POLLUTION AND TEMPERATURE DATA"
tab _merge
drop _merge

```

```

*****
*****
**RUN MULTIPLE IMPUTATIONS ANALYSIS AND STORE RESULTS
*****
*****
mi estimate, dots: clogit casecontrol edbsm atemp24ave, group(id)
sort id date
replace mimodsamp5ro=e(N) if _n==`i'
mat define b = e(b_mi)
replace mipeffect5ro=b[1,1] if _n==`i'
replace miteffect5ro=b[1,2] if _n==`i'
mat define V = e(V_mi)
replace mipse5ro=sqrt(V[1,1]) if _n==`i'
replace mitse5ro=sqrt(V[2,2]) if _n==`i'
list mimodsamp5ro mipeffect5ro mipse5ro miteffect5ro mitse5ro if _n==`i'
*****
**RUN COMPLETE CASES ANALYSIS AND STORE RESULTS
*****
*****
clogit casecontrol edbsm atemp24ave, group(id)
sort id date
replace ccmodsamp5ro=e(N) if _n==`i'
replace ccmomiss5ro=e(N_drop) if _n==`i'
replace ccmomissg5ro=e(N_group_drop) if _n==`i'
replace ccpeffect5ro= b[edbsm] if _n==`i'
replace ccpse5ro= se[edbsm] if _n==`i'
replace ccteffect5ro= b[atemp24ave] if _n==`i'
replace cctse5ro= se[atemp24ave] if _n==`i'

*****
*****
**RESTORE FILES BACK TO ORIGINAL STATE
*****
*****
**UNSET MI FORMAT AND KEEP ONLY THOSE VARIABLES WITH OUTCOME DATA AND SIMULATION RESULTS
mi unset
keep id dod date diffdays casecontrol edbsmon n* miss* ccmomiss* ccmomissg* ccpeffect* ccpse*
ccteffect* cctse* ccmodsamp* mipeffect* mipse* miteffect* /*
*/ mitse* mimodsamp* impeffect* impse* imteffect* imtse* imodsamp*
save "OUTCOME FILE IN CASE-CROSSOVER FORMAT", replace
**EXPOSURE DATA FILE - KEEP ALL THE COMPLETE EXPOSURE DATA ONLY
use "EXPOSURE FILE CONTAINING COMPLETE POLLUTION AND TEMPERAUTRE DATA", clear
mi unset
keep date n edbsmon casecontrol edbs atemp24ave atemp24ave1to3 season month year y1990 y1991 y1992
y1993 y1994 jan feb mar apr may jun jul aug sep oct nov dec dow dowsun /*
*/ dowmon dowtue dowwed dowthu dowfri dowsat edbsd1 edbsd2 edbsd3 edbsd4 edbsd5 edbsd6 edbsd7 edbsd8
edbsd9 edbsd10 edbsd11 edbsd12 edbsd13 edbsd14 edbsd15 edbsd16 edbsd17 /*
*/ edbsd18 edbsd19 edbsd20 edbsd21 edbsd22 edbsd23 edbsd24 edbsd25 edbsd26 edbsd27 edbsd28 edbsd29
edbsd30 edbsp1 edbsp2 edbsp3 edbsp4 edbsp5 edbsp6 edbsp7 edbsp8 edbsp9 /*
*/ edbsp10 edbsp11 edbsp12 edbsp13 edbsp14 edbsp15 edbsp16 edbsp17 edbsp18 edbsp19 edbsp20 edbsp21
edbsp22 edbsp23 edbsp24 edbsp25 edbsp26 edbsp27 edbsp28 edbsp29 edbsp30
save "EXPOSURE FILE CONTAINING COMPLETE POLLUTION AND TEMPERAUTRE DATA", replace
noisily display in yellow "`i'," _c
}

```

To perform the different MCAR and MAR scenarios replace the ‘missing data generation’ section in the above code with the appropriate examples of each scenario below.

```

*****
***MISSING DATA SIMULATION WHEN ARTIFICALLY ADDING A MCAR 15% OF MISSING DATA
*****
local r = 0
while `r' < 204 {
  local d = (1+int((1359-1+1)*runiform()))
  replace edbsp=1 if n==`d'
  count if edbsp==1
  local r = 0 + r(N)
}
sort n

*****
***MISSING DATA SIMULATION WHEN ARTIFICALLY ADDING MAR 15% OF MISSING DATA
***SUMMER:WINTER RATIO OF 75:25
*****
local r = 0
**randomly identify (204) 15% missing, of which (51) 25% in Winter 75% Summer
while `r' < 51 {
  **randomly choose dates within winter
  local dw = (1+int((718-1+1)*runiform()))
  replace edbsp=1 if ns==`dw' & season==0
  di `dw'
  count if edbsp==1
  *count number of missing
  local r = 0 + r(N)
}
**randomly pick additional 204 missing summer
while `r' < 204 {
  **randomly choose dates within summer
  local ds = (719+int((1359-719+1)*runiform()))
  replace edbsp=1 if ns==`ds' & season==1
  di `ds'
  count if edbsp==1
  *count number of missing
  local r = 0 + r(N)
}
drop ns
tab edbsp season, row
sort n

*****
***MISSING DATA SIMULATION WHEN ARTIFICALLY ADDING MAR 15% OF MISSING DATA
***MISSING BLOCKS MEAN BLOCKS = 7 DAYS WITH S.D. = 8 DAYS
*****
local r = 0
while `r' < 204 {
  **randomly choose block length
  local l = round(rnormal(7,8),1)
  di `l'
  *randomly choose start date of block
  local d = (1+int((1359-1+1)*runiform()))
  replace edbsp=1 if n==`d'
  **If the length is greater than 1 replace the following days of length "1" with missing
  if `l'>1 & `l'<204-`r' {
    di `l'
    replace edbsp=1 if n>`d' & n<=`d'+`l'
  }
  count if edbsp==1
  local r = 0 + r(N)
}
**check blocks length
sort n edbsp
gen misseq=1 if edbsp==1 & edbsp[_n-1]==0
replace misseq=misseq[_n-1]+1 if edbsp==1 & edbsp[_n-1]==1
tab misseq if edbsp==1
gen missblock=misseq if edbsp==1 & edbsp[_n+1]==0
gsort -n
replace missblock=missblock[_n-1] if edbsp==1 & edbsp[_n-1]==1
sort n
tab missblock if misseq==1
drop misseq
drop missblock
sort n

```

```

*****
***MISSING DATA SIMULATION WHEN ARTIFICALLY ADDING MAR 15% OF MISSING DATA
***MISSING BLOCKS MEAN BLOCKS = 7 DAYS WITH S.D. = 8 DAYS
***SUMMER:WINTER RATIO OF 75:25
*****

local r = 0
**randomly identify (204) 15% missing, of which (51) 25% in Winter 75% Summer
while `r' < 51 {
  sort ns
  **randomly choose dates within winter
  local dw = (1+int((718-1+1)*runiform()))
  replace edbsp=1 if ns==`dw' & season==0
  di `dw'
  **randomly choose length of missing block
  local l = round(rnormal(7,8),1)
  di `l'
  sort n
  **If the length is greater than 1 replace the following days of lenght "l" with missing
  if `l'>1 & `l'<=51-`r' {
    di `l'
    replace edbsp=1 if ns>`dw' & ns<=`dw'+`l'
  }
  count if edbsp==1
  *count number of missing
  local r = 0 + r(N)
}
**randomly add 75% missing in summer
**randomly pick additional 204 missing summer
while `r' < 204 {
  sort ns
  **randomly choose dates within summer
  local ds = (719+int((1359-719+1)*runiform()))
  replace edbsp=1 if ns==`ds' & season==1
  di `ds'
  **randomly choose length of missing block
  local l = round(rnormal(7,8),1)
  di `l'
  sort n
  **If the length is greater than 1 replace the following days of lenght "l" with missing
  if `l'>1 & `l'<=204-`r'{
    di `l'
    replace edbsp=1 if ns>`ds' & ns<=`ds'+`l'
  }
  count if edbsp==1
  *count number of missing
  local r = 0 + r(N)
}
tab edbsp season, row
*check that the blocks are of appropriate length
drop ns
sort n edbsp
gen misseq=1 if edbsp==1 & edbsp[_n-1]==0
replace misseq=misseq[_n-1]+1 if edbsp==1 & edbsp[_n-1]==1
tab misseq if edbsp==1
gen missblock=misseq if edbsp==1 & edbsp[_n+1]==0
gsort -n
replace missblock=missblock[_n-1] if edbsp==1 & edbsp[_n-1]==1
sort n
tab missblock if misseq==1
tab missblock season if misseq==1
tab edbsp season, row
drop misseq
drop missblock
sort n

```