Causal Modelling of Survival Data with Informative Noncompliance

A thesis submitted to The University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Medical and Human Sciences

2011

Lang'o Taabu Odondi

School of Medicine

Table of contents

List of Tables

List of Figures

Abstract

The University of Manchester; Lang'o Taabu Odondi; PhD; May 2011;

Causal modelling of survival data with informative noncompliance

Noncompliance to treatment allocation is likely to complicate estimation of causal effects in clinical trials. The ubiquitous nonrandom phenomenon of noncompliance renders per-protocol and as-treated analyses or even simple regression adjustments for noncompliance inadequate for causal inference. For survival data, several specialist methods have been developed when noncompliance is related to risk. The Causal Accelerated Life Model (CALM) allows time-dependent departures from randomized treatment in either arm and relates each observed event time to a potential event time that would have been observed if the control treatment had been given throughout the trial. Alternatively, the structural Proportional Hazards (C-Prophet) model accounts for all-or-nothing noncompliance in the treatment arm only while the CHARM estimator allows time-dependent departures from randomized treatment by considering survival outcome as a sequence of binary outcomes to provide an 'approximate' overall hazard ratio estimate which is adjusted for compliance. The problem of efficacy estimation is compounded for two-active treatment trials (additional noncompliance) where the ITT estimate provides a biased estimator for the true hazard ratio even under homogeneous treatment effects assumption. Using plausible arm-specific predictors of compliance, principal stratification methods can be applied to obtain principal effects for each stratum. The present work applies the above methods to data from the Esprit trials study which was conducted to ascertain whether or not unopposed oestrogen (hormone replacement therapy-HRT) reduced the risk of further cardiac events in postmenopausal women who survive a first myocardial infarction. We use statistically designed simulation studies to evaluate the performance of these methods in terms of bias and 95% confidence interval coverage. We also apply a principal stratification method to adjust for noncompliance in two treatment arms trial originally developed for binary data for survival analysis in terms of causal risk ratio. In a Bayesian framework, we apply the method to Esprit data to account for noncompliance in both treatment arms and estimate principal effects. We apply statistically designed simulation studies to evaluate the performance of the method in terms of bias in the causal effect estimates for each stratum. ITT analysis of the Esprit data showed the effects of taking HRT tablets was not statistically significantly different from placebo for both all-cause mortality and myocardial reinfarction outcomes. Average compliance rate for HRT treatment was 43% and compliance rate decreased as the study progressed. CHARM and C-Prophet methods produced similar results but CALM performed best for Esprit: suggesting HRT would reduce risk of death by 50%. Simulation studies comparing the methods suggested that while both C-Prophet and CHARM methods performed equally well in terms of bias, the CALM method performed best in terms of both bias and 95% confidence interval coverage albeit with the largest RMSE. The principal stratification method failed for the Esprit study possibly due to the strong distribution assumption implicit in the method and lack of adequate compliance information in the data which produced large 95% credible intervals for the principal effect estimates. For moderate value of sensitivity parameter, principal stratification results suggested compliance with HRT tablets relative to placebo would reduce risk of mortality by 43% among the most compliant. Simulation studies on performance of this method showed narrower corresponding mean 95% credible intervals corresponding to the the causal risk ratio estimates for this subgroup compared to other strata. However, the results were sensitive to the unknown sensitivity parameter.

Declaration

No portion of the work referred to in the thesis has been submitted in 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 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 and/or Reproductions described in it may take place is available from the Head of the School of Medicine.

Dedication

for the memory of my late dear mother

Acknowledgements

First and foremost my deepest gratitude goes to my supervisor, Dr. Roseanne McNamee, for her wholesome mentorship through invaluable advice, guidance and statistical expertise during my PhD studies. Roseanne's readiness to help and challenge at the same time forms the key pillars that shaped this eventful and fruitful journey. For lack of better words, I can only say thank you Roseanne for putting up with my occasional failure to meet deadlines. I also thank my advisor, Prof. Graham Dunn, for his readiness to assist at any time by providing useful directions/insights which ensured I never lost the big picture nor sacrificed details.

I wish to thank the whole Biostatistics Group; staff, friends and fellow students for their individual and collective help and support throughout my studies. My special thanks goes to The University of Manchester for the scholarship to pursue my PhD. Specifically, I wish to thank James Power, the Faculty's postgraduate recruitment officer, for all the assistance towards securing the scholarship.

My greatest appreciation goes to my family for their belief in me by giving me the confidence to trust my instincts and courage to pursue my dreams. Special thanks to my late dear mother for the determination, cheer and generosity in raising me and providing me with the very best. Sorry mum you never lived to see me graduate but I trust you share (from yonder) the joy of the whole odyssey. I also hope this work will make my late brother Gilbert proud.

Most importantly my thanks goes to supervisors of my life and my singular twin pride: Hellen and Nigel (a.k.a *Boss*) whose giggles and chuckles embodies pure beauty epitomized in innocence. Your safe pair of love, affection and understanding provided the irreplaceable peace of mind and encouragement during both happy and trying moments. True, *nothing lasts forever*, you shared in the pains of the challenge and now I invite you to celebrate the ultimate prize.

And finally, I sincerely thank anyone who I may have inadvertently forgotten to mention by name but have enriched my life academically and otherwise during my tenure as a student. I owe it to all of you and for all your hearts and minds please receive a king-size thank you.

Preface

I graduated from Moi University, Kenya, with a BSc (Hons) in Mathematics and Statistics (1995) and later an MPhil in Biostatistics (1999) before proceeding to Belgium where I studied and graduated with an MSc in Applied Statistics (2002) at Limburgs (now Hasselt) University. I taught High School Mathematics after my undergraduate degree and I was appointed a Tutorial Fellow in the Department of Mathematics and Statistics at Moi University after my MPhil studies. Before coming to Manchester for my PhD studies, I was an Assistant Lecturer of Biostatistics at McMaster University in Ontario, Canada.

Publication and presentation related to this work

- Publication

  Odondi, L. and McNamee, R. (2010). Performance of statistical methods for analysing survival data in the presence of non-random compliance. *Statistics in Medicine*, 29 : 2994 – 3003

- Work in progress

  "Bayesian structural models adjusting for noncompliance in survival data with two-active treatments"

  In preparation for the *Biostatistics* Journal

- Conference presentation

  "Evaluating performance of causal models for survival data with nonrandom compliance"

  ISCB 2009 (Prague)

List of abbreviations

ACE	Average causal effect
ACM	All-cause mortality
AFT	Accelerated failure time
AIC	Akaike information criterion
C-Prophet	Complier PROPortional Hazards Effect of Treatment
CALM	Causal accelerated life model
CACE	Complier average causal effect
CHARM	Causal hazard ratio adjustment regression model
EER	Extended exclusion criteria
ER	Exclusion criteria
HR	Hazard ratio
HRT	Hormone replacement therapy
ITT	Intention-to-treat
IV	Instrumental variable
LASSO	Least absolute shrinkage and selection operator
MAR	Missing at random
MCAR	Missing completely at random
MLE	Maximum likelihood estimation
MNAR	Missing not at random
MRCD	Myocardial reinfarction or cardiac death
OLS	Ordinary least squares
PCR	Principal component regression
PH	Proportional hazards
PLSR	Partial least squares regression
PMLE	Penalized maximum likelihood estimation
RCM	Rubin causal model
RCT	Randomized controlled trial
RR	Relative risk
SMM	Structural mean model
SUTVA	Stable unit treatment value assumption

Chapter 1

Introduction and Motivation

1.1 Introduction

We begin this introductory chapter by briefly outlining the broad context of our research question followed by a synopsis of the motivating study/data. The next section presents a discussion of causation in medical research and a review of research design focussing on randomized controlled clinical trials and spelling out key design features, their classification, merits and limitations as contrasted with observational studies. This is followed by a brief review of reasons of association and the interplay between causal inference and statistics. The next section provides a review of the framework of counterfactual causal modelling by first tracing the origin of the use of counterfactuals in empirical research in the first quarter of the last Century followed by an exposition of the use of counterfactuals and posttreatment variables to address noncompliance to treatment assignment in order to obtain well-defined causal estimates. An outline of key assumptions implicit in causal modelling using counterfactuals is then presented with a discussion of opposing viewpoints between proponents and critics of causal modelling using counterfactuals. The section ends with a skeletal review of alternative formulations of causal models. After a brief outline of the similarities between noncompliance and missing data phenomenon, the subsequent section reviews propensity scores as a

method of adjusting for confounding and extends it to building compliance prediction models (propensity to comply). This is followed by a review of different methods of estimating treatment effects starting with intention-to-treat which ignores compliance information then per-protocol analysis that evaluates treatment as received. The section ends with a brief description of methods which adjust for noncompliance while still retaining ideals of randomization starting with instrumental variables and then 'complier average causal effects' estimation techniques. These two techniques are considered as an entry point to principal stratification which is reviewed as a general framework for adjusting for noncompliance in both one and two (possibly multiple) treatment arms. The chapter concludes with a statement of broad aims and specific objectives of the present work and finally an outline of the whole thesis.

1.2 Research problem in context

Most research questions in the health sciences are causal in nature, for example, what is the efficacy of a given treatment in a given population? A primary objective of most medical research studies is to estimate causal effects when comparing two or more treatments (interventions). Effective randomization of subjects between the treatments groups plays a central role in permitting such statistical comparison (Lachin, 2000). In an ideal world we would expect subjects to perfectly comply with their respective treatment assignments. But reality is different and noncompliance to treatment assignment due to treatment discontinuation, switching or subject dropout from the study is a common feature in clinical trials. Non-adherence to treatment allocation may take the form of either all-or-nothing or partial noncompliance. Intention-to-treat (ITT) is the gold standard for efficacy analysis under perfect compliance to treatment allocation. The ITT is an as-randomized analysis where the treatment groups are compared as assigned and the resulting ITT estimators enjoy the benefit of randomization by preserving the baseline comparability between treatment groups as it retains all randomized subjects. But in the presence of treatment noncompliance, the ITT is likely to underestimate the treatment efficacy by mixing the effects of

treatment compliers and non-compliers (White and Pocock, 1996). Noncompliance is mostly a nonrandom phenomenon (Nagelkerke et al., 2000) and it is a challenge adjusting for such potentially informative entities. We note here that using the term noncompliance may inadvertently confer/imply moral judgement and some authors suggest alternative terminology like departures from treatment allocation (White, 2005). However, the present work will use noncompliance in its simple sense to mean failure to comply with assigned treatment.

Many practical-problem studies involve time-to-event data, and such data are common in many medical fields including clinical trials, cohort studies and epidemiology studies. Such studies are commonly referred to as survival analysis, for example studying time to death or myocardial reinfarction or re-emergence of a cancer tumour among others. While many methods have been developed to estimate causal effects in the presence of treatment noncompliance for continuous outcomes, less attention has been paid to analysis of survival data (Nie et al., 2011). The Cox (1972) proportional hazards model is considered the standard tool in analysis of survival data from randomized trials. Although the ITT estimates effectiveness (treatment effect in usual or ordinary conditions) even when there is noncompliance with treatment, the ITT estimates efficacy (treatment effect in ideal conditions) only under perfect compliance. But noncompliance is ubiquitous (and often informative) which makes it unrealistic to consider non-compliers in a trial as a random subset of all patients under study.

In the presence of random noncompliance, simple regression techniques that adjust for noncompliance would be adequate in making causal inference. But the nonrandom nature of noncompliance is likely to complicate drawing inference about causal effects if present in any trial. While per-protocol and as-treated analysis are the most commonly used methods to supplement ITT in evaluating efficacy (Little et al., 2009; White, 2005), these post-hoc analyses are bereft of the benefits of randomization. They may be biased due to selection effects which may be evident in different compliance behaviour in different arms. One solution is provided by application of the principal stratification framework (Frangakis and Rubin, 2002) which permits adjustment based on a posttreatment variable like a noncompliance status, which retains the ideals of randomization to produce well-defined causal effect estimates.

The problem of efficacy estimation is even more complicated in the presence of two-active treatment arms where the ITT may provide a biased estimator even under homogeneous (uniform) treatment effects assumption (Aalen, 1998; Baker and Kramer, 2005). Possible noncompliance in both arms compounds the identification problem for such trials (Brittain and Lin, 2005; Chiba, 2009). It is a challenge adjusting for possible noncompliance in both arms due to resulting identification problem. For example, double-blind placebo-controlled randomized clinical trials with active treatment may be complicated by two levels of noncompliance: a part from simple noncompliance in both active and placebo arms, an imperfect blinding may present an additional differential compliances in the two arms (Jin and Rubin, 2008). But baseline covariates which are plausible predictors of compliance may be used to address such identification problems. Using such plausible predictors in a Bayesian approach, we can develop marginal compliance models for each treatment arm and apply principal stratification methods to construct causal models which provide principal effects for each stratum (Roy et al., 2008).

In the next section we provide a brief summary of the study which motivated the present work. We will analyse data from this study with the objective of addressing the problems outlined above, i.e. adjusting for noncompliance in one and two treatment arms.

1.3 Motivating data: The Esprit study

The onset of menopause is often characterized by diminishing production of oestrogen hormones due to a decline in ovarian function. The resulting physiological and biological changes among such women's bodies commonly manifest themselves through unpleasant menopausal symptoms like vasomotor (mostly hot flushes and night sweating), urogenital atrophy, insomnia, depression, fatigue, skin changes and increased irritability (Pecorelli and Fallo, 1998). Poor management of these symptoms can impact negatively on the body leading to low quality of life among such women for the better part of the last third of their lives (Hill, 1996; Rees, 2005). Hormone replacement therapy (HRT) is a treatment for oestrogen-deficiency

Table 1.1: Esprit study: baseline characteristics (adopted from Cherry et al. 2002)

	HRT tablets		Placebo	
	Number	Mean (SD) or number (%)	Number	Mean (SD) or number (%)
Age at admission (years)	513	62.3 (5.2)	504	62.9 (4.9)
Age at end education (yrs)	498	14.9 (1.2)	479	14.9 (1.2)
Age at last menses (years)	487	46.3 (5.8)	476	46.6 (5.7)
BMI (kg/m^2)	507	26.8 (5.1)	500	26.7 (5.3)
Last occupation manual	499	286 (57%)	483	292 (60%)
Smoker at admission	513	276 (54%)	503	264 (52%)
Normally drinks (> 1 unit alcohol/week)	513	197 (38%)	504	177 (35%)
White	511	496 (97%)	503	489 (97%)
History				
Angina	512	140 (27%)	504	136 (27%)
High blood pressure	513	237 (46%)	504	211 (42%)
Stroke	513	39 (8%)	504	36 (7%)
Diabetes	513	79 (15%)	503	74 (15%)
Fracture (in previous 10 years)	512	71 (14%)	503	97 (19%)
Hysterectomy	513	140 (27%)	504	105 (21%)
Oral contraceptive use	509	187 (37%)	497	185 (37%)
Used HRT (> 12 months before admission)	512	62 (12%)	502	51 (10%)

symptoms which is mainly administered in two broad forms depending on whether a woman has her uterus intact or not: unopposed oestrogen (oestrogen taken by itself) for those who have had hysterectomy (removal of the uterus) or oestrogen with progestin for the non-hysterectomized. The addition of progestin is meant to counteract the effects of estrogen on the uterus like endometrial cancer, i.e. cancer of the lining of the uterus.

Although observational studies showed benefits of HRT in lowering rates of coronary heart disease (Grady et al., 1992; Grodstein and Stampfer, 1995), recent clinical trials (Barnabei et al., 2005; Cherry et al., 2002; Grady et al., 2002) failed to confirm such beneficial effects among postmenopausal women. While HRT has been shown to offer no significant benefit or harmful effect on cardiovascular risk reduction, HRT is still commonly prescribed to address other postmenopausal symptoms like osteoporosis (Cho and Mukherjee, 2005; Rees, 2005). The oEStrogen in the Prevention of ReiInfarction Trial (Esprit) (Cherry et al., 2002) was one of the studies which revealed no HRT benefit among postmenopausal women in England and Wales. The present work uses data from the Esprit study to motivate the methods and analyses considered herein. Table 1.1 provide a summary of some of the baseline characteristics for each arm of the Esprit data.

The aim of the Esprit study was to ascertain whether or not unopposed oestrogen reduces the risk of further cardiac events in postmenopausal women who survive a first myocardial infarction. Recruitment was conducted from 35 hospitals in England and Wales and involved women aged between 50 and 69 years. A total of $14,773$ hospital records were inspected out of which $11,652$ did not meet inclusion criteria for myocardial infarction. Out of the $3,121$ who met the inclusion criteria for myocardial infarction, $2,104$ were excluded mostly because of previous myocardial infarction, history (previous use) of oestrogen, or vaginal bleeding. Ultimately, the study comprised a total of 1017 women with 513 and 504 randomized to HRT treatment and placebo arms respectively and monitored over two-year period.

Previous observational studies suggested that hormone replacement therapy prevented cardiac events but there had been no randomized clinical trial on the same when the Esprit study was started in 1996. Evidence from observational studies showed that unopposed

oestrogen increased the risk of endometrial cancer although the addition of progesterone reversed this effect. Additionally, there was evidence that oestrogen reduced risk of fractures. There was also evidence of increased risk of breast cancer among the long term users (> 5 years). Organizers of the Esprit trial argued that the benefits (in terms of cardiovascular deaths) of giving unopposed oestrogen to a group at high risk of second infarction could outweigh any increased risk of endometrial cancer. The primary outcomes were reinfarction or cardiac death and all-cause mortality. Although ITT analysis of the data has been previously published (Cherry et al., 2002), the analysis took no account of compliance data. The present work utilizes compliance data and apply specialist methods in analysis to adjust for noncompliance in the active treatment arm of the the Esprit data while considering two outcomes: all-cause mortality (ACM) and myocardial reinfarction or cardiac deaths (MRCD). Assuming no carryover effects nor treatment switches, we define compliers as those participants who actually took HRT tablets up to a day before experiencing event of interest (ACM/MRCD) or end of interval/study, whichever occurred first. The rate of compliance with HRT treatment was generally poor (42% on average). Since we consider several competing methods of analysis of data, we will apply statistically designed simulation studies to evaluate the performance of these methods in terms of bias, root mean-squared error and 95% confidence intervals coverage.

Most studies designed to investigate efficacy often consider compliance with the experimental treatment only while ignoring compliance with the control. Ignoring the compliance data for the control group is likely to produce estimates that sacrifice either lack of bias or precision at the expense of the other, i.e. precision-bias tradeoff (Sommer and Zeger, 1991). For example, double-blind placebo-controlled randomized clinical trials with active treatment present complications beyond simple noncompliance (Jin and Rubin, 2008). The first complication is presented in the form of (ordinary) simple noncompliance to treatment assignment for both active and placebo arms. A second complication may manifest itself in the presence of imperfect blinding, e.g. detectable positive or negative side effects of the active treatment which may unconceal blinding. Such a scenario is possible in the Esprit data where bleeding

may breach blinding because it could make victims aware of their active HRT treatment assignment. The Esprit data also contained information on placebo compliance which may be useful in extending simple compliance to account for differential compliances in both active treatment and placebo arms. Considering the Esprit data as a two-active treatment trial and to address identification problem, we will select plausible separate predictors of compliance for HRT treatment and placebo arms and apply principal stratification framework using the Roy et al. (2008) model for survival data to adjust for noncompliance in both arms of the Esprit data. We will also apply statistically designed simulation studies using Bayesian techniques to evaluate the performance of this method in terms of bias and 95% credible intervals.

1.4 Causation in medical research

While philosophical origin of causal modelling may date as far back as the time of Aristotle, philosopher David Hume (1739) is credited as the first to posit a wholly empirical definition of causality that laid the foundation for the modern concept of causality. The application of causation in medical research is often focussed on the search for cause-effect relationships in diseases (Belin and Normand, 2009). Causal research is commonly conducted using either observational studies or randomized clinical trials. Although clinical trials are considered the gold standard in establishing causation due to their ability to eliminate systematic pretreatment differences between the treatment groups, situations arise when clinical trials are not feasible or ethical and epidemiologists often revert to observational studies to test hypothesis about causation. However, any discussion of observational studies must remain cognisant to an advise by Cochran (1965) of the need to imagine how the study would be conducted if it were possible to do it by controlled experimentation.

In epidemiological studies, John Snow[1] is recognized as the first to conduct an observational study to investigate possible causes of a cholera epidemic in London in 1853. Snow established that cholera death rates were much higher among people in areas who drew water

[1] `http://www.johnsnow.matrix.msu.edu/`

from downstream River Thames (after possible contamination with sewerage) compared to those from areas upstream who were supplied with water by the Lambeth Company.

On the other hand, James Lind[2] is credited with performing the first randomized clinical trial in 1774. As a naval officer and surgeon aboard the Salisbury, a ship carrying people afflicted with scurvy, he selected twelve patients with similar symptoms of scurvy (e.g. putrid gums, the spots and lassitude, with weakness of the knees), allocated a range of exposures to pairs of sufferers. The result was a discovery of greater improvement among those who received oranges and lemons each day, hence the well-known relationship between lack of vitamin C as a cause of scurvy. We note the fidelity with basic ideals of randomized clinical trials in Lind's experiment like using patients with similar symptoms to minimize selection bias.

Medical studies are mostly conducted to examine associations between exposures or risk factors and disease outcomes, for example, assessing the effect of HRT treatment on alleviating postmenopausal symptoms. The early examples above show the pioneering work of aggregating data and comparing groups to help identify causes. But comparison of groups may reveal an association that is not necessarily causal. Apparent association can be attributed to various factors that are not necessarily causal in nature. According to Jepsen et al. (2004), reasons for associations in medical research may be attributed to one or more of the following: bias, confounding, chance or an indication of a causal relation. We will briefly review each of these components of association (see Section 1.6). But before examining association between variables, we need to have a study designed and executed to produce relevant data.

In attempt to obtain a true causal relationship between exposure and outcome, study design and analysis phases need to be tailored to exclude other forms of association besides causation. This underscores the importance of study design and in the next section we review epidemiological study designs for cause-effect studies with focus on randomized controlled trials by outlining key design features, main classifications, their strengths and limitations as contrasted with observational studies.

[2] `http://www.jameslindlibrary.org/`

1.5 Research designs

A well-structured (objective) study design is central to the success of any medical research (Wang, 2002). The two main types of research design are experimental and observational studies (Boslaugh and Watters, 2008; van Belle et al., 2004). The quality of any research is a function of its design: only an objective design is likely to produce credible inference. A statistically designed study introduces the necessary ingredients that enhances the quality of the research data (Vandenbroeck et al., 2006). In medical research, a study design and background subject-matter knowledge is as important as the data it generates. While a poor analysis applied to a well-designed study can be salvaged by an appropriate reanalysis, no amount of sophisticated statistical manipulation can redeem a poorly designed study (Campbell et al., 2007; Robins, 2001). According to Stephenson and Babiker (2000), the range of medical studies can be broadly categorized according to either their purpose or study designs. A clinical purpose may be to determine the incidence or prevalence of disease in a population, to identify causes of disease or those at high risk of disease, to describe the natural history of disease, to prevent the onset of disease or alter the course of disease in individuals or populations. However, the present work considers medical studies classified by study design where design here is used to mean the whole range of process from contemplating, collecting, organizing to analyzing data prior to seeing outcome data (Rubin, 2008).

1.5.1 Randomized controlled clinical trials

The hallmark of an experimental study is the fact that the investigator (fully) controls the allocation or assignment of treatments or exposure (Stolberg et al., 2004). Intervention studies are often conducted to evaluate the efficacy of a treatment or therapy. Here efficacy is used in the Cochrane (1972) sense to mean a measure of the benefit derived from an intervention under the ideal conditions of an investigation by answering the question, "*does the practice do more good than harm to people who fully comply with the recommendations?*" Intervention studies include experimental studies and randomized controlled trials. While

experimental studies may include animals and plants, a distinguishing feature of clinical trials is the fact that they involves human subjects (Laine et al., 2007).

According to the World Health Organization[3], a clinical trial is any research study that assigns human participants to one or more health-related interventions with the principal objective of evaluating effects of specific health outcomes. Intervention as used here may narrowly refer to treatment or in a wider context may include any form of clinical application offered to study participants with a possible effect on their health status. Clinical trials have revolutionized the practice and conduct of medical research in terms of the way diseases are detected, prevented, treated, and early death avoided (Hackshaw, 2009). In addition to playing a central role towards establishing efficacy for any new pharmaceutical product, clinical trials are also commonly used in academia and public sector in evaluating existing interventions, programs and therapies among other non-commercial products. No matter the hue, a defining feature of good clinical trials is the rigour in scientifically designing and executing an experiment so as to generate clinical data that can be used to evaluate one or more interventions on patients' population.

Randomized controlled trials (or simply clinical trials) are considered the gold standard for efficacy in medical research (for evaluation of therapeutic or preventative interventions) because the randomization procedure ensures that any comparisons between arms of the trial reflect causal effects and are not affected by any confounding bias (Cuzick et al., 1997; Fischer-Lapp and Goetghebeur, 1999; Nagelkerke et al., 2000; Stuart, 2010). Besides providing an unbiased evaluation of the intervention by avoiding confounding from other factors, a good clinical trial minimizes variability in the estimate of treatment effect. While the goal of a superiority trial is to prove benefits of a new treatment over an existing one or placebo, non-inferiority or equivalence trials are often designed to evaluate similarity in efficacy between two/more treatments with the objective of identifying a cheaper one and/or one with less side-effects (D'Agostino Sr et al., 2003). For the present work we consider data from superiority studies.

[3]World Heath Organization. International Clinical Trials Registry Platform: `http://www.who.int.ictrp/about/details/en/index.html`

1.5.2 Classification of clinical trials

Designing a good clinical trial is premised on asking an important research question and answering it reliably (Pocock, 1983). For example, a therapeutic question is likely to be considered important if the target disease is common, the outcome major (e.g. death), the treatment is widely practicable and the question on its efficacy and safety has not already been answered reliably. Although multiple terms have been used to describe different types of randomized controlled trials, they may be best classified according to the objectives they are designed to address. According to Jadad and Enkin (2007), randomized controlled trials can be broadly classified according to either the aspects of intervention under investigation or the way in which the participants are exposed to the intervention.

Clinical trials classified according to manner of exposure and response to the intervention trials include parallel, crossover and factorial designs. The word parallel here could as well be viewed to imply simultaneous comparison. Parallel designs is the most commonly used design and is distinguishable by the fact that each group of participants is exposed to only one of the study interventions (Pocock, 1983). In contrast, crossover design refers to a study in which each of the participants are sequentially given all of the study interventions (Senn, 2002). While treatments are randomly allocated to participants in a parallel design, for a crossover design it is the order in which the participants receive each of the study interventions that is randomized. A defining feature of crossover design is the need of washout period between subsequent treatment so as to avoid contamination while evaluating efficacy. This characteristic makes crossover design more appropriate only for interventions that last a short time within the patient. It also makes the design unsuitable if any of the treatments is a cure in which case administering another treatment after the cure is self-defeating towards evaluating treatment efficacy.

Another classification of clinical trials according to exposure is factorial designs which besides permitting evaluation of two or more interventions, also allows evaluation of their combination (interaction) and against a control (Jadad and Enkin, 2007). For example, a

2×2 factorial designed to evaluate the effects of unopposed oestrogen and combined oestrogen and progestogen in reducing postmenopausal symptoms generates four sets of data for analysis: data on women who received none of the interventions, women who received unopposed oestrogen, women who received combined oestrogen and progestogen, and women who received both unopposed oestrogen and combined oestrogen and progestogen. We note that it is possible to design more complex factorial designs with multiple factors. In comparison, factorial designs provides more information than parallel designs because it includes possible interaction in addition to effects of individual treatment.

The types of clinical trials used to evaluate different aspects of interventions can take the form of explanatory or pragmatic trials. While explanatory trials are designed to establish if an intervention works under ideal circumstances, on the other hand pragmatic trials are designed to determine whether the intervention works under ordinary/usual practice. They may be distinct by design but both exploratory and pragmatic designs are complementary and most clinical trials combine both elements in investigating efficacy and effectiveness. Owing to the extensive demands for safety by regulatory bodies, exploratory and pragmatic clinical trials for pharmaceutical drugs are often rigorously implemented in four phases: I, II, III and IV (Senn, 2007). We now discuss each of these phases in the next section.

1.5.3 Phases of clinical trials

Clinical research is often conducted in a series of steps and generally classified according to the phase of development of the drug. Although all clinical phases may not be relevant to some classes of drug and individual phases may not be clearly delineated, four broad phases (I, II, III, and IV) are commonly used to define any comprehensive clinical trial (Chow and Liu, 2004). The rigour during the four phases is meant to address all aspects of a new drug in terms of safety, effectiveness and consistent quality. For example, in the United States of America, the approval of a new drug by the Federal Drug Agency (FDA) requires extensive testing and evaluation of the drug through a series of all the four phases. We discuss the four

phases below by spelling out the respective questions they are designed to address.

Phase I clinical trials is defined by the first time a new treatment is tested on humans, usually healthy volunteers (Hackshaw, 2009). Basically phase I trials are designed to answer the question: *is the treatment safe*? This phase is often small scale involving only a few participants and the trial is conducted at a single site or only at few different locations. The primary aim of a phase I clinical trial is often to investigate the new drug's metabolic and pharmacologic actions so as to find an acceptable dose in terms of safety, tolerance absorption, distribution, efficacy and side effects of a new treatment. Phase I drug trials are also referred to as dose finding trials.

Following promising results (reasonably safe in people) from a phase I trial, phase II trials are often conducted to answer the question: *does the treatment work*?. It is during phase II when the drug is administered to target patients, i.e. those suffering from the disorder for which the drug is intended. The goal of a phase II trial is to investigate the feasibility and level of activity of the drug or treatment by determining efficacy (effects on the target patients who comply with treatment) and safety in a small number of closely supervised patients. By administering varying doses of the new treatment, phase II trials are designed to provide more information about the effective dosage of a drug, the severity of the side effects, and how to manage the side effects. Compared to phase I, phase II trials usually involve more participants. Phase II trials are also referred to as safety and efficacy trials. Phase II trial produces data that can be used to design phase III trials (Hackshaw, 2009).

When data from phase II studies indicate that potential benefits from a new treatment outweighs possible hazards, in terms of safety and efficacy, an extended clinical trial (phase III) is designed to provide additional information that can be used to evaluate the overall risk-benefit relationship of the drug. Phase III trials are designed to address the question: *does the new treatment work better than the standard/placebo treatment*? Phase III trials are principally conducted to ascertain whether the drug confers clinical benefit in the disease states for which effectiveness can be claimed with an acceptable incidence and nature of adverse effects (Piantadosi, 2005). The key features of randomization and use of a compar-

ison (control) group here form important ingredients towards providing a definitive answer on whether the new treatment is better than the control, or the treatment is of similar effectiveness but with other advantages e.g. tolerable side effects. Also phase III are often double-blind trials, i.e. neither the patient nor the investigator knows which treatment is administered. Compared to phases I and II, phase III trials are generally large trials enrolling subjects at a wide variety of locations. Phase III trials are also referred to as comparative treatment efficacy trials. This report considers Esprit data which was a product of phase III trial.

Once a treatment has been approved for the general public following phases I, II and III, phase IV trials are usually conducted to answer the question: *is the treatment safe over time*? Phase IV trials involves a drug with an approved indication, formulation and route of administration. A phase IV trial is a post-marketing or surveillance study designed to extend the information developed in pre-marketing trials, i.e. obtain additional information on the risks associated with the drug/treatment, its benefits and its optimal use. The primary aim of a phase IV trial is to evaluate the long-term safety and effectiveness of a treatment. Phase IV trials are also referred to as expanded safety trials and usually involve a very large number of subjects. Overall the number of subjects in a trial usually increases with the progress in phases, i.e. while phase I usually involves fewer subjects/volunteers, phase IV trials often involves the largest number of subjects.

1.5.4 Key design features of clinical trials

In addition to inclusion and exclusion criteria protocol spelling which subjects are eligible and illegible respectively in a clinical trial, three defining components of any clinical trial include control group, randomization and blinding (Chow and Liu, 2004; Hackshaw, 2009). In the next section we provide a brief discussion of each of these components. We however point out that there is no control in a two-active arms trials where instead a standard treatment with known efficacy may be used as a comparator to the experimental treatment.

Role of control group

The use of comparable groups of study subjects in a clinical trial where one (treated) group receives the intervention while the other (control) group receives no/standard intervention allows us to estimate treatment effects. As a result of the comparison, treatment effects from clinical trials are always relative by definition, i.e. we only define efficacy of a treatment relative to another, standard treatment or placebo.

The increased attention given to subjects in a research project may influence their response and hence study findings. This so-called Hawthorne effect refers to the tendency of people to alter their behaviour when they are subject to special attention in a research setting (Hertogh et al., 2010). The use of a control group, especially in masked trials, evenly 'distributes' the Hawthorne effect on the trial findings between the study groups. The use of a control group tends to compensate for non-treatment related changes in disease status. In general, the choice of a control intervention depends on available alternative treatments, for example, it is unethical to give a patient placebo if an established/standard treatment exists because doing so denies some subjects a known health benefit.

Merits of randomization

As early as the first quarter of last Century, both Neyman (1923) and Fisher (1925) identified randomization as a cornerstone in scientific experimentation. And as discussed above (Section 1.5.1), controlled clinical trials derive their strengths from randomization procedures which plays a central role in ensuring that bias does not weaken the study results (Stolberg et al., 2004). Such biases (commonly) include human choices, beliefs or any other factors besides those being studied that can affect results of a clinical trial. For example if physicians or participants themselves choose the group, treatment assignments might be personally influenced and therefore unevenly slanted toward one side or the other resulting in selection bias. Random allocation of subjects ensures that the intervention and control groups are as similar as possible with respect to important determinants of outcome, both known and unknown. As a result randomization generates comparable intervention groups

which are alike in all important aspects except for the intervention each group receives. In the process randomization prevents selection bias and insures against accidental bias by producing comparable groups and eliminating the source of bias in treatment assignments (Altman, 1991; Mathews, 2006). Randomization also provides a basis for the statistical inference/methods used in data analysis by permitting the use of probability theory to express the likelihood of chance as a source for the difference between outcomes thus underscoring the role of statistics in seeking cause-effect relationship.

However, despite its ability to protect a trial against selection bias and facilitate comparability among groups, randomization does not guarantee that all groups at baseline are evenly matched for all known and unknown risk factors. So while randomization may effectively eliminate systematic differences between study groups, it may fail to eradicate some group differences due to chance. But despite such limitations, randomization remains a central pillar in trials by conferring one of the most important advantages that clinical trials have over observational studies where statistical adjustment can only be made from known confounders (Friedman et al., 1998). As Angrist and Pischke (2009) correctly point out, randomized trials may not be the panacea for a trouble-free trial but in principle they solve selection bias which is one of the most important problems that arises in many empirical research.

Importance of blinding

Granted every person is likely to have remarkable psychological abilities to affect her own health. Knowledge of which treatment is given may result in bias because of possible psychological effects among participants and researchers in a clinical trial who may have expectations associated with a particular treatment. Such bias may be minimized by concealing the treatment given to each subject (Hackshaw, 2009). This is often achieved by use of double-blind trials where both the subjects and any researcher involved in administering and/or assessing the treatment is unaware of the treatment given. A well executed double-blind design protects against ascertainment biases by blinding both parties to the assigned treatment. We note that the terms blinding and masking are often used interchangeably in the literature

with proponents of the later arguing use of the former projects a negative reference.

But blinding may not be practical in some situations like when dealing with an emerging epidemic where the afflicted need prompt treatment. In such situations where it is not possible to conceal (blind) trial interventions, it may be more meaningful to use objective outcome measures that are devoid of personal opinions of both participant and researcher. For example for a trial evaluating hypnotherapy for smoking cessation, we could scientifically measure serum or urinary nicotine as a marker of current smoking status instead of using a questionnaire on self-reported quitting habit which is prone to biases, e.g. reporting bias or recall bias. For a double-blind placebo control trials with an active treatment which causes side effects, a breached blinding is likely to introduce two levels confounding in the form of simple and differential confounding (Jin and Rubin, 2008).

1.5.5 Limitations of clinical trials

Properly executed randomized controlled clinical trials are capable of providing the strongest empirical evidence in any medical study by providing strong evidence of causality (Chow and Liu, 2004; Hackshaw, 2009; Mathews, 2006; Piantadosi, 2005; Pocock, 1983). However, conducting clinical trials may not be feasible in some situations due to ethical, financial, legislative, time or logistical constraints (Hackshaw, 2009; Hernán, 2004). For example, it is immoral to give a treatment that is known to be worse or to withhold a treatment that is better than standard practice or placebo. It is also generally unacceptable to conduct randomized experiments of harmful exposures like cigarette smoking. Also in a trial comparing heart transplant to medication, it is unlikely to gain approval of an ethical committee to randomly allocate heart transplants. This is simply because hearts are in short supply and the committee would most likely favour assigning them to subjects who are more likely to benefit from the transplant, rather than assigning them randomly among potential recipients.

Safety analysis is equally important (if not more) as efficacy analysis in the development of a new drug (Quan et al., 2008). While safety of a new treatment is often evaluated alongside

efficacy, it would be difficult to design a clinical trial with the exclusive objective of assessing a treatment's safety without violating the fundamental principles of the Helsinki Declaration on Ethical Principles for Medical Research involving Human Subjects which prioritizes the health of participants. Such ethical considerations based on the Helsinki Declaration may limit the exploration of scientific questions that clinical trials can address. Finally, some treatments may need long time to manifest adverse side effects and consequently clinical trials not lasting long enough may fail to detect such effects and may pose serious health risks.

From discussions above, randomization in clinical trials may be considered a necessary but not sufficient condition to address bias which could be introduced by other factors like subversion of randomization (e.g. treatment switches and noncompliance). Randomization only ensures that within the limits of chance variation, there are no systematic differences between the two groups in known and unknown prognostic factors so that any difference in outcome can be reasonably attributed to the effect of the intervention. Fidelity with protocol implementation is key to producing good randomized controlled trials, a fact that makes them the best way to estimate causal effects but still susceptible to flaws, i.e. not perfect thus *a bronze standard* (Berk, 2005). Although generally acceptable tools for causal inference, randomized controlled clinical trials often suffer from two major complications in the form of treatment noncompliance and dropouts which can only be addressed by incorporating additional assumptions (Jadad and Enkin, 2007; Mealli et al., 2004). And noncompliance to treatment allocation imparts to a clinical trials some of the characteristics of an observational study where the subjects themselves decide which group to enter (Heitjan, 1999). The next section provide a brief review of observational studies and their designs.

1.5.6 Observational studies

As discussed above, causal questions are probably best answered in an experimental setting. But reality often presents researchers with situations in which the only feasible option is data from observational studies. Incorporating simple realistic assumptions together with statisti-

cal aspects of the data can help reveal a putative causal association from observational studies (Jewell, 2003). In an observational study the investigator only observes the outcomes due to exposure on the study subjects e.g. she observes the course of a disease or the relation between risk factors and outcomes. A defining characteristic of an observational study is the fact that the investigator plays no role in assigning exposure to the study subjects. This makes observational studies vulnerable to methodological problems like selection bias. Although observational studies may not be ideal, sometimes they may be the only tool available to a researcher to evaluate effectiveness of an intervention provided confounding (see Section 1.6) can be effectively controlled. From the foregoing the first challenge would logically appear to be designing an objective observational study that may enable us make valid causal inference.

1.5.7 Designs for observational studies

Since randomized controlled experiments are the accepted gold standard in establishing causality, an observational study design can only be considered objective if it possesses the appealing features of randomized experiments, i.e. a reliable observational study should be designed to approximate randomized experiments as closely as possible. According to Rubin (2007), such a study must be designed using only baseline information to create balanced/similar subgroups in both the treated and control arms. And to ensure the objectivity of the design, the design process must be performed without any access to outcome data. This is discernable from the fact that a design informed by outcome data is prone to subjective construction that is likely to introduce other forms of bias at the design stage hence compromising the study's integrity.

Many methods have been developed to objectively create subgroups of similar treated and control units that are balanced with respect to baseline covariates. Propensity scores (see Section 1.10) is one such method where a subject's likelihood for inclusion is modelled as the probability of receiving experimental treatment conditional on baseline covariates. This makes the design objective since the balance confers on it the benefits of randomized

controlled experiments. In fact Heitjan (1999) remarks that data from such an objectively designed observational study may be conceptualized as coming from a broken randomized controlled experiment (e.g. due to noncompliance). Designs with similar features as those of randomized controlled experiments except that the probabilities of treatment assignment depend on covariates are often collectively referred to as regular designs (Rubin, 2004). Referring to them as regular designs confers on them acceptability to be analyzed and interpreted causally just like randomized controlled experiments. As Rosenbaum (2010) observes, the quality (overall integrity and validity) and strength of evidence provided by an observational study is determined largely by its design, i.e. there is no substitute to a good study design. While Rubin (2008) rightly conclude that *for objective causal inference, design trumps analysis*, Freedman et al. (2010) broadly remarks that a well-designed observational study can be very informative.

Observational studies often take different forms. Three of the most common types of observational studies include cross-sectional, cohort and case-control studies. The next section provide a brief description of each type spelling out their respective strengths and limitations.

Cross-sectional studies

A cross-sectional study identifies at a point in time individuals with a defined disease, risk factor, or other condition of interest. This is probably why Last (2001) equates cross-sectional designs to prevalence studies which allows us to estimate the burden of disease using prevalence measures. A defining feature of cross-sectional studies is that both exposure and outcome are measured simultaneously. As we will see (Section 1.6.1), this phenomenon falls short of Hill's (1965) temporality criterion that cause precedes effect, hence this is one of the reasons that makes cross-sectional studies less attractive for making causal inference in medical research. But the descriptive nature of cross-sectional studies makes them generally suitable for data exploration purposes, i.e. providing a basis for subsequent extensive cohort or case-control studies (Checkoway et al., 2004; Jepsen et al., 2004). A limitation of cross-sectional studies is that they may identify only risk factors for prevalent, rather than

incident, disease. For example, people with a prevalent disease often have a benign and long lasting form of disease that is neither fatal nor readily treated. As a result, inference based on their data may lack external validity, i.e. difficult to generalize.

Cohort studies

Cohort studies are probably the most commonly used in medical studies (Rothman et al., 2008; van Belle et al., 2004). A cohort may be considered to mean a group of individuals identified on the basis of a common experience or characteristic; the cohort is usually monitored over time from the point of assembly. Cohort studies are typically designed by specifying two or more groups in a population that are free of the disease but differ by their amount of exposure to a potential cause of the disease, for example, an epidemiological cohort study may be a follow-up of exposed and non-exposed defined groups that compares disease rates during the time covered. A major advantage of cohort studies is that they allow a researcher to investigate multiple outcomes within the same study, e.g. all-cause morality and myocardial reinfarctions in Esprit study considered for the present work.

While statisticians mostly use the term longitudinal studies to mean repeated measures, epidemiologists and social scientists on the other hand often use prospective cohort studies and panel studies respectively (Sterne and Tilling, 2002; Toh and Hernán, 2008; Xing et al., 2003). A defining feature of these studies is that they involve measurements obtained from the same individuals on repeated occasions that provide a means for assessing changes over time. Such a characteristic makes cohort studies one of the principal research strategies employed in the medical and social sciences (Goldstein, 1979; Nesslroader and Baltes, 1979). Cohort studies are more likely to reveal true efficacy than cross-sectional observational studies because of their ability to exclude time-invariant unobserved individual differences in addition to observing the temporal order of events (Diggle et al., 2002). In the presence of measured explanatory variables that are correlated with unit-specific unobservables, we can exploit the strength of cohort studies to obtain consistent estimates of causal parameters (Halaby, 2004). Randomized controlled trials as discussed above may be considered a special subset of cohort studies

owing to their prospective nature (Jadad and Enkin, 2007; Piantadosi, 2005; Rothman, 2002).

Loss to follow up is one of the common problems encountered in cohort studies. It can lead to selection bias if those lost to follow up are systematically different from those traced in terms of their exposure and disease status. An additional limitation of the cohort design is in dealing with rare diseases where we would either need a large sample or long follow up to accumulate enough cases to have sufficient power to make meaningful inferences. Cohort studies in such circumstances become very expensive and time consuming (Kuper and Gilbert, 2005) even though Hulley et al. (2001) point out that using large sample sizes in such situations minimizes chance as an explanation for the observed findings. Next we briefly discuss case-control which may be suitable to study rare diseases.

Case-control studies

A distinctive feature of case-control study is that individuals are selected according to disease or outcome status rather than exposure status. People with the disease or outcome of interest are selected as cases, and a suitable group of individuals without the disease are selected as controls. The first step of identifying the cases makes case-control designs suitable for studying rare outcomes which would otherwise require considerably large sample sizes in cohort designs, for example. In such situations, retrospective studies of rare conditions may be more efficient than prospective studies because individuals experiencing the rare outcome can be found in patient records rather than following a large number of individuals to find a few cases. In general, case-control studies are generally relatively cheaper and quick to carry out than cohort studies in this respect (Breslow, 1996; Breslow and Day, 1980).

Case-control studies have the intuitive appeal as a means of investigating etiology owing to their key (guaranteed) feature that inferences about the association between exposure and disease depend entirely on the exposure preceding the disease, a key criterion among Hill's guidelines to assess cause-effect relationship (Section 1.6.1). We can think of a case-control as a logical extension of a case series where the addition of a control group allows the frequency

of exposure in cases to be expressed relative to people who are disease free (Stephenson and Babiker, 2000). This intuitive appeal of case-control may however mask a problem of how to select the most appropriate control group. Given that the exposure has already happened, selecting controls who are more/less likely than the cases to have been exposed for reasons unrelated to the outcome of interest will result in a biased association (e.g. odds ratio) between exposure and disease. Case-control studies are commonly used as the method of choice for the prompt investigation of a suspected adverse drug reaction in addition to giving an efficient way of collecting covariate information for medical studies of rare diseases.

The fact that cases may report exposures differently from controls has the potential to introduce recall bias in case-control studies. However, this problem may be less serious if the exposure is objectively assessed (for example, height), accurately recalled (for example, age), or verified (for example, treatment received). Also selection of controls can introduce selection bias. Despite these limitations, Kuper and Gilbert (2005) point out that case-control studies can provide the same information as cohort studies in a shorter time and at lower cost if designed and executed thoughtfully.

With the study designed and executed, the treatment has probably caused some effects. But what maybe the other possible explanations for the underlying cause-effect relationship?

1.6 Reasons for association

Bias

Bias is a systematic error in a medical study that results in an incorrect estimate of the association between exposure and risk of a disease (Altman, 1991; Hennekens and Buring, 1987). Basically bias is the deviation of results from the truth. Since inaccuracies during data collection are inevitable in any medical research, it is vital for every researcher to evaluate the role of potential bias as an alternative explanation for an observed association, or lack of one, when interpreting any study result. There are many types of bias but selection and information bias

may be considered the two most common types (Rothman, 2002; Rothman et al., 2008). The propensity to select a study subject from one group and not the other gives rise to selection bias because the resulting study populations being compared are strictly not comparable, i.e. not a random sample of the population under study. Information or observation bias on the other hand occurs if non-comparable information is obtained from each study group. Selection bias error results from systematic differences in characteristics between those who take part in a study and those who do not. On the other hand information/measurement bias is systematic error arising from inaccurate measurement or classification of subjects on study variables resulting in inaccurate or incomplete data that allows introduction of false association.

Although bias may be unavoidable in practice, a researcher need to strive to evaluate the magnitude and impact of any bias on the study results. Following recommendation by Zaccai (2004), each medical research must be considered on its own merit in the context of its study population since there is no universal or simple formula/template for assessing bias. The presence of bias does not necessarily invalidate a research but valid estimation must ensure fidelity between assumptions and their associated statistical properties (Rubin, 2010).

Confounding

Granted, confounding is both an ubiquitous and an enigmatic phenomenon in medical research (Weisberg, 2010). Confounding may be described as a mixing/blurring of effects which is caused by a third factor that is associated with both the exposure and independently affects the risk of developing the disease. Formally, confounding can be defined as a distortion of the estimated effect of an exposure on an outcome, caused by the presence of an extraneous factor associated both with the exposure and the outcome (Greenland and Morgenstern, 2001; Last, 2001). Formally, a true confounder must satisfy three conditions (Hammal and Bell, 2002):

(i) It must be a risk factor for the disease in question (affect outcome),

(ii) It must be associated with the exposure under study (otherwise the variable becomes a covariate if it is only associated with disease but not the exposure) and

(iii) It must not be an intermediate step in the causal pathway between the exposure and the disease.

The criteria above are necessary but may not be sufficient (symmetric) in the sense that while all confounders must satisfy them, a variable satisfying the criteria does not necessarily qualify to be a confounder (McNamee, 2003). For example, this may be discernable from the fact that a variable satisfying all the above conditions ignores the possibility that stratifying on one variable may transform another variable into a confounder hence complicating identification of the right confounder(s).

Confounding manifests itself in many forms, for example, *confounding by indication* is a common problem in medical studies which results in selection bias (Ashby et al., 1998; Lok, 2008). Under confounding by indication, those subjects with more severe medical condition are more likely to be prescribed a particular medication. Measurement of and control for confounding may be considered two of the most difficult challenges in any medical research (Rothman et al., 2008). Depending on the magnitude and direction of association with the exposure and disease, a confounder may lead to either overestimation or underestimation of an effect. In some (extreme) circumstances, confounding may even change the apparent direction of an effect (Julious and Mullee, 1994; McNamee, 2003).

Although bias and confounding may work together, it is important to note that they may work in same or different directions towards underestimating or overestimating treatment effects. While any medical research is susceptible to bias at any of the three stages (design, execution and analysis), confounding can be addressed either at the design stage or during analysis (Boslaugh and Watters, 2008; Tai and Iliffe, 2000). At the design stage, confounding can be dealt with through different methods including randomization, restriction, and matching. On the other hand standardization, stratification and multivariate analysis are the common statistical modelling techniques used to address confounding at the analysis stage. Each of these method has its strength and limitations and none can comprehensively address confounding in isolation (Greenland et al., 1999). But whether we control confounding at the design or analysis phase, the first step should always be to identify possible confounders since as Datta (1993) correctly observe *'you cannot exclude the explanation that you have not considered'*.

Time-varying confounding

Systematic inclusion of the element of time in medical studies may help strengthen the causal inference by capturing the effects of variables that change with time. Time-varying confounding is an important and challenging aspect of medical research that is often ignored in standard statistical analysis (Bray et al., 2006). A patient's present medical state may be a product of numerous past decisions. The course of her treatment might be influenced by her past state which in turn was influenced by previous treatment decisions. A standard statistical adjustment using the patients state in the past is inadequate and results in bias since such analysis ignores information on the effect of past treatment (Lok, 2008).

A covariate is considered a time-dependent confounder if it both predicts the future treatment and the future outcome, conditional on past treatment. Formally, a covariate is defined to be a time-varying confounder for the effect of exposure on outcome provided it satisfies the following three conditions (Fewell et al., 2004; Mark and Robins, 1993; Tilling et al., 2002):

(i) Past covariate values predict current exposure,

(ii) Past exposure predicts current covariate value and

(iii) Current covariate value predicts outcome.

In a study to examine the effects of highly active antiretroviral therapy (HAART) on human immunodefficiency virus (HIV)-related death, CD4 counts in people living with HIV is an example of a time-varying confounder because low CD4 is more likely to prompt being put on HAART while low CD4 is a risk factor for AIDS and death. Another example of time-varying confounder is *diagnosis of diabetes* in a study of the effect of diet on risk of coronary heart disease (Young et al., 2010). Here *diagnosis of diabetes* is a time-dependent confounder because a diagnosis of diabetes not only affects a patient's future dietary choices but is also a risk factor for coronary heart disease. The prior diet also affects risk of diabetes in the future.

Related to time-varying confounder is the concept of time-modified confounding as introduced by Platt et al. (2009). The key difference between the two is the fact that *values* of the confounding variable change over time for time-varying confounding while the *effects* of the

confounder change over time for time-modified confounding. In essence time-modified confounding can occur with a time-invariant or time-varying covariate whereas time-varying confounding occurs only with time-varying covariates. Specifically for a time-varying covariate, time-varying confounding and time-modified confounding may therefore occur simultaneously.

In general the presence of confounders does not necessarily invalidate a medical study but instead their presence indicates the need for the researcher to quantitatively measure the confounder with which to assess its size and direction (Hammal and Bell, 2002). Confounding can be present even when an association between a covariate and either the exposure or outcome is not statistically significant. The magnitude of confounding may be unrelated to the magnitude of the p-value for the associations between a confounder and either exposure or outcome. These observations may be considered as justification for the futility of attempts to institute statistical tests for assessing confounding (Sonis, 1998).

Chance

Variations due to chance are an unavoidable consequence of sampling, but the effects can be minimized by having a study that is sufficiently large. If a study does not have a statistically significant result, a researcher is often left wondering whether this is because the sample was not big enough to detect it or there really is no difference. Because of constraints like time and cost, we often conduct studies using samples whose results we then project(extrapolate) to make inference on an entire population. Random variation in any population manifests itself when some chance factor results in the study outcomes not being representative of the ultimate true values, even if bias and confounding are non-existent. As a remedy, variations from the true values may be minimized by using large sample sizes and/or sufficiently longer studies (Zaccai, 2004).

Researchers commonly use statistics in the analysis phase to evaluate the effects of chance by quantifying the degree to which chance may account for the results observed. P-value and confidence intervals are the most commonly used accuracy statistics in this regard. For example, 95% is the most commonly used confidence interval in medical studies and a p-value

of 0.05 is often taken as the test size. For a significance test, a two-sided test size of 0.05 translates to a 95% confidence interval. We however, note that statistical significance does not necessarily translate into clinical significance. Commonsense and clinical experience are necessary considerations to separate meaningful (although statistically insignificant) association from a statistically significant association that has no clinical/logical meaning (Jepsen et al., 2004). The original ITT results of Esprit study revealed no statistically significant effects for HRT treatment compared to placebo, but this may have been due to small sample size. In attempt to investigate causal relationships, we will perform efficacy analysis to explore potential principal effects for individual strata. Such analysis may be informative and augment the knowledge gained from the ITT results.

Causation

At the heart of understanding most scientific and medical research enquiries lies the search for causation (Desousa and Murrells, 2005). Avoiding bias, controlling for confounding and ruling out chance guarantees internal validity of an association study. Provided the effects of bias, confounding and chance are ruled out then causation becomes a possible explanation for the exposure-disease association. The mantra *correlation does not imply causation* is now an accepted truism among most quantitative and qualitative researchers. Determining when a correlation is causal and when is it a result of confounding forms the basis of causal inference. As Longford (2008) points out, causation can only be established from direct intervention that excludes other sources of association.

Randomized controlled clinical trials are the ideal designs to examine causation because they often effectively address other forms of association besides causality. But observational studies may be less able to eliminate the alternative explanations of bias or confounding. Failure to account for bias, confounding and chance can substantially threaten the validity and quality of any medical study at all its phases. But the presence of these factors however, doesn't necessarily imply that a study is scientifically unacceptable (Zaccai, 2004). Instead the study must be put in its right medical context. For example reporting bias in a study may not necessarily mean that data was doctored to reflect a predetermined outcome/interpretation.

1.6.1 Bradford Hill's criteria of causation

While acknowledging that we cannot explicitly prove causality, Sir Bradford Hill (1965) outlined systematic criteria originally for epidemiological studies involving observational data for using scientific judgement that allows use of statistical association to infer causation. Hill's landmark 'checklist' is often referred to as the Bradford Hill's criteria of causation. He listed nine guidelines including strength, consistency, specificity, temporality, dose response, biological plausibility, coherence, experimental evidence and analogy. The experimental evidence criterion (a scientific demonstration of cause-effect relationship under controlled conditions) is implicit in a controlled clinical trial since it is experimental by design (and possibly by execution and analysis). While temporality criterion (exposure must precede the disease) is always satisfied for a clinical trial, dose response may not be always possible in the case of one dose. The strength criterion (causation is strengthened if the association between the exposure and disease is large) may be satisfied in trials through randomization which rules out confounding and other associated spurious relationships. The coherence criterion which ensures that the possibility of causation does not contradict established facts may be satisfied in a trial's phases II which involves subjects with target medical condition. According to Cole and Frangakis (2009), the consistency criterion (replicating results of a study using different methods at different times) is guaranteed for by design in clinical trails since the investigator controls allocation. But this is better argued in the knowledge that trials are mostly designed to address a specific objective/disease in a target population and time. Also the specificity criterion (one cause, one effect) may not apply for trials like the Esprit which may be used to investigate multiple outcomes. Finally both biological plausibility (propensity to accept a causal relationship premised on plausible biological mechanism) and analogy (the likelihood to accept comparable arguments for causation) criteria may be assumed true for controlled clinical trials although not easily provable.

We point out that although Hill's criteria were originally proposed for epidemiological studies, many modern epidemiologists (e.g. Rothman et al. 2008) strongly refute the necessity of some of the considerations. But to be fair Hill himself also acknowledged the causal

limitation in using his guidelines by concluding that "*None of my nine viewpoints can bring indispensable evidence for or against the cause-and-effect hypothesis . . .*" Although Hill's guidelines may not be watertight in proving cause-effect relationship, they may be considered a pioneer criteria that provide a basis for search of causation in modern medical research. As Phillips and Goodman (2006) points out, considering both extremes of Hill's criteria may be counterproductive towards the search of causation, i.e. while the overtly enthusiastic users are prone to over-interpret Hill's criteria, cynics who often readily dismiss the criteria as flawed are likely to under-interpret them. A middle ground may be an objective compromise where Hill's criteria are objectively applied as a basis of scientific thinking towards inferring causation in controlled randomized trials. The underlying objectivity is what probably makes Hill's criteria be considered a basis of the pragmatic intention-to-treat principle that evaluates treatment effectiveness (Freedman, 2006; Newell, 1992).

1.7 Causal inference and statistics

But what is the role of statistics in search for causation? Central to many empirical or scientific research is the search for cause-effect relationship (Dunn and Everitt, 1995) or what is often referred to as causal modelling (Heckman, 1996; Hirano and Imbens, 2001). Research in medicine and social sciences in recent years has led to phenomenal growth in causal modelling as an extension of associational models in standard statistics. Most medical research involve investigating complex multivariate relationships using data that are influenced by many factors. A primary objective of evidence-based medical research is to untangle vital causal relationships from clinical trials data which aids in illuminating intervention policies and treatments in management of diseases (Hackshaw, 2009; Rossi, 2010). The success of such research is premised upon making educated and realistic assumptions about the structure of the data generation processes in order to obtain reliable results from such studies. Modern empirical research has resulted in data explosion with causal inference as a main interest which evaluates interventions towards effective management of emerging and chronic

diseases in addition to informing policy formulation. Using structural equations and associated directed graphs to describe causal relationships, Pearl (2009b) provides a powerful defense for the necessity and legitimacy of causality as a subject of inquiry in statistics in particular and empirical research in general.

While making causal inference is a multifaceted task which no single mechanical solution may sufficiently solve. According to Heckman (2000), statistics alone cannot overcome our inability to directly measure causal relations and Pearl (2009a) concurs that causal claims cannot be exhaustively substantiated by statistical associations alone. Instead conclusions about causal effects can be considered dependent (to a greater degree) on interpretation, theorizing or assumptions that are brought into play along with empirical data or observations. Additionally it may be considered an accepted consensus among researchers that causal inference cannot be simply and directly made from empirical data, regardless of whether the data is collected through ingenious research designs or summarized by advanced statistical models (Blossfeld and Rohwer, 1997). Objectivity and the need to account for all relevant factors are key to effective causal modelling. It may be counterproductive to put too much faith in the promise of causal modelling approaches characterized by radical subjectivity. As a result causal models necessarily need to acknowledge (and account for) the complexity of real life to sustain objectivity by integrating the ingredients of intuition, logic and common sense.

The beauty of causal modelling is probably illustrated best by Wasserman (1999) in his remark that '*behind every adjusted association lurks a causal interpretation.*' Association as a product of statistical analysis makes statistics a handy and vital tool to infer causation. But Wasserman (1999) was probably too harsh with his (restrictive) edict that there are only two types of statisticians: those who do causal inference and those who lie about it. Standard statistical inference only permits conclusion about observed association which can be attributed to many reasons as discussed above (see Section 1.6). Although statistical inference enables us rule out chance as a reason of association, the procedures do not provide any information about which variable causes the other, or whether the apparent relationship between the two variables is due to one being a confounder. A statistically designed study is

likely to provide a meaningful and valid solution to answer a causal question as a product of an objective study design properly executed to produce quality data. Causal inference may then be viewed as a product of both qualitative (critical) reasoning and quantitative analysis.

The world is stochastic rather than deterministic and using probability theory allows statistics to model such non-deterministic events, i.e. probability provides a language to communicate the logic of uncertainty. Statistics is an indispensable tool in appreciating (via models) the often enormous variability of biological data encountered in medical research (Aalen and Frigessi, 2007; Campbell et al., 2007). This variability-statistics interplay is probably captured best in the definition of a statistical model by Robins and Greenland (1986) as '*a mathematical expression for a set of assumed restrictions on the possible states of nature*'. Implicit in the definition is inclusion of assumptions as an acknowledgement of the limitation to represent the often complex true state of nature. Statistical inference is essentially an exercise in using observed data to learn/infer into what we do not observe (parameters). Probability theory plays a central role in this endeavor which is also manifested in the central role of randomization in designing experiments to establish causation. Causal inference can be viewed as a combination of science and statistics where science dictates model and statistics measures the magnitude of effect, for example, in estimating a treatment effect, the effect should be both statistically concise and clinically sensible to be meaningful and useful. Perhaps by obeying the first commandment of statistical inference according to Driscoll (1977), *"Thou shalt not hunt statistical inference with a shotgun"*, an analysis must not blindly seek and interpret statistical significance as clinical significance. As a discipline, statistics is an important cog in the wheel of causal inference which makes it a necessary (but not sufficient) tool to infer causation. But as Pearl (2009a,b) rightly observed causal inference is such a broad concept it need not be limited to the probabilistic language of statistic. From the foregoing, it may be intuitively deduced that efforts to estimate causal effects probably constitutes the most important application of statistics (Weisberg, 2010).

The framework of potential outcomes allows construction of structural models (science and statistics) that facilitates valid causal inference. The next section provides a review of

this concept and the implicit assumptions that permit causal inference.

1.8 Causal modelling with counterfactuals

Counterfactuals are an intuitive concept of causal modelling which is based on the potential outcomes approach that addresses the *what if* question. Considering a two-armed trial, the counterfactual model presupposes that subjects under study have two theoretical outcomes: one that would be observed if they were subjected to the treatment of interest and the other one that would be observed if they were to be subjected to the alternative treatment. Causal estimates then become comparisons of the potential outcomes that would have been observed under different exposures of units to treatments. This appealing *contrary-to-fact* concept to causal modelling was introduced by Rubin (1974), a fact that probably led Holland (1986) to name them as Rubin causal models (RCM). However, we point out that although Rubin (1974, 1977) formalized the modern statistical framework of counterfactuals, his models builds on earlier concepts introduced by Neyman (1923) and Fisher (1926) who considered potential yields of crop varieties on different plots of land whereby the plots were randomly allocated to the varieties. Well, the modern concept of counterfactuals may be traced back to philosopher David Hume (1748) who wrote

> "we may define a cause to be an object followed by another ... where, if the first object had not been, the second never had existed."

Implicit in Hume's definition lies the concept of conditional relationship built on a thought experiment, i.e. hypothesizing on what would have happened under conditions contrary to actual conditions.

Although the terms counterfactuals and potential outcomes are often used interchangeably, the meaning may differ depending on discipline of application. The use of the term *potential outcomes* implicitly implies a prospective consideration with respect to treatment assignment. In contrast, the term counterfactuals is often used to define observed out-

come and the outcomes which would have been observed (counter-to-fact) that the subject received an alternative treatment and hence denotes the outcomes retrospectively to assignment (Hernán, 2004). According to Rubin (2006a), the prospective implication in potential outcomes makes it more appealing since at the design stage no well defined potential outcome is counterfactual while at least some of the potential outcomes are factual at the analysis stage. We note that these two definitions are similar in meaning and can be algebraically shown to be identical. In the present work we will use potential outcomes and counterfactuals interchangeably.

The three basic ingredients used to define causal effects are units, treatments and potential outcomes (Rubin, 2004). These are the basic building blocks of a design which Rubin (1975) collectively referred to as primitives. Here a unit refers to a physical object of experimentation at a particular place and time, for example, the 1017 postmenopausal women identified for our Esprit study. Treatment refers to the intervention assigned, for example, HRT medication/tablets. Associated with every unit in a group are two potential outcomes, for example, $Y(1)$ and $Y(0)$ if a woman were to receive HRT tablets and placebo respectively.

To define a general model for causal effects, let A represent the treatment variable whose effect on outcome Y we are interested in for a given population. Let i denote the unit of observation or subject so that A_i denotes the observed level of the treatment in subject i whose effect we wish to measure. Rubin causal models assumes that for each subject there exists treatment specific outcomes $Y(a)$ for each $a \in A$, where a denotes a hypothetical level of treatment. We point out that here Y may be a scalar or vector and its components may be continuous or discrete. Further let X denote measured pretreatment covariates and let $V \in X$ denote specified pretreatment covariates whose associations with outcome are of interest, i.e. all elements of V are in X but the converse may not be necessarily hold true hence we may classify those elements in X but not in V as nuisance variables.

We call $Y(a)$ a potential outcome because, until the time a decision is made about treatment, the outcome $Y(a)$ only remains potentially (latently) observable. If a subject receives treatment level a $(A = a)$ then $Y(a)$ is observed, otherwise $Y(a)$ is not ob-

served and is hence counterfactual. Let $Y_i(a)$ denote the outcome Y that we would observe in subject i if she receives level a of treatment. Causal effects are defined as comparisons of different potential outcomes $Y_i(a)$ for the same units i under different treatment levels a and a'. This is a defining difference between standard statistical inference and causal inference in the sense that while the former involves a comparison of different units under static conditions, the later goes a step further to compare the same unit that is changed by an intervention (treatment) in time, i.e. making inference on aspects of the data generation process (Pearl, 2001, 2009a). Formally we say there is a causal effect of treatment level a versus treatment level a' in subject i at the time where treatment is assigned if the outcomes differs under both conditions, i.e. $Y_i(a) \neq Y_i(a')$. The magnitude of this effect can be defined in different ways, for example, the magnitude of the individual-level causal effect of the treatment may take the form of difference

$$\tau_i = Y_i(a) - Y_i(a') \qquad \text{or the ratio} \qquad \tau_i^* = \frac{Y_i(a)}{Y_i(a')}, \tag{1.1}$$

provided the outcome $Y_i(a')$ is non-zero. We note that the choice of the measure τ_i affects the interpretability of a summary measure of individual effects, e.g. quantifying the average causal effect in terms of risk difference or risk ratio. Similarly, the choice of τ_i affects the interpretability of heterogeneity of individual effect magnitudes like causal interaction (Höfler, 2005). Here we will consider the quantity τ as the general measure of causal effect.

We however note that in reality we cannot directly compare potential outcomes $Y_i(a)$ and $Y_i(a')$ $(a \neq a')$ for any individual subject i because the outcomes $Y_i(a)$ and $Y_i(a')$ are not simultaneously observable on subject i. For example in the case of a treatment and a control, we can only observe one outcome and not both, i.e. we cannot observe the potential outcome under the treatment state for those observed in the control state (and vice versa). Consequently, we can never know the individual-level causal effects given by Equation (1.1) above. This predicament makes it impossible to make causal inference without making (generally untestable) assumptions. This dilemma is what Holland (1986) referred to as the '*fundamental problem of causal inference*'.

If we have a data set $\{Y_i, Z_i\}_{i=1}^n$, i.e. a simple random sample of size n from the population of interest where the variables Y_i and Z_i represent outcome and treatment assignment variables respectively, then individual observations on the outcome Y_i would follow the simple observation rule (Morgan, 2001; Morgan and Winship, 2007):

$$Y_i = Z_i Y_i(a) + (1 - Z_i) Y_i(a'),$$

where for a binary Z : $Z = 1$ if randomized to treatment arm and 0 otherwise and A denotes the observed level of treatment. We note that the set $\{Y, A, Z\}$ constitutes the observed data whereas the full data needed to estimate (1.1) is composed of the set $\{Y(a), Y(a'), A, Z\}$: the distribution of the observed Y_i contains only half the information contained in the distributions of the theoretical potential outcome variables. As a result, we cannot use the observed variables Y_i and Z_i to identify the population distributions of either $Y_i(a)$ or $Y_i(a')$. For a statistical solution, researchers typically focus attention on estimation of the average causal effect (ACE) often defined as

$$\text{ACE} = \bar{\tau} = E[Y(a) - Y(a')] = \bar{Y}(a) - \bar{Y}(a'), \tag{1.2}$$

where $\bar{Y}(a)$ and $\bar{Y}(a')$ are population-level means of the corresponding individual-level potential outcomes. The average causal effect quantity $\bar{\tau}$ is probably the most basic quantity of interest in almost every causal study investigating the effects of a treatment. ACE is the average gain in outcome that would be observed if a randomly selected individual were subjected to the treatment of interest instead of taking an alternative treatment. Under proper randomization (and perfect compliance), we can estimate $\bar{\tau}$ by comparing the means of Y in the two arms (Rubin, 1978), i.e. under these conditions, the difference in sample means $\bar{y}(a) - \bar{y}(a')$ can be considered as an unbiased estimate of ACE $\bar{\tau}$. This difference is the standard intention-to-treat (ITT) estimator for the effect of treatment assignment Z on the outcome Y (see ITT estimation, Section 1.11.1).

1.8.1 Key causal modelling assumptions

We now outline some of the main assumptions implicitly implied in defining counterfactuals for causal inference that enables us to estimate the ACE as given by Equation (1.2) above and its variants discussed later. First is the so-called overlapping distribution assumption: the fact that every subject must have the potential (non-zero probability) to receive treatment: $0 < \Pr(A_i = a) < 1$. Second is the consistency assumption which helps relate the observed data to the counterfactual data: $Y(a) = Y|\ a$ for the observed $A = a$, and that the subject's observed outcome is their counterfactual outcome, i.e. $Y = Y(a)$ (dropping the subscript i for simplicity). Implicit in the consistency assumption is the concept of exchangeability which makes it possible to estimate treatment effects under no unmeasured confounding and non-informative censoring assumptions (Cole and Hernán, 2008; Greenland and Robins, 1986; Lindley and Singpurwalla, 2002). In the presence of noncompliance to treatment allocation, exchangeability assumption posits that any counterfactual outcome under any treatment level a is independent of the treatment actually received A, i.e. $Y(a) \perp A$, where $\perp$ denotes statistical independence (Dawid, 1979).

In the presence of noncompliance to treatment allocation, and to retain the ideals of randomization, causal modelling is predicated on five key assumptions (Angrist et al., 1996). First is the randomization assumption (i.e. Z is randomly assigned) which posits independence between assignment to treatment arm and potential outcomes, potential treatment received and baseline covariates including pretreatment variables, i.e. $Z \perp \{Y(a), Y(a'), X\}$. However, it is important to note that $Z \perp \{Y(a), Y(a'), X\} \neq \{Z \perp Y, X\}$: treatment assignment is not independent of observed outcome since the observed outcome Y is a function both of counterfactuals and treatment assignment, i.e. the observed data are many-to-one transformation of full data (Tsiatis, 2006). This is discernable from the fact that we expect/want the distribution of Y to depend on Z if the treatment is effective. Randomization plays the key role to permit making causal inference by enabling the conceptualization of causal questions in terms of real or hypothetical (counterfactual) manipulation, a fact that probably justifies the edict '*there is no causation without manipulation*' (Holland, 1986; Rubin, 1978).

In relation to consistency assumption above, randomization of the treatment is expected to induce exchangeability hence enabling us estimate treatment efficacy. The other four assumptions of causal inference include stable unit treatment value assumption (SUTVA), exclusion restriction assumption, monotonicity assumption and non-zero average causal effect of randomization mechanism Z on treatment receipt A.

One assumption implicit in the definition of counterfactuals which even randomization by itself may not justify is what Rubin (1978) referred to as '*no interference between treatment units*' assumption. This assumption underlies what Cox (1958a) earlier referred to as the independent action of treatment assumption. But Rubin (1980) formally called it the stable unit treatment value assumption (SUTVA). Under SUTVA, we assume no interaction between subjects so that potential outcomes of a subject are independent of the exposure of all other subject, i.e. if $z = z'$ then $A_z = A_{z'}$ for subject i. Moreover if $z = z'$ and $a = a'$ then $Y_z(a) = Y_{z'}(a')$ for subject i, where z, z' and a, a' are two different vectors respectively corresponding to randomization assignment and treatment received by subject i. As a result the SUTVA assumption ensures consistency that makes it possible to identify potential outcomes by ensuring that the potential outcome of a certain treatment will be the same regardless of the treatment assignment mechanism (Rubin, 1986). But we note that despite its plausibility, SUTVA assumption may be violated in some situations like when dealing with contagious disease like influenza where treatment effects may not be independent for patients in close proximity to each other. In fact Frölich (2003) argues that due to possible interactions, departures from SUTVA can increase proportionately with increase in the number of experimental units.

The exclusion restriction assumption posits that there is no direct effect of treatment assignment (randomization) Z on the mean outcome Y except through treatment actually received A. This assumption ensures that, for example, there is no anti-placebo effect which is the tendency to feel worse after being unable to tolerate a treatment believed to be beneficial. A placebo effect on the other hand is the tendency to feel better from an inert treatment. The exclusion restriction therefore implies that

$$Y_z(a) = Y_{z'}(a) \quad \forall \quad z, z'. \tag{1.3}$$

Assuming that the outcome Y is independent of the randomization mechanism Z enables us to simplify notation so that $Y_z(a) = Y(a)$. Next is the monotonicity assumption which implies that there is no access to treatment for subjects randomized to the control arm of the trial. Generally this assumption implicitly implies that there are no defiers, i.e. people who would take the treatment opposite to what they are randomly assigned: $\Pr(A' = 0) = 1$ if A' represents placebo. Finally, the non-zero average causal effect of randomization Z on the treatment A assumption which posits that $E(A - A') \neq 0$, i.e. the treatment of interest A is assumed to have a measurable effect compared to the alternative treatment A'.

In general, causal effects may be identified under three sufficient assumptions (Cole and Frangakis, 2009): exchangeability (no unmeasured confounders), positivity (existence of a non-zero probability to receive treatment) and consistency. From the foregoing, we can argue that provided a researcher can credibly theorize that potential outcomes exist, then the advantage of using counterfactuals lies in the fact that it allows the narrowing of focus of causality to a single cause-effect relationship in addition to providing a clear definition of causal effects and quantification of treatment effects. Only the theoretical leap in which researcher hypothesizes about the existence of missing data (potential outcomes) and how that such data is related to the data guarantees causal inference, i.e. unlike standard statistical techniques, data may not necessarily speak for itself with causal models but only through the (mostly) unverified assumptions made by the researcher, i.e. the science.

1.8.2 Criticisms of counterfactual models

Generally, every branch of scientific research has its proponents and critics and counterfactual causal modelling is no exception to that truism. Causal modelling using counterfactuals has its enthusiasts and critics (Greenland, 2004). Most criticisms of the counterfactual approaches is premised on the fact that in considering causes of past events, the models invoke

distributions for events that never occurred and hence cannot be observed. As a consequence, the critics argue that some important features of these distributions remain empirically untestable. Some authors point out as a weakness the fact that some causal inferences based on counterfactuals depends entirely on untestable assumptions (Dawid, 2000). But on the other hand others like Greenland et al. (1999) consider this 'untestability' property of counterfactuals as its principal strength arguing that it is natural to employ (untestable) assumptions to address almost any important causal question which would otherwise remain unanswerable without these assumptions, i.e. there would be no progress without assumptions. And Greenland (2004) adds that counterfactuals in fact help expose the limitations of statistical inference which infers causation by addressing only the magnitude of associations and the associated average causal effects while failing to account for the causal mechanisms underlying those effects. Various assumptions like no confounding, specific statistical distributions and independence are some of the realistic (albeit often untestable) assumptions that can be considered *necessary evil* for causal inference. In fact Greenland et al. (1999) argues that besides constructively aiding precise formulation of assumptions needed to identify causal effects statistically, the counterfactual approach can also aid in developing techniques for meeting those assumptions. In addition to giving simple and clear explanation of causality, models based on counterfactuals provide a rigorous theory for clarifying estimation of treatment effects that have been adjusted for confounders which forms an integral component of causal inference (Wasserman, 2000). But as Sobel (2000) remarks, the subject of causation itself may be controversial, but the practical need for causal modelling may be considered even more overwhelming.

Critics of use of counterfactuals in causal inference often criticize the approach for including structural elements that cannot in principle be identifiable by randomized experiments which is considered the gold standard for making causal inference. Considered critically, this may be a sound opposition because, for example, we cannot determine correlation among potential outcomes since no two potential outcomes $Y(a)$ and $Y(a')$ from distinct interventions $a \neq a'$ can be observed on one unit. In fact according to Dawid (2000) the approach is

less scientific because since such correlation are unobservable, they can only be considered hypothetically. Referring to the framework 'fatalism', Dawid considers the counterfactual approach to causal modelling not only unnecessary but also cumbersome. Instead Dawid (2004) propose Bayesian causal modelling based on probability that is devoid of 'distracting' metaphysical ingredients. However, we note that Bayesian inference too involves elicitation of priors which is often considered a subjective decision by mainstream frequentist statisticians.

In his criticism with regard to basing causal inference on unobserved counterfactuals, Dawid (2000) illustrated his non-counterfactual cause-effect formulation by decomposing causal inference into two components: *effects of causes* and *causes of effects*. The distinction may be better understood by considering his two prototype sentences:

1. I have a headache. Will it be gone if I take aspirin?
2. My headache has gone. Is it because I took aspirin?

The first sentence is a common medical (hypothetical) consideration aimed at predicting the future to address a *what if* scenario. In that sentence we observe that Dawid's effects of causes is equivalent to causal effects evaluated by counterfactuals as a comparison between responses to different treatments/exposures. On the other hand, the second sentence is a retrospective examination of the past, akin to legal search of evidence. This sentence essentially defines the treatments themselves which cause the observed responses. The rhetorical decompositions may be considered a philosophical criticism of counterfactuals but a common thread connecting both arguments is the quest to empirically evaluate the impact of an intervention on a response. No matter the hue of causality, the knowledge of causes is key to understand, predict and inform the right intervention.

According to Arjas (2001), use of counterfactual models present the *causal dilemma*: the contradiction of formalizing causal models on mental/hypothetical (and subjective) constraints instead of pure scientific reasoning. According to him the contradiction lies in the fact that while the ingredients of counterfactuals are presented in terms of random variables, by its very nature this leads to conditioning on unobservables hence violating the very basic

tenets of probability theory. While Arjas (2001) considers use of mental constraints a limitation in causal modelling, according to Höfler (2005) the strength of counterfactuals lies in their ability to link metaphysical component of causality to empirical research.

Despite opposing viewpoints, counterfactuals are extensively used and appreciated in many fields of causal modelling. As Lewis (1973) correctly observed decades ago *all statements about causality can be understood as counterfactual statements*, both from a practical and commonsense point of view. In fact in its defense, counterfactual enthusiasts argue that the approach stimulates insights that not only help in recognition of shortcomings of previous methods but also promote development of new, intuitive and more generally valid methods (Wasserman, 2000). According to Wasserman (1999), the unobserved potential outcome is an intuitive causal modelling approach that can also be equivalently handled by considering it as a missing data problem, i.e. like estimation in the presence of 50% missing data. Although we do not observe them, causal inference is still possible because we can make logically sound conditional inferences about counterfactuals (Maldonado and Greenland, 2002). In the presence of complex (but ubiquitous) settings like noncompliance, Rubin (2006a) argues that causal inference is best understood using counterfactuals. Also both Geng (2003) and Kluve (2004) agree that counterfactual models provide the most precise definition and description of causal effects. The unifying framework of counterfactuals may be discerned from Pearl's (2009b) observation that most statistical approaches to causal modeling have equivalent counterfactual formulations. In general, causal modelling in the counterfactual framework has been applauded as one of the innovative intellectual developments over the past few years (Sekhon, 2009). Also as a language of inquiry, counterfactuals may not be more that a form of *scientific common sense* (Phillips and Goodman, 2006). But we note that obtaining the average causal effects often involves making some (mostly) untestable assumptions (see Section 1.8.1).

1.8.3 Alternative formulations of causal effects

Pearl (1995, 1998) developed causal diagrams as an alternative formulation of causal models. Causal diagrams is an integration of graph and probability theories which gives the formulation its alternative name of graphical models for causal diagrams. This framework of causal modelling is based on directed acyclic graphs. A directed acyclic graph (DAG) is a graph made up of nodes connected by directed edges (arrows) such that there are no cycles and no edges from a node to itself (Lauritzen, 1996, 2001). The objective is to use DAGs to represent causal relationships in terms of probabilistic dependencies between random variables. Using graph theory terminologies, each node in a directed acyclic graph represents a variable: $X \rightarrow Y$ implies variable X causes Y or equivalently X is a parent of Y, and Y is a child of X. A causal path from X to Y ($X \rightarrow Y$) means X is an ancestor of Y, and Y is a descendant of X. The relation $X \leftarrow Z \rightarrow Y$ mean Z is a common cause of both X and Y. In general DAGs convey the implicit causal assumptions through their missing arrows: an arrow linking X to Y represents potential direct causality while missing links are exclusion restrictions Pearl (1998).

Directed acyclic graphs can be used to encode conditional independence structures hence inducing conditional probability (exchangeability) which is a central concept in causal modelling by allowing us address identification problem in estimation of causal effects. This is due to the causal Markov assumption which states that if X is not a cause of Y and Y is not a cause of X, then X and Y are independent conditional on their common causes (Spirtes, 2005). In DAG language, the causal Markov assumption states that each variable is independent of its non-descendants and non-parents, conditional on its parents in the true causal graph. The concept of d-separation (Pearl, 1988) provides the correct connection between a causal DAG and probability distributions. D-separation is a Markovian representation that involves checking whether a set of vertices Z blocks all connections of a certain type between X and Y in a graph:

$$X \rightarrow Y \rightarrow Z \qquad \xrightarrow{\text{d-separation}} \qquad X \perp Z|_{\text{d-sep}}\, Y,$$

i.e. X is independent of Z conditional on Y in all distributions represented by the DAG. While causal Markov assumption allows us to deduce statistical independence given d-separation, it does not allow us to deduce statistical dependence in the absence of d-separation, what is also referred to as d-connection (Robins and Hernán, 2009).

Also by adopting do-calculus notation (Pearl, 2009b), the causal effect of an intervention Z on outcome Y can also be represented by $\Pr[Y|\ \mathrm{do}(z)]$. An equivalent representation to this in the counterfactual formulation is provided by $\Pr[(Y(z)=y]$:

$$\mathrm{ACE} = E[Y|\ \mathrm{do}(z)] - E[Y|\ \mathrm{do}(z')],$$

where the ACE is identifiable if it can be estimated from observed data.

Finally, probability theory plays a central role in statistics by allowing manipulations for valid inference and Suppes (1970) pioneered probabilistic causality which were also later expanded by Eells (1991). Recently Steyer et al. (2000a,b, 2002) proposed an alternative formulation of counterfactuals using classical probability theory. In introducing probability causal model, they however point out that they are '*not re-inventing the wheel*' by presenting a new theory but instead exploiting the wealth of probability theory to minimize ambiguity in understanding the original counterfactual formulation. They argue that the probabilistic formulation produce stochastic causal models unlike the deterministic nature of the original formulation. Arjas and Parner (2004) also propose use of Bayesian probabilistic causal models based on the general framework of marked point processes. We point out that we have outlined these alternative formulations of causal model for completeness only since the present work only applies the counterfactual framework of causal modelling.

1.9 Noncompliance and nonresponse

Complex, often unpredictable behaviours and dynamic lifestyles probably makes humans the most difficult units for experimentation. Efron (1998) perhaps captures this phenomenon

more succinctly in his observation that

> "There could not be worse experimental animals on earth than human beings; they complain, they go on vacations, they take things they are not supposed to take, they lead incredibly complicated lives, and, sometimes they do not take their medication."

A clinical trialist's task is probably complicated most by Efron's last observation which is often manifested in the form of noncompliance to treatment assignment. Many trials often suffer the twin complications of noncompliance and nonresponse (Bellamy et al., 2007; Horiuchi et al., 2007). Non-participation and drop-out may introduce bias whose magnitude depends on how strongly its determinants are related to the respective parameter of interest, e.g. hazard ratio (Höfler et al., 2005). These biased estimates may potentially threaten the validity and credibility of the resulting causal effect estimates. These two apparently similar complications can occur separately or simultaneously. The twin complications of noncompliance and missing data collectively affect causal effect estimation (Dunn et al., 2003; Frangakis and Rubin, 1999; Mealli et al., 2004; O'Malley and Normand, 2005; Peng et al., 2004; Rubin and Zell, 2010). Ideally fully observed treatment compliance would be adequate to handle missing outcome data with less complexities (Jo et al., 2010). But in real trials, noncompliance information is always incomplete since we do not observe compliance for subjects randomized to the control intervention.

It may be discerned from the foregoing that noncompliance to treatment is a common phenomenon that is likely to complicate analysis of data from randomized studies (Dunn and Goetghebeur, 2005). While it may be possible to closely monitor laboratory experiments and ensure satisfactory compliance to treatment, it is often difficult to control compliance behaviour in large scale field experiments involving intensive treatment regimes. Imperfect compliance has the net effect that the received treatment does not always agree with assigned treatment. Noncompliance may arise from a variety of reasons either related to the treatment or otherwise. Indeed evaluation of treatment effect is even complicated further if the reason for noncompliance is related to the treatment itself like stopping medication due to adverse side effects. When it is practically and ethically permissible, encouragement designs can be

incorporated to help increase the level of compliance (Frangakis et al., 2002), e.g. offering support information for pregnant women to stop smoking. While the present work considers methods of accounting for noncompliance, but because of their similarity in occurrence and effects on treatment effects estimation, we briefly describe the key features of missing data to illustrate some of the its similarity with noncompliance.

Missing data

A common feature in data are codes indicating lack of response such as don't know, refused or intelligible. Dealing with such data poses a challenge and applying any missing data technique requires establishing whether an underlying *true* value exists and, if so, whether that value is unknown (Schafer and Graham, 2002). Missing data then refers to data values that were intended to be collected but were not available for some reason. Missing (incomplete) data is a common feature of cohort studies and more often than not it is the rule rather than the exception. For example, subjects originally recruited into the study fail to participate in one or more of the subsequent waves. The two main reasons that cause subjects to be lost to follow up are death and dropout (Daniels and Hogan, 2008; Molenberghs and Kenward, 2007).

Unlike *unbalanced* studies arising by design, the missing data mechanism (the reason of the missingness, i.e. description of possible relationship between measured variables and probability of missing data) should be considered in the analysis of cohort data. The mechanism by which data is lost is critical to understanding the impact of missing data on the conclusions, model validation and making reliable inferences. There are two main features distinguishing missing data mechanisms in cohort data. The first one is monotonic missingness where some subjects drop out from a study, for example, as a result of an adverse treatment effect or lack of efficacy of the study treatment or simply the refusal of the subject to continue the study. On the other hand, some data may be missing intermittently, for example because of an illness, an invalid measurement or forgetfulness. This results in a non-monotone pattern. Cohort studies generally suffer from both of these types of missingness and more often than not the collected data are incomplete with a non-monotonic structure.

The classification proposed by Little and Rubin (2002) is based on the relationship between the mechanism leading to complete or incomplete data (the missing data process) and the mechanism controlling the actual value of the response of interest (the response process).

We adopt the terminologies of Little and Rubin (2002) in which a non-response process is said to be missing completely at random (MCAR) if the missing data mechanism is independent of both unobserved and observed data whereas we say data is missing at random (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed missing not at random (MNAR). In the context of likelihood inference, and when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, MCAR and MAR are collectively referred to as *ignorable* while the non-random MNAR process is referred to as *nonignorable*.

Let us define Y^* to denote the complete set of measurements which would be obtained were there no missing values. We then partition this into $Y^* = \{Y^{(o)}, Y^{(m)}\}$ where $Y^{(o)}$ denotes the observed measurements and $Y^{(m)}$ the lost (missing) ones. We observe the similarity in representation here with the relationship between observed Y and potential outcomes $\{Y(0), Y(1)\}$. Further, we define R as a missing indicator, i.e.

$$R_i = \begin{cases} 1 & \text{if } Y_i \in Y^{(o)} \\ 0 & \text{if } Y_i \in Y^{(m)}. \end{cases} \tag{1.4}$$

From the above formulation it can be shown that the missing data indicator R plays a similar role as a noncompliance posttreatment variable S. The effects of noncompliance can be evaluated by modelling it as missing data problem (Little and Rubin, 2000; Lumley, 2010; Rubin, 2006b; Tsiatis, 2006; Wasserman, 1999; Xie and Heitjtan, 2004). While classifying a particular missing data mechanism as MCAR, MAR or MNAR is often not clear-cut and may not be done with certainty, this classification is useful in providing a guide to the relevant analysis that would account for missing data. Apart from complications in data handling and analysis, analysis in the presence of missing data can cause loss of efficiency and/or introduce

bias (Barnard and Meng, 1999). The challenge is even greater for survival data where event (e.g. death) from competing risks, for example, may pose problems similar to nonignorable missing data mechanism (Baker, 1998). MCAR may be equated to random noncompliance in which case an ITT analysis (see Section 1.11.1) which ignores any information on compliance to treatment assignment would be adequate. However, we note that although we obtain unbiased estimates from such an analysis that excludes subjects with incomplete data (noncompliers), the estimates will not be necessarily efficient (Enders, 2010; Feudjo-Tepie et al., 2006).

In general, appropriate modelling of noncompliance and/or missingness is key to establishing the right cause-effect relationship. From all of the reasons of association discussed above (see Section 1.6), confounding is ubiquitous (Rothman et al., 2008) and it can be controlled at the design stage by randomization or at the analysis stage by suitable adjustments. The next section provides a brief review of propensity scores as one of the methods commonly used to adjust for confounding. However, we note that although propensity scores were originally developed for observational studies, the concept may be extended to compliance (often nonrandom) in controlled clinical trials.

1.10 Adjusting for confounding using propensity scores

Applying standard statistical methods (e.g. multiple and logistic regressions) to data with time-varying confounders produce biased estimates (Cox, 1958a; Robins et al., 2000). On the other hand ignoring them altogether results in residual confounding (Robins, 1989b). Also statistical adjustment for confounders may not only fail to reduce bias but even increase it (Robins, 1998b; Robins et al., 2000). Special methods have been developed to address such situations. Although most of these methods were developed for observations studies, some of them like the propensity scores have been modified for application to clinical trials data (Jo and Stuart, 2009).

The concept of propensity scores was introduced by Rosenbaum and Rubin (1983) as a

method of dealing with confounding caused by nonrandomized assignment of treatments in observational studies. The propensity score for an individual can be defined as the conditional probability of being treated given the individual's covariates (Last, 2001). Essentially the propensity scores method replaces a collection of covariates with a single covariate that is a function of the original ones. For example, for an individual i $(i=1,\ldots,n)$ with vector X_i of baseline covariates, the propensity score is the probability $\mathrm{e}(\mathrm{x}_i)$ of being treated $(Z_i=1)$ versus not being treated $(Z_i=0)$:

$$e(\mathrm{x}_i) = \Pr(Z_i = 1 | X_i = \mathrm{x}_i),$$

where Z is the assignment indicator and is assumed independent given the covariates X. The implication here is that conditioning on the propensity score, each experimental unit has the same probability of being assigned to treatment just like in a randomized experiment. Using logistic regression analysis, for example, we can estimate a propensity score (probability) that a subject would have been treated on the basis of the measured covariates. In designing a study, we discussed in Section 1.5.7 the importance of using baseline information to balance both treated and control groups. Propensity scores is one such technique that can be used to accomplish the task such that subjects in treatment and control groups with (nearly) equal propensity scores will tend to have the same distributions of the covariates used and can be considered similar. The resulting propensity score can then be used in three different ways to adjust for the uncontrolled assignment of treatments as a (Rosenbaum, 2002):

(i) matching variable,

(ii) stratification variable, and

(iii) continuous variable in a regression model, i.e. analysis of covariance (ANCOVA).

Matching and stratification are the most commonly used methods in controlling for confounding (Rubin, 2006c). Under matching method, we select comparison groups with similar background for example we match non-smokers with other non-smokers and smokers with other smokers. On the other hand, for stratification we start by first dividing the population into strata composed of homogeneous members from which we obtain the comparison groups

within each stratum. The stratification technique therefore ensures separate evaluation of association between the exposure and disease only within homogeneous categories or strata of the confounding variable.

While matching can be very laborious because of the continuous scale of the propensity scores, the Achilles heels of stratification may be the *curse of dimensionality* arising from situations when we have to estimate many parameters in the presence of many strata. Estimating many parameters from many strata some of which may be empty may result in reduced efficiency. Propensity scores can be used to effectively address this problem because it reduces all the variables into a single score. According to Klungel et al. (2004), results from such propensity scores estimation may provide clear causal interpretations without the need of further model-based adjustments. Covariate balancing using propensity scores can considerably reduce bias in observational studies (D'Agostino, 1998) such that even when a model is misspecified, the bias due to misspecification of the propensity scores may be relatively smaller compared to misspecification in standard multivariable regression model (Klungel et al., 2004).

A lot of studies have been conducted to demonstrate the advantages of propensity scores approach to causal modelling. In their seminal paper, Rosenbaum and Rubin (1983) demonstrated that propensity scores adjustment is less sensitive to assumptions about the functional form of the association of a particular covariate with the outcome (e.g., linear or quadratic).Dehejia and Wahba (1999) showed how to estimate the impact of treatment on a manpower training program observational study data using propensity scores methods in addition to assessing the efficacy of the propensity scores methodology. According to Dehejia and Wahba (1999), the propensity scores approach not only allows us to estimate the treatment effect by exploring observed information contained in variables, but even in the presence of important but unobservable covariates, applying propensity scores methods may offer useful diagnostics on the quality of the comparison group. Using simulations studies, Cepeda et al. (2003) demonstrated the superiority of propensity scores over logistic regression with respect to model robustness and empirical power. However, with respect to the empirical coverage probability, bias, and precision, they found the propensity scores method

to be superior only when the number of events per confounder was low. They found logistic regression to perform better on the criteria of bias and coverage probability when there were more events per confounder. This observation can be attributed to the fact that summarizing covariates to a single score may mask information and working with averages can potentially conceal true data characteristics, i.e. curse of averages.

1.10.1 Selecting (compliance propensity) predictors

Although propensity scores were developed to adjust for confounding from measured variables in observational data where there is no randomization in treatment assignment, the concept may be extended to model phenomenon in randomized controlled clinical trials that exhibit traits similar to confounding. Compliance to treatment allocation (often a nonrandom) is one such phenomenon. Compliance information can be a useful ingredient in estimating treatment. Using baseline covariates, Follmann (2000) obtained the compliance score (propensity to comply) from a logistic regression and used the compliance score-treatment interaction in a two stage-regression procedure to estimate effect of treatment among potential treatment compliers. However, we note that Follmann (2000) used data from the treatment group to estimate the compliance score hence ignoring placebo compliance. The present work investigates the effects of accounting for compliance with placebo allocation on causal estimates by applying the concept to compliance scores to model propensity to comply in individual arm of treatment.

Using propensity scores to select plausible predictors of compliance is an exercise in variable selection which is often fraught with many problems. The challenge often lies in the decision on which variables to select for the propensity model (Brookhart et al., 2006). In ideal situations, subject matter with a detailed understanding of treatment assignment mechanism would be a guide on choosing suitable variables but noncompliance being mostly a nonrandom phenomenon is influenced by many factors, both treatment related and extra-treatment. With no benefit of such knowledge and confronted with a large collection of base-

line covariates, the challenge translates to choosing which terms to include so as to produce a prediction model of compliance that is both well calibrated and discriminative and less optimistic. The present work will address this issue (Chapter 3) hitherto not discussed/explored much before by anyone else.

1.11 Estimating treatment effects

A principal objective in causal inference is to estimate treatment effects. In this section we present methods for estimating treatment effects starting with the intention-to-treat estimation which ignores any form of compliance information and then consider two methods (instrumental variables and CACE estimation) which are special cases of the general principal stratification framework which can be extended to adjust for noncompliance in two (and possibly multiple) arms and other complicated applications as discussed in the subsequent sections.

1.11.1 Intention-to-treat

Intention-to-treat (ITT) analysis is a strategy that compares the study groups in terms of the treatment to which they were randomly assigned, irrespective of the treatment they actually received or other trial outcomes. ITT estimate is a measure of the average causal effect of randomization Z on outcome Y, i.e. for two subjects randomized to Z and Z' arms, the ITT may be defined as

$$\text{ITT} = E[Y(Z) - Y(Z')] = E[Y(Z)] - E[Y(Z')]. \tag{1.5}$$

The ITT is variously referred to as average treatment effect or average causal effect (Freedman, 2006) Essentially the ITT is a measure of treatment effectiveness that can be used to address Cochrane (1972) question; *what would be the effect of offering treatment under ordinary conditions*? ITT analysis is performed according to assigned treatment group with

no regard to protocol deviations and participant compliance or withdrawal. Most public health policies are formulated based on the ITT measure because it reflects the effect of treatment under ordinary conditions. However, Hernán and Robins (2006) points out that such a measure is prone to vary according to local conditions. The validity of ITT estimates is premised on the fact that it retains the ideals of randomization principle that permits comparisons between the treatment arms.

However, in the presence of noncompliance, ITT analysis may not be appropriate for studies designed to evaluate treatment efficacy or equivalence. Performing ITT analysis on such studies may potentially lead to misleading inferences. For example, Sheng and Kim (2006) demonstrated that using ITT analysis for therapeutic equivalence trials increases the chance of erroneously concluding equivalence. They attributed this result to the fact that the direction and magnitude of changes in the type one error rate and power of the study depend on the patterns of noncompliance, event probabilities, the margin of equivalence and other (unobservable) factors. The assumption that ITT analysis operates under ordinary conditions is rarely satisfied in practice because noncompliance and withdrawals are more of the norm (practice) rather than exception in real life trials. The effectiveness of randomization is only equivalent to efficacy under perfect treatment compliance. But in the presence of noncompliance, ITT analysis may mask confounding (mostly due to selection bias) while comparing the treatment arms hence potentially leading to misleading conclusions mostly by biasing (underestimating) efficacy towards the null (Greenland et al., 2008; Loeys et al., 2005) and/or loss of power (Becque and White, 2008). Critics of ITT analysis often consider it an attempt to define away (trivialize) a serious epidemiological problem, especially when treatment received is the subject of scientific interest for a study (Greenland and Morgenstern, 2001).

ITT analysis is considered the gold standard in estimating treatment effects (Fischer-Lapp and Goetghebeur, 1999; Hackshaw, 2009; Pocock, 1983), i.e. ITT analysis of randomized controlled trials provide the *best unbiased inference with regard to causal knowledge* (Goetghebeur and Shapiro, 1996). However, by comparing the randomized treatment groups regardless of compliance status the method is likely to fail in revealing a treatment's true therapeutic

effects. Because it is an as-randomized analysis that ignores the compliance information, the ITT estimate is a measure of the effect of treatment randomization rather than the effect of treatment for those who actually received it. The strength of ITT lies in the fact that its estimator is protected from selection bias by randomized treatment assignment.

However, for one-active treatment trials, ITT analysis generally provide a distorted (diluted) measure of the effect of the treatment itself because it averages effects from both compliers and noncompliers, which often produces underestimated treatment efficacy (White and Pocock, 1996). The problem is compounded for a two-active treatment arms trial where the ITT may provide a biased estimator even under homogeneous treatment effects assumption (Baker and Kramer, 2005). For example, a double-blind placebo-controlled randomized clinical trials with an active treatment may present two levels of noncompliance: simple compliance due to non-adherence to treatment allocation and possible differential noncompliance due to imperfect blinding (Jin and Rubin, 2008). These two levels of noncompliance together are likely to result in biased ITT estimates. As a result there is need for methods that account for potential noncompliance in both one and two treatment arms.

1.11.2 Per-protocol and as-treated analysis

To supplement ITT analysis, per-protocol and as-treated methods are the two commonly used secondary analyses in estimating treatment efficacy (Little and Rubin, 2000; Lui, 2011; Marcus and Gibbons, 2001; White, 2005). However, these secondary analysis methods can potentially produce seriously biased estimates of treatment effects owing to selection bias because the subjects compared cannot be considered a random sample of the population under study (McNamee, 2009; Pocock and Abdalla, 1998). Per-protocol analysis compares participants who did not deviate from protocol as outlined in the research design, i.e. censors noncompliers. But such groups may be systematically different; for example, sicker participants may be more likely to take their medication than their relatively healthy counterparts and the reasons for deviation may differ between the two groups. As a result the compara-

bility is lost unless the deviations from protocol are genuinely random (Sheiner and Rubin, 1995). Also by discarding entire records of patients that violate the protocol, per-protocol analysis may be wasteful of information/data and resources (Greenland et al., 2008).

As-treated analysis on the other hand evaluates the difference in outcome between those groups classified by treatment received. As a result as-treated analysis disregards assignment in favour of receipt of treatment hence also violating randomization principle. The validity of both per-protocol and as-treated analyses may be questionable because they both lack the benefits of randomization (Imbens and Rubin, 1997). With randomization violated, confounding factors associated with switching from the assigned treatments can potentially corrupt the causal interpretation of treatment effects. Recently McNamee (2009) derived expressions for the bias in per-protocol and as-treated estimates under nonrandom noncompliance.

1.11.3 Instrumental variables estimation

Regression techniques based on ordinary least squares may be considered the powerhouse of most statistical analysis conducted to seek association between variables. However, ordinary least squares fail to account for hidden bias which is prevalent in most medical studies hence making it difficult to draw valid causal inference. Despite the elegance of propensity scores (see Section 1.10), they can only control for measured confounding. As a solution, the concept of instrumental variables was borne out of structural equation modelling and has been a popular technique in econometrics since the 1920s (Wright, 1928).

Instrumental variables were originally introduced to improve ordinary least squares by adjusting for hidden selection biases in observational studies and estimate causal treatment effects on the outcomes of interest (Heckman et al., 2006). The instrumental variable techniques have gained extensive use in medical research in estimation of treatment effects in the presence of noncompliance in observational studies (Rosenbaum, 2002). Instrumental variables provide the simplest and most robust solution to both aspects of noncompliance, i.e. in the presence of treatment dilution and/or treatment switch (Angrist, 2006). But just like

in all aspects of causal modelling, effective use of instrumental variables is premised on some untestable assumptions. For example, the strong ignorability assumption (treatment assignment is random given observed covariates) allows use of instrumental variables to produce consistent treatment effects estimates (Sobel, 2000).

The strong ignorability assumption is a common feature among econometricians where it is usually referred to as selection on observables. Heckman and Robb (1985, 1988) and Heckman and Hotz (1989) proposed a number of methods to estimate treatment effects that correct for this type of selection bias. By incorporating suitable modelling conditions, they applied the methods to longitudinal workforce data. These studies probably form the basis of modern instrumental variable techniques. We can use instrumental variables to estimate how much the variation in the treatment variable that is induced by the instrument (and only that induced variation) affects the outcome measure. The induced variation is what econometricians often refer to as exogenous variation which is said to identify the desired estimate.

An instrumental variable may be intuitively thought of as a device used to mimic statistical pseudo-randomization towards controlling for residual confounding. With that cue, actual randomization in a randomized controlled trial becomes a special case of instrumental variable (Newhouse and McClellan, 1998). A simple illustration of this is assigning people to treatment or control groups at random on the basis of tossing a fair (unbiased) coin. The outcome of the coin toss (heads or tails) becomes the instrumental variable since it induces variation in the treatment variable. The resulting quasi-experiments may sometimes have more desirable features compared to designed experiments in situations where the study conditions may be more representative of real-world settings to the extent that the use of designed experiments is less representative of the participants e.g. studies in which participants are exclusively volunteers (Luellen et al., 2005).

A key advantage of instrumental variables is the fact that they can be used to control unmeasured confounding (bias), i.e. suitably chosen instrumental variables produce consistent estimates of the average causal effects even in the presence of unmeasured confounding (Angrist et al., 1996). These results however only hold true provided the chosen instrument

satisfies the above instrumental variable properties. A research would have to follow and adhere to very strict protocol for an identified instrument to have such properties, e.g. perhaps only a double-blind clinical trial suffice:

(i) due to the fact that the trial participants are more likely to receive treatment if they were assigned to treatment (adherence to protocol, i.e. no treatment switches/defiers),

(ii) because treatment assignment affects the outcome only through the value of the treatment itself (exclusion restriction assumption) and

(iii) since the random assignment itself is a suitable instrument.

For a given outcome Y and treatment A, an instrumental variable Z may be applied using a structural two-stage regression strategy (Angrist et al., 1996):

$$Y_i = \beta_0 + \beta_1 A_i + \epsilon_i, \quad \text{where} \quad \text{Cov}(A, \epsilon) \neq 0 \quad \text{and} \tag{1.6}$$

$$A_i = \alpha_0 + \alpha_1 Z_i + \nu_i, \quad \text{where} \quad \text{Cov}(Z, \nu) = 0, \tag{1.7}$$

with the assumption $\alpha_1 \neq 0$, i.e. existence of a non-zero effect of instrument. While most studies consider dichotomous A and Z, we note that in principle A can be continuous and Z constitute multiple arms.

Ordinary least-squares estimation of equation (1.6) produce biased and inconsistent estimates of the effect of A if A_i is endogenous (i.e. endogeneity occurs when A_i is related to ϵ_i). This problem may be addressed by using the instrument variable Z_i in equation (1.7) to estimate $\hat{A}_i$ (e.g. $\hat{A}_i = \alpha_0 + \alpha_1 Z_i \equiv$ the predicted value of A_i) which is then substituted in equation (1.6) instead of the actual A_i variable. Using $\hat{A}_i$ given Z_i provide unbiased estimate of the impact of A_i on Y_i. It may be useful to think of equation (1.7) as 'purging' A of potentially confounding influence (Linden and Adams, 2006), i.e. $\text{cov}(A, \epsilon) \neq 0$ implies confounding. In effect equation, (1.7) is an expression of our lack of knowledge in which group is assigned to which treatment and the instrumental variable Z explains why one group is

treated and the other is not. By substituting Equation (1.7) in Equation (1.6) we obtain

$$\begin{aligned} Y_i &= (\beta_0 + \beta_1\alpha_0) + \beta_1\alpha_1 Z_i + (\beta_1\nu_i + \epsilon_i) \\ &= \beta_0^* + \beta_1^* Z_i + \zeta_i, \quad \text{Cov}(Z, \zeta) = 0. \end{aligned}$$

The instrumental variable (IV), slope β_1^*, may then be estimated by ordinary least squares regression techniques, i.e. by taking covariances with Z on both sides of the Equation (1.6),

$$\widehat{\text{IV}} \equiv \hat{\beta}_1^* = \frac{\text{cov}(Z, Y)}{\text{cov}(Z, A)}. \tag{1.8}$$

We observe that in the case of a perfect instrument (random assignment), a perfect relationship exists between Z and A and the parameter $\alpha_1 = 1$, in which case the ITT estimator and the instrumental variable estimator coincide, i.e. under perfect compliance. However, this ideal case is rarely the case in practice since no randomization can be perfect, instead most trials are plagued with noncompliance to treatment allocation. Using two equations to describe the problem above implicitly satisfies the important assumption that randomization Z has no effect on outcome Y other than through its effect on treatment A (since $\text{cov}(Z_i, A_i) = 0$). This is the exclusion restriction assumption (Angrist et al., 1996) as discussed previously (see Section 1.8.1). An additional assumption that $\alpha_1 \neq 0$ implies that there is no treatment defiers, i.e. no subjects taking treatment opposite to their assignment.

The problem of instrumental variable method may be compounded by small variations in instrument measure between comparison groups. For example, when variation in the likelihood of receiving a particular therapy is small between groups of patients based on an instrumental variable, differences in outcome due to this differential use of the treatment may be very small and hence difficult to assess. Furthermore the treatment effect may not be generalizable to the population of patients whose treatment status was not determined by the instrumental variable hence compromising the model's (external) validity. We note that this problem mimics a characteristic of clinical trials where estimated treatment effects may not be generalizable to a broader population. This limitation of instrumental variables may

be viewed both as a curse and a blessing. It is a blessing in the sense that identifying and using an appropriate instrumental variable improves the strength of observational studies which will compare favourably to the gold standard clinical trials. But it is a curse if we were to extrapolate results obtained by using inappropriate instrumental variables that may lead to making invalid causal inferences. In general, instrumental variables provide unbiased efficacy estimates compared to the ITT, per-protocol and as-treated approaches (Kim, 2010).

Although instrumental variables may be more useful in observational studies, we note that the present study considers a double-blind randomized controlled trial and hence we have a valid instrument (randomization), which is associated with received treatment and is most unlikely to affect the outcome (moratlity/myocardial reinfarction) other than through treatment received and shares no common causes with the outcome by virtue of randomization.

1.11.4 CACE: Complier average causal effect estimation

The complier average causal effect (CACE) estimator can be used to estimate treatment efficacy in the presence of noncompliance. We note that CACE is also referred to as local average treatment effect (LATE) in the econometrics literature (Imbens and Angrist, 1994; Wooldridge, 2010) but here we will use CACE. In their seminal paper, Angrist et al. (1996) demonstrated CACE as a valid estimate of the effect of treatment among the subpopulation who would comply with their treatment assignment, i.e. provided there are some compliers, CACE can be considered as the ITT estimate among the subgroup complying with treatment. Given a two-armed placebo controlled trial and by using the potential outcome definition of causal effects, we can classify participants into four latent (potential) complier types: compliers, always-takers, never-takers and defiers, which Frangakis and Rubin (2002) also refers to as principal strata:

- *compliers* are subjects who would adopt whatever treatment they were assigned,

- *always-takers* would always take the active treatment regardless of their treatment assignment,
- *never-takers* would always take the control treatment regardless of their treatment assignment and
- *defiers* would take the treatment opposite to what they were assigned.

If we let S_i denote the compliance type of subject i and based on a subject's joint values of potential treatment received $A_Z, Z \in \{0,1\}$ (for a two-armed trial), Angrist et al. (1996) defined four mutually exclusive subgroups that results from cross-classification of randomization and treatment received of subjects as follows:

$$S_i = \begin{cases} 1, & \text{if } (A_0, A_1) = (0,1) \\ 2, & \text{if } (A_0, A_1) = (0,0) \\ 3, & \text{if } (A_0, A_1) = (1,1) \\ 4, & \text{if } (A_0, A_1) = (1,0), \end{cases} \tag{1.9}$$

where types $1, 2, 3, 4$ refers to compliers, never-takers, always-takers and defiers respectively. These compliance groups is what Frangakis and Rubin (2002) originally referred to as basic principal strata (see next Section) while Little et al. (2009) referred to them as principal compliance stratum. We will use compliance types and strata interchangeably for the present work. For a simple case, the four unique subgroups may be collectively classified (collapsed) into two of compliers ($S=1$) and noncompliers ($S\neq 1$) (Cheng et al., 2009; Zhang, 2004).

Key to correct CACE formulation is the distinction between the definition of latent/true compliance (potential true complier under both treatments) and observed compliance (compliance under the treatment actually assigned). Since we can only observe compliance status for the assigned treatment, a subject's full compliance status is incompletely observed, i.e. unidentifiable. For example, a subject complying with the active treatment assignment may be a complier or an always-taker and a non-complier to the treatment of interest may be a never-taker or a defier. Also if a subject is assigned to the control and complies, then

that subject may be a complier or a never-taker. But a subject assigned to the control who switches to the active treatment may be either an always-taker or a defier.

Stratification on a latent variable (compliance status) S is key to addressing the identification problem outlined above. The four strata defined by Equation (1.9) above includes all subjects where the strata are defined by values of S under both (potential) treatment assignments such that a subject will belong to the same stratum no matter her treatment allocation. Because principal compliance (unlike observed compliance) is independent of the assigned treatment, it can be validly used as a stratification variable in treatment comparisons, i.e. the compliance status S is independent of the treatment allocation hence induces exchangeability. To distinguish between true and observed compliance, we observe that a subject's fidelity with allocated active treatment results in 'observed compliance' (complier or never-taker) while true compliance is (at least partially) unobserved. For a simple case, CACE may be defined as the average causal effect for the subpopulation of compliers ($S=1$). For example, for two compliers in the treatment arm who actually receive treatment, a CACE estimate may be defined as

$$\text{CACE} = E\left[Y_1(1) - Y_1(0) | S = 1\right]. \tag{1.10}$$

The method of estimating CACE is not immediately explicit because the compliance status S of individuals is generally unknown. However, we can estimate CACE from data under SUTVA and random assignment of treatment assumptions together with the additional assumptions (Angrist et al., 1996): exclusion restriction, monotonicity and at least some compliers (to ensure a nonzero denominator). As discussed previously (Section 1.8.1), the exclusion restriction assumption posits that the treatment assignment only acts on the outcome through the treatment received A. For a general setting (e.g. a/a' arms) this implies that for never-takers and always-takers, whose adopted treatment is the same regardless of which treatment is assigned (i.e. no effect of randomization), the outcome Y is the same regardless of which treatment is assigned, i.e. $Y(a) = Y(a')$ if $a = a'$. As a result these two groups play no role in determining causality because they are not relevant for comparing the

target treatments (Palmgren and Goetghebeur, 2004; Sheiner and Rubin, 1995; Sommer and Zeger, 1991). This may be discerned from the fact that both groups of compliers would not change their behavior with respect to the target treatments. While the monotonicity of treatment assignment and treatment actually received implies assuming no defiers (Imbens and Angrist, 1994), the non-zero denominator assumes that the population of interest includes some compliers. Under these assumptions, Angrist et al. (1996) derived a CACE estimate as

$$\begin{aligned} E\left[Y(a) - Y(a')\right] &= E\left[Y(a) - Y(a')|S = 1\right]\Pr(S = 1) + \\ &\qquad E\left[Y(a) - Y(a')|S \neq 1\right]\Pr(S \neq 1), \\ \text{so that} \qquad \text{ITT} &= [\omega_s \times \text{CACE}] \; + \; [(1 - \omega_s) \times \text{effect on noncompliers}], \end{aligned} \tag{1.11}$$

where ω_s is the estimated proportion of compliers in the treatment arm.

Because the exclusion restriction assumes that the effect of treatment on outcome is only through the treatment actually received, it is then reasonable to assume no treatment effect among noncompliers. And with the monotonicity assumption, the treatment effect for always-takers and never-takers can be considered to be identically zero because they cannot be induced to change treatment status through variation of treatment allocation (Frölich, 2003). Using the exclusion restriction and no defiers assumption, the last term on the right hand side of the equation above is zero:

$$E\left[Y(a) - Y(a')|S = 1\right] = \frac{E\left[Y(a) - Y(a')\right]}{\Pr(S = 1)} \qquad \text{so that} \qquad \widehat{\text{CACE}} = \frac{\widehat{\text{ITT}}}{\hat{\omega}_s}, \tag{1.12}$$

i.e. CACE can be estimated from the ITT effect divided by the proportion of compliers. We observe that the CACE estimate (1.12) provides a better efficacy estimate in the presence of noncompliance and trivially CACE and ITT are equivalent in the presence of full compliance.

There is a relationship between CACE (1.12) and the instrumental variable estimate given by Equation (1.8) (Dunn et al., 2003; Jo, 2002b; Zhang, 2004). Let the sample means $[\bar{y}(a) - \bar{y}(a')]$ be an unbiased ITT estimate (assume randomized treatment allocation). Also let $\omega(a)$ represent the proportion of participants in the treatment arm who received the new

treatment of interest and let $\omega(a')$ be the proportion of participants in the control arm who adopt the new treatment. Then Angrist et al. (1996) showed that $\omega(a)$ is an unbiased estimate of the proportion of compliers or always-takers (those who received the treatment they were assigned), and $\omega(a')$ is an unbiased estimate of the proportion of always-takers (those who adopt the treatment when assigned the control). Then $[\omega(a)-\omega(a')]$ is an unbiased estimate of the proportion of compliers, $\Pr(S=1)$ and an unbiased estimate of CACE may then be given by

$$\widehat{\text{CACE}} \equiv \text{IV} = \frac{\bar{y}(a) - \bar{y}(a')}{\omega(a) - \omega(a')}, \tag{1.13}$$

i.e. the instrumental variable (IV) is equivalent to the estimated ITT effect divided by difference in the proportions that received the new treatment in the new treatment and control arms. Equation (1.13) may provide a measure of efficacy, for example, the risk difference due to taking HRT tablets compared to placebo. We observe that the instrumental variable estimator given by Equation (1.13) is an equivalent expression (sample analogue) of the CACE estimator given by Equation (1.12) where the difference $[\omega(a)–\omega(a')]$ represents the proportion of compliers ω_s. The two quantities may be identical but the CACE estimate does not suffer many of the possible limitations implicit in instrumental variables estimation. Using potential-outcome formulation may even lead to more efficient CACE estimators compared to the instrumental variable estimator (Little and Rubin, 2000).

1.11.5 Principal stratification

We observe from CACE estimation discussed above that using a posttreatment variable to create strata induces (conditional) exchangeability which enables us to identify estimands of interest which are well-defined causal effects. Frangakis and Rubin (2002) introduced the principal stratification as a unifying framework for this approach of adjusting on post-treatment variable (i.e. intermediate response pattern) which embodies characteristics of an experimental unit and treatment, for example, compliance status S in a randomized trial encodes both efficacy and compliance behaviour. For CACE, the subdivision given by Equation

(1.10) is an example of principal stratification into four principal strata: compliers, always-takers, never-takers and defiers. We can also obtain two principal strata of compliers and noncompliers where the later is formed by combining all the other three strata. The four strata is what Frangakis and Rubin (2002) referred to as basic principal stratification while they referred to the possible permutations as principal stratifications. We note that while there may be multiple principal stratifications (e.g. four or two as defined above), there is a unique basic principal stratification with respect to a given posttreatment variable.

The principal stratification framework permits comparison of potential outcomes under different assignments within principal strata to produce principal effects. A principal strata is defined by two important properties (Frangakis and Rubin, 2002):

(i) the strata is not affected by treatment allocation and

(ii) comparison of principal effects within the strata produces well-defined causal estimates.

The definition implies that given baseline covariates X and a bivariate posttreatment variable $S \in (0,1)$, then $Z \perp \{S(0), S(1), Y(0), Y(1)|X\}$ which implies that potential outcomes are independent of the treatment assignment given the principal strata: $\{Y(0), Y(1)\} \perp Z|S(0), S(1), X$. As a result the treated and control units can be compared conditional on a principal stratum.

These properties demonstrate principal stratification as a powerful framework that is now extensively applied in a wide variety of problems in causal inference. For example, in studies to adjust for noncompliance, Roy et al. (2008) evaluated the effects of supervised exercise to promote smoking cessation, Frangakis (2004) and Frangakis et al. (2004) evaluated the effects of needle exchange programs in reducing HIV transmission among injection drug users while Zhang et al. (2009) studied the impact of job training programs on re-employment. Principal stratification has also been used to address the prevalent twin problems in clinical trials of noncompliance and nonresponse or missing data (Mealli and Rubin, 2002). For example while Dunn et al. (2003) applied principal stratification in a mental health study, Frangakis and Rubin (1999) used simulation studies in a principal stratification framework applied to survival data to address all-nothing compliance followed by missing outcome. While Barnard

et al. (2003) used Bayesian approach to address noncompliance with missing outcome data to evaluate efficacy of New York City school voucher feeding program, Jin et al. (2010) recently extended the study by using principal stratification framework to account for both missing covariates and outcomes in the presence of complicated noncompliance. Also by using simulation studies that assumed all-or-nothing noncompliance, O'Malley and Normand (2005) demonstrated the robustness of maximum likelihood estimation to departures from outcome distribution outcomes and exclusion restrictions. Egleston et al. (2010) recently applied principal stratification in sensitivity analysis to account for abstinence in the estimation of smoking cessation intervention effects.

The flexibility of the principal stratification framework has led to its extension to complex settings like *truncation by death* (Frangakis et al., 2007; Rubin, 2006a; Zhang and Rubin, 2003) where death occurs before a primary outcome of interest is recorded hence resulting in censored records/measures, i.e. the outcome of interest is not observed and is not meaningful (undefined) for subjects who die. For example, in a study of HRT effects on five-year myocardial reinfarction survival, some women may die before five years having not suffered reinfarction by the time of death, say, due to breast cancer at three years. The complication here is not limited to the undefined survival time to myocardial reinfarction for such a woman but by the fact that probability of her death from cancer may itself be affected by HRT treatment. Application of principal stratification framework here would be based on cross-classification of potential surrogates $S(a)$ and $S(a')$ and not the unobserved S hence permitting stratification on a bivariate survival outcome which is not affected by treatment receipt. Mattei and Mealli (2007) proposed a principal stratification-based model to jointly address three complications including noncompliance, missing outcomes and truncation by death. Also in attempt to show the link between missing data due to death from nonresponse and data truncation by death, Kurland et al. (2009) recently demonstrated the common origin of both analysis in factorizations of the distribution of longitudinal data and survival information. Even methods hitherto developed for observational studies like propensity scores have been modified to adopt the principal stratification framework. For example,

using baseline covariates as principal scores to address identification, Jo and Stuart (2009) recently applied propensity scores in a principal stratification framework to estimate causal effect within strata under principal ignorability assumption which posits that compliance status is independent of potential outcome given observed information, i.e. they assumed covariate information was sufficient for estimating causal effects.

Simple methods would be adequate to adjust for random noncompliance if noncompliers can be considered a random sample of the population under study (McNamee, 2009). But as discussed above, this is seldom the case owing mostly to selection bias among other factors. Noncompliance is often a nonrandom phenomenon in the sense that it is likely to be related to a subject's risk of survival for example. Accounting for such informative noncompliance is a challenge and the present work applies principal stratification in Chapter 6 to adjust for possible informative noncompliance in two arms in the Esprit data while in Chapter 8 we apply statistically designed simulation studies to evaluate performance of the method in terms of bias and 95% credible intervals. But before then we introduce in the next chapter the fundamentals of survival data and some specialist methods of adjusting for potential noncompliance in one and two treatment arms of trials investigating time-to-event outcomes.

1.12 Aims and objectives of present work

1.12.1 Broader aims

- Analyze Esprit data and adjust for noncompliance in one and two arms.
- Compare performance of statistical methods for analysing survival data in the presence of random and nonrandom noncompliance in one active-treatment arm only.
- Evaluate performance of Roy et al. (2008) method which adjusts for noncompliance in two treatment arms.

1.12.2 Specific objectives

1. Review literature on causal modelling of treatment effects in the presence of noncompliance (principal stratification).

2. Perform an in-depth analysis of the Esprit data using specialist methods that account for noncompliance in one (active) treatment arm for both all-cause mortality and myocardial reinfarction or cardiac deaths outcomes.

3. Consider how predictors of compliance should be selected by a review of general literature on model selection for prediction and select plausible separate predictors of compliance for HRT treatment and placebo arms for the Esprit study.

4. Apply the principal stratification method by Roy et al. (2008) for survival data to adjust for noncompliance in two treatment arms and apply it to analyse Esprit (adjust for noncompliance in both HRT treatment arm and placebo arm) and estimate causal effects among subpopulations characterized by different potential compliance behaviour patterns, i.e. compare causal estimates for each stratum to overall causal estimates for each arm (active and placebo arms).

5. Apply statistically designed simulation studies to evaluate performance of six statistical methods for dealing with noncompliance in the context of a randomized controlled trial comparing an active treatment and control in terms of survival when non-compliers comply for part of their treatment period.

6. Apply statistically designed simulation studies in the context of a randomized controlled trial comparing two-active treatments in terms of survival to evaluate performance of Roy et al. (2008) method, i.e. bias due to noncompliance in two treatment arms.

1.13 Outline of the thesis

This report is organized in nine chapters. In this chapter, after the introduction and motivating data, we presented a brief review of study designs focussing on key concepts of randomized controlled clinical trials, outlining their strengths and limitations and contrasting them with observational studies. After a synopsis of reasons for association, the next section provided a review of the concept of counterfactuals and its role in causal modelling, key assumptions and the use of principal stratification framework to estimate causal effects. Chapter 2 presents a review of key features of survival data and the specialist causal modelling methods for survival data which adjusts for noncompliance in one and two treatment arms. For adjustments in one arm, we consider methods for all-or-nothing and partial compliances. A brief description of methods which adjust for noncompliance in two arms is followed by a review of the principal stratification method by Roy et al. (2008) which adjusts for noncompliance in two-active treatments' trials. In Chapter 3, we review some techniques of model selection for predicting compliance to treatment assignment in each arm. After outlining the strengths and limitations of standard stepwise regression procedures, we review penalized regression methods and measures to evaluate performance of selected prediction models. The next five chapters provide results from analysis of the Esprit study and simulation studies. While Chapter 4 presents the first analysis of Esprit data using specialist methods adjusting for noncompliance in one treatment arm, Chapter 5 presents the second analysis where we develop separate prediction models of compliance for each treatment arm and Chapter 6 presents an application of the principal stratification (Roy et al. 2008 model) to analyse the Esprit data using Bayesian approach to adjust for noncompliance in both treatment arms. Chapter 7 presents a simulations study comparing the performance in terms bias, root mean squared error and 95% confidence interval coverage of the statistical methods applied in Chapter 4. Chapter 8 presents the second simulations study using Bayesian methods to evaluate the performance of the Roy et al. (2008) method applied in Chapter 6 in terms of bias due to noncompliance in two treatment arms and 95% credible intervals. Finally Chapter 9 presents discussions, novelty of present study, conclusions and possible extensions with future work.

Chapter 2

Causal Modelling of Survival Data with Noncompliance

2.1 Introduction

This chapter reviews methods for modelling survival data in the presence of noncompliance in one and two treatment arms. The first section provides definition of key features of survival data followed by an outline of the Cox (1972) proportional hazards models and accelerated failure time models. The next section presents an outline of the relationship between the two models which will be useful when comparing their performance using simulation studies in Chapter 7. We also provide an outline of the relationship between hazard ratio and relative risk which will be used in Chapter 6 for the Roy et al. (2008) method and the evaluation of its performance in Chapter 8. The next section reviews the specialist methods used to adjust for all-or-nothing and partial compliance. The final Section provides a brief literature review of methods for adjusting compliance in two treatment arms followed by a comprehensive review of the principal stratification method by Roy et al. (2008) which adjusts for noncompliance in two arms which is applied to the Esprit data in Chapter 6 and its performance evaluated via simulations in Chapter 8.

2.2 Key features of survival data

The choice of the primary endpoint may have a large impact on the design of a study and a substantive change of method of analysis. Data from clinical trials in which the outcome represents time-to-event produces what is commonly referred to as survival data. Survival data differs from Normally distributed data in three distinct ways:

(a) the survival time is strictly positive unlike normal data that can take any value (in the range $-\infty,+\infty$),

(b) survival data are generally not symmetrically distributed and often positively skewed in contrast to normal data which is symmetric and

(c) there exists censoring or partially observed outcomes unlike normal data that often have fully observed outcomes.

2.2.1 Censoring

Censoring is a defining characteristic of survival data and takes different forms: right, left or interval censoring. Right censoring, also referred to as progressive censoring, is the most common case of the three types of censoring (Collet, 2003). Right censoring can arise from numerous causes, for example,

(1) There may be no event reported on a patient by the end of the study,

(2) The patient may have been lost to follow-up,

(3) The patients withdraws from the study and

(4) Patient experiences failure from alternative (competing) risk.

Causes $2-4$ can be either informative or non-informative depending on whether the form of censoring is dependent on the unobserved failure time. Censoring, even if non-informative, must be accounted for in the analysis. Otherwise, results might be misleading and/or biased.

Right censoring are of two types, I and II, depending on what is considered a random variable: time to end of study or the number of failures (events) at the end of study. In type I, the researcher pre-specifies the number of failures ($r \leq n$, where n is the number of subjects) of interest at the beginning of the study and only terminates the study once this number is realized. The process thus sets the time to end of study as a random variable. This type of censoring commonly occurs in industrial or animal experimentation where items or animals are put on test and observed until failure (O'Quigley, 2008). In contrast, in type II right censoring, the researcher pre-specifies the time of terminating the study at which time the number of failures are determined. This number of failures ($r \leq n$) now becomes the random variable. Right censoring perhaps derives its name from the fact that the times of failure to the right (larger than T) are missing. Continuing with a study until a specified number of events are realized may lead to an open-ended random tests. While this may be appealing in indicating preference for type I right censoring, the process may not be practical due to limited resources like time and costs. Type II censoring has the significant advantage in situations where it is important to know a priori the number of failures (e.g. to achieve a pre-specified power of study). We however note that the censored (partially observed) information still provides useful information (Lindsey, 2004) so that if we let U be the observed time then we can define $U = \min(T, C)$, where T and C denote survival time and underlying censoring time respectively, i.e. U is the minimum (what comes first) between T and C. An example of right censored data can be realized when a patient leaves town before an event of interest occurs, i.e. $T > U$.

On the other hand, for left censored data a failure time is only known to be before a certain time, e.g. we may know that a particular patient died sometime before the first month but not exactly when this death occurred. Formally, a time to event T for a specific subject in a study is considered to be left censored if it is less than a censoring time C_l ($C_l < C$), i.e. the subject has already failed (realized event of interest) before being observed in the study at time C_l. All we know for such subjects is that they have experienced the event (failed) sometime before time C_l, but their exact event time is unknown. The exact lifetime T will be

known if and only if $T \geq C_l$. Thus in contrast to right censoring above, the observed time U is defined as $U = \max(T, C_l)$. An example of left censoring is when a physician discovers cancer during a routine visit for a different condition altogether hence the actual time of development of cancer is unknown and the patient sustains 'event' before initial time of observation.

Finally interval censoring is often used to reflect uncertainty as to the exact times the units failed within an interval. Lack of constant monitoring may be a cause of such type of data, e.g. if we monitor patients after every three months and we realize that a particular patient was alive at the last evaluation but dead by the next evaluation then the only information we have is that she failed in the 3-6 months interval of time. Probably due to this nature of periodic monitoring is what makes engineers in reliability studies refer to data that is interval censored as inspection data. An example of interval censoring is data obtained when a patient after surgery sustains event between two scheduled post-operation visits, i.e. the event is only known to have occurred in between the two visits. In general, right censored data is more prevalent in survival analysis of medical studies than left and interval censoring (Lee and Go, 1997). Data from the Esprit study were right censored (type II).

According to Marubini and Valsecchi (1995), the principal aims of survival analysis in the biomedical field include:

(i) Estimation of failure time distributions, i.e. summarize the distribution of survival times,

(ii) Compare the distributions of survival times among competing treatments so as to find the best treatment and

(iii) Prognostic evaluation of different variables, i.e. explore and understand the relationship between survival time and important covariates.

2.2.2 Survival function and hazard rate

The two functions extensively used to describe survival times and which are of central importance in the analysis of survival data are the survival function and the hazard functions

(Everitt and Pickles, 1999; O'Quigley, 2008). The survival function $S(t)$ describes the probability of an individual surviving beyond time t, i.e., the probability of experiencing the event of interest after time t. Mathematically we can define $S(t) = \Pr(T > t) = 1 - F(t)$.

The time scale t is often set so that $t = 0$ refers to the beginning of follow-up. $S(t)$ has the three properties:

(a) $S(0) = 1$ which implies that a patient is presumed alive at the beginning of the study,

(b) $\lim_{t\to\infty} S(t) = 0$, i.e. given a sufficiently long study period all subjects studied will definitely experience the event (fail) or loosely speaking no patient lives forever and

(c) $S(t)$ is a monotonic, non-increasing, and nonnegative function implying that the probability of survival diminishes as the study progresses.

The probability density function $f(t)$ is related to $S(t)$ by

$$S(t) = 1 - F(t) = \int_t^\infty f(u)\ du,$$

where by definition in statistical theory $F(t) = \Pr(T \leq t) = \int_0^t f(u)\ du$ and $f(t) = F'(t)$. A $S(t)$-time plot produces the survival curve which begins at $S(0) = 1$ and decreases to 0 as t increases to infinity.

An important aspect of the survival distribution is the hazard function h(t) which is also known as the hazard rate or intensity or force of mortality and is defined as

$$\mathrm{h(t)} = \lim_{\delta t\to 0} \frac{\Pr(t \leq T < t + \delta t | T \geq t)}{\delta t}, \tag{2.1}$$

i.e. h(t) is the *instantaneous* (failure) rate of developing the event of interest in an arbitrarily short time interval δt, provided the subject is still at risk at time t (has not fallen ill before time t). From the above Equation (2.1), $\mathrm{h(t)}\delta t$ can be thought of as the approximate probability of an individual who has not experienced the event by time t experiencing the event in the next instant following t. However, we note that although $\mathrm{h(t)}\delta t$ is a probability,

technically the hazard rate h(t) is not a probability in the proper sense because it may take on values greater than one, i.e. it has no upper bound.

Alternatively, the hazard rate can be interpreted as the average number of events in a unit interval of time, hence why it is also commonly referred to as intensity. The incidence rate then validly approximates the hazard rate that assumes piecewise constant rate, a scenario that is realistic for short periods. In such applications incidence rates can be regarded as estimates of a limiting (theoretical) hazard rate h(t), which epidemiologists often refer to as the incidence intensity or force of morbidity (Rothman et al., 2008). Strictly speaking however, incidence and hazard rates do not always coincide.

For a continuous random variable T the following relationships between S(t) and h(t) holds:

$$\mathrm{h}(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t), \quad \text{so that} \quad S(t) = \exp\left[- \int_0^t \mathrm{h}(u)\, du \right].$$

For all the three aims of survival analysis listed above, interest often lies in estimating the distribution of failure time T, the hazard h(t) and modelling the relationship between them and a subset of relevant covariates X. Inference from survival data analysis often assumes noninformative censoring, i.e. T is independent of any mechanism which causes the individual's survival time to be censored at C. This assumption is necessary so as to make the distribution of T become identifiable from the distribution of the observables. Practically the assumption implies that the survival experience of censored individuals can be estimated by using data on the uncensored individuals. This implies that if we consider a group of all patients with same values of relevant prognostic factors, then a patient whose survival is censored at time c is assumed representative of all other patients in that group who have survived until that time.

2.3 Accelerated failure time and proportional hazards models

Survival data is commonly modelled using either accelerated failure time (AFT) models or proportional hazards (PH) models (Cleves et al., 2008; Collet, 2003). While PH models are popular in medical research by comparing treatments using a hazard ratio, AFT models estimate a factor by which a harmful/beneficial treatment decelerates/accelerates life of those taking them. In the next section, we briefly describe both PH and AFT models and provide a skeletal outline of their equivalence for special cases for Weibull distribution.

2.3.1 Accelerated failure time models

Since time is only meaningfully considered positive, most regression models to a function of the survival time assumes that the regression model is applied to the logarithm of the time. These models form the class commonly referred to as accelerated failure time (AFT) models. The nomenclature is probably due to the fact that the effect of a covariate is to multiply the time-scale on which events occur, i.e. the models mimics a clock running either slower or faster than usual (to failure). In health sciences, AFT may be usefully interpreted in terms of the speed of progression of a disease (Everitt and Pickles, 1999). The simplest AFT model assumes that the survival times are exponential random variables (i.e. constant survival rate). However, other parametric models such as the Gamma, Log-normal and more general Weibull distributions can also be used.

An AFT model comparing two treatments can algebraically be represented by

$$S_1(t) = S_0\left(\varphi_i t\right),$$

where φ_i often assume the form $\exp(X'\beta)$. The parameter φ_i is commonly interpreted as the accelerating factor for subject i compared with baseline patient group. Alternatively, considering treatment A as the only (time-invariant) covariate, a simple AFT may be represented as

$$\log T = -\beta' A + \epsilon, \tag{2.2}$$

where T denotes survival time, A the treatment indicator, β regression parameter to be estimated and ϵ is a random variable with a (parametric) distribution not dependent on A, e.g. ϵ often taken to assume an extreme value distribution: $f(\epsilon) = \exp[\epsilon - \exp(\epsilon)]$. We note that for Gaussian errors with no censoring, the AFT model (2.2) reduces to a linear regression of log-survival time. On the other hand, a semiparametric variant of the model (2.2) would not specify a parametric family for the distribution of ϵ. It may be worth noting the fact that while parametric AFT models can be fitted using maximum likelihood estimation in a number of statistical packages, the semiparametric model requires more advanced algorithms (Lin and Ying, 1995; Vittinghoff et al., 2005).

Interpretation of regression parameters for AFT models is composed of a dummy two-arm treatment. For example, $\exp(\beta)$ can be interpreted as the multiplicative factor by which the underlying survival time is multiplied such that if $\exp(\beta) = 3$, then on average patients on new a treatment ($A = 1$) would take a duration three times shorter to experience the event of interest (fail) compared to patients on standard/placebo treatment ($A = 0$). But this interpretation may also be considered a limitation of the AFT model in cases where treatment levels are continuous or even a mixture of both dummy and continuous scales. A solution to this limitation is using the Cox PH model that can also account for time-varying covariates.

2.3.2 Proportional hazard models

The Cox (1972) proportional hazard (PH) model is probably the most widely used method in survival data modelling (Lee and Go, 1997; O'Quigley, 2008; Pan, 2001). A defining feature of this model is its demonstration that we can estimate the relationship between the hazard rate and explanatory variables without having to make any assumptions about the shape of the baseline hazard function unlike the parametric models. Since the distribution of survival times is not necessarily specified, the Cox model is an example of a semiparametric model.

The Cox's PH models the hazard of event (e.g. death) as a function of covariates. In the Cox PH regression model for survival data, t measures a convenient time axis and the

hazard h(t) at time t. A Cox PH model to evaluate the effect of treatment A is given by

$$h(t|A) = \mathrm{h}_0(t)\ \exp(\beta A), \tag{2.3}$$

where $h(t|A)$ is the hazard of failing (e.g. death at time t given the treatment A) and $\mathrm{h}_0(t)$ is the baseline hazard, i.e. the underlying hazard in the presence of no treatment A. An attractive aspect of the Cox model is the fact that $\mathrm{h}_0(t)$ need not be specified.

2.3.3 Relationship between PH models and AFT Models

As discussed above the main difference between PH and AFT models is effect of treatment on hazards and time respectively: whereas a PH model assumes a multiplicative effect of a covariate on the hazard, an AFT model assumes a multiplicative effect of a covariate on time. In general the parameters of the PH and AFT models may not be comparable since they are based on different scales, i.e. hazard ratio for PH models and time ratio for AFT models. An important difference is that while the baseline hazard function for a parametric PH model need not be completely specified, the baseline hazard function for AFT model needs to have a complete parametric specification.

A relationship between PH and AFT would be useful for comparison of performance between methods which provide results in terms of either formulation. For constant covariates, the PH and AFT models coincide only for Weibull distribution (Collet, 2003; Cox and Oakes, 1984): there is a direct correspondence between Weibull PH and AFT in terms of hazard ratio and time ratio parameters respectively:

$$\begin{aligned} \exp(\beta) = \varphi^p = [\exp(-\vartheta)]^p\,, \qquad & (2.4) \\ \text{so that} \qquad \beta = -p\ \log\varphi, & \end{aligned}$$

i.e. the hazard ratio is equivalent to the negative logarithm of the accelerated life parameter hence for the exponential distribution (p=1), $\beta = -\log\varphi$.

2.3.4 Hazard ratio-relative risk relationship

When applying the Roy et al. (2008) model for survival data (Chapters 6 and 8), we will use relative risks (RR) to approximate hazard ratios (HR). Although HRs and RRs are often used for time-to-event and binary outcomes respectively, for short follow-up and small event rates, they can be shown to be algebraic approximations of each other (Symons and Moore, 2002; Wang, 2002). The proportional hazards model given by Equation (2.3) provides the HR $\exp(\beta)$ as a measure of treatment effect on the treated compared to the untreated. We can also define the probability of experiencing an event (e.g. death) in the time interval $[0, T]$ for the treated and untreated group respectively as

$$P_1 = 1 - \exp[-\mathrm{H(T)}e^{\beta}] \qquad \text{and} \qquad P_0 = 1 - \exp[-\mathrm{H(T)}],$$

where $\mathrm{H}(T) = \int_0^T h(t)dt$ is the the cumulative hazard function. The RR may then be defined as

$$\mathrm{RR} = \frac{P_1}{P_0} = \frac{1 - \exp[-\mathrm{H(T)}e^{\beta}]}{1 - \exp[-\mathrm{H(T)}]} = \frac{1 - (1 - P_0)^{e^{\beta}}}{1 - (1 - P_0)} \approxeq \frac{1 - [1 - \mathrm{H(T)}e^{\beta}]}{1 - [1 - \mathrm{H(T)}]} = \exp(\beta) = \mathrm{HR}.$$

Given that every individual is prone to react differently to any intervention, such heterogeneity in survival data may be captured using frailty models as described briefly in the next section.

2.4 Heterogeneity and frailty

Assuming constant effects of treatment among subjects may be unrealistic in a heterogeneous population. An individual's genetical/biological and physiological composition is likely to influence her way to uniquely respond to an intervention. Heterogeneity of treatment effects reflects patient diversity in risk of disease, responsiveness to treatment, vulnerability to adverse effects, and utility for different outcomes. Formally, heterogeneity of treatment effects refers to the differential response to the same treatment by different patients (Kravitz et al., 2004), i.e. a condition where some patients are more likely to experience a larger number of events

than others due to some unknown, unmeasured or unmeasurable reason. From the foregoing, unbiased efficacy estimation would require an effective way to account for such individual variation among a group randomized to an intervention. Vaupel et al. (1979) introduced the concept of frailties in survival data as a convenient way to to explicitly account for possible unobserved heterogeneity (and correlations) among failure times. For the proportional hazards model, frailty can be defined as a latent random effect that enters multiplicatively on the hazard function h(t) (Cleves et al., 2008), i.e. if we let each individual to have a frailty η: individuals with a high value of η implies a high frailty and vice versa, then conditional on the frailty, the hazard rate of an individual is often assumed to take the form

$$h_i(t|X,\eta) = h_0(t)\exp(X_i^{'}\beta + \eta_i)$$

or equivalently

$$h_i(t|X,\eta) = \eta_i^{'}\ h_0(t)\exp(X_i^{'}\beta).$$

The frailty approach provides a convenient statistical modelling technique that enables us to model possible heterogeneity (caused by unmeasured covariates) and/or address deviations from proportional hazards assumption in models for survival data (Keiding et al., 1997; Wienke, 2010). Considering a population as homogenous may not only be unrepresentative (or unrealistic) but also analysis of such data is likely to produce biased efficacy estimates (O'Quigley, 2008). Introducing frailty amounts to considering a heterogeneous sample, i.e. subjects with different hazards.

Frailties are often introduced into a model as unobservable positive quantities which assume specified distribution with standardized mean and variance to be estimated from data, for example, from the frailty model above, $\eta_i \sim N(0,\sigma^2)$. Since frailty cannot be negative, the Gamma (along with log-normal) distribution is one of the most commonly used frailty distribution (Clayton, 1978; Vaupel et al., 1979). Useful frailty distributions should have an explicit Laplace transform (Aalen et al., 2008) and from a computational and analytic perspective, the Gamma frailty distribution fits well with failure time data because it is easy

to derive the closed form expressions of its various survival functions due to the simplicity of the Laplace transform (O'Quigley, 2008). We will use a Gamma-distributed random variable for hazard rates in our simulation studies (Chapters 7 and 8) to reflect possible implicit heterogeneous treatment effects among subjects.

2.5 All-or-nothing compliance methods

Noncompliance to treatment allocations often takes one of two forms (White, 2005): all-or-nothing or partial compliance. Under all-or-nothing compliance, a subject is assumed to have discontinued medication immediately after randomization or at a time of experiencing a pre-specified event of interest. Partial compliance on the other hand utilizes the time till treatment discontinuation or switching and hence accounting for attendant benefits derived from medication up to stoppage time. We note that all-or-nothing compliance may fail to utilize time till stoppage of medication/intervention and any information of possible corresponding benefits of treatment taken till then. In this section we provide brief reviews of three specialist methods of analysing survival data in the presence of nonrandom noncompliance in one (active) arm. While the structural proportional hazards (C-Prophet) method adjusts for all-or-nothing compliance, the Causal Accelerated Life Model (CALM) and the Causal Hazard ratio Adjustment Regression Model (CHARM) methods adjust for partial compliance.

2.5.1 C-Prophet: Structural (causal) proportional hazard models

The ITT method is considered the gold standard due to its validity under the null. As discussed above in section (2.3.2), for survival data if $Z = 1$ denotes the active experiment arm, then the Cox proportional hazard model

$$\mathrm{h}(t|Z) = \mathrm{h}_0(t)\exp(\beta Z),$$

is commonly used to estimate treatment efficacy. This approach is valid but we note that

when experimental treatment has no effect ($\beta = 0$) then the survival distributions coincide on both randomized arms. However, in the presence of noncompliance, β may fail to reveal the true treatment effect by mixing (diluting) the effect on compliers with the absence of effect on non-compliers.

All-or-nothing compliance is a common method that evaluates treatment effects by introducing a binary treatment indicator U which indicates whether a subject received treatment ($U=1$) or nothing ($U=0$). The Cox model would then become

$$\mathrm{h}(t|U) = \mathrm{h}_0(t)\exp(\beta U),$$

where, as before $\mathrm{h}(t|U)$ is the hazard rate for failure at time t given the exposure. However, when compliance is selective (nonrandom - individuals who comply are prognostically different from those who do not) then the parameter β may carry no causal interpretation.

To estimate treatment efficacy in the presence of noncompliance, Loeys and Goetghebeur (2003) proposed the Complier PROPortional Hazards Effect of Treatment (C-Prophet) as a structural proportional hazards method to model all-or-nothing compliance. Besides assuming uninformative (independent) censoring, C-Prophet satisfies the modelling assumptions as outlined earlier:

(a) SUTVA: treatment received U, survival time T and randomization Z are assumed independent and identically distributed (iid) so that the potential outcomes for each subject are unrelated to treatment or outcome experienced by other individuals,

(b) Randomization assumption: $(U, T) \perp Z$

(c) No access to treatment on the control arm: $\Pr(U=0)=1$ for all subjects and

(d) Exclusion restriction: $\Pr(T^1 > t|U = 0) = \Pr(T^0 > t|U = 0)$, where T^0 and T^1 denote counterfactual survival times in the control and active arms respectively.

To estimate the causal effect of treatment actually received, Loeys and Goetghebeur (2003) proposed the causal (structural) C-Prophet model

$$\mathrm{h}(t|Z_i = 1, U_i = u) = \mathrm{h}(t|Z_i = 0, U_i = u)\exp(\psi_0 u), \tag{2.5}$$

where U_i is the potential all-or-nothing treatment for subject i, i.e. the treatment that would have been observed had subject i been randomized to the active treatment arm. But we note that U_i is only observed on the experimental arm while it is latent on the control arm. The model assumes noninformative censoring conditional on compliance for the experimental arm but censoring is assumed unconditionally noninformative in the control arm (Goetghebeur and Loeys, 2003). While the first part is a common assumption in 'associational' survival models, the second part of unconditional noninformative censoring for the control arm component fits counterfactual definition for causal interpretation.

The C-Prophet estimate (log hazard ratio ψ_0 in model (2.5)) is a measure that compares survival under experimental and potential control conditions in the treatable subgroup $\{U_i=1\}$. By comparing survival in the two arms conditional on all-or-nothing compliance i.e. contrasting

$$\Pr(T^0 > t|U = 1) = \Pr(T_i > t|Z = 0, U = 1)$$

with

$$\Pr(T^1 > t|U = 1) = \Pr(T_i > t|Z = 1, U = 1),$$

$\exp(\psi_0)$ provides the causal proportional hazard effect estimate among the compliers in the treatment arm (subpopulation). A negative and positive ψ_0 respectively implies a beneficial and harmful effects of the active treatment in the treatable subset. C-Prophet estimation is predicated on the exclusion restriction assumption (Goetghebeur and Loeys, 2003), i.e. in the subgroup $\{U_i = 0\}$ that would not have been treated when assigned to experimental treatment, no effect of assignment on survival is assumed. The C-Prophet model (2.5) may be considered a special case of the G-estimation with two level of treatments (G-estimation is a general method for adjusting for time-varying confounding, (Robins et al., 1992)).

2.6 Partial Compliance Methods

2.6.1 CALM: Causal Accelerated Life Models

Robins and Tsiatis (1991) introduced the structural accelerated failure time model for survival data which accounts for time-dependent departures from randomized treatment in either arm and relates each observed event time to a potential event time that would have been observed if the control treatment had been given throughout the trial.

Let T, Z and A denote survival time, randomization arm and treatment receipt indicator respectively. Then the difference $T^{Z=1} - T^{Z=0}$ would provide an effective measure of treatment allocation. But the fundamental problem in randomized trials posits that we cannot jointly observe both potential survival times for one individual (Holland, 1986). However, under suitable assumptions a statistical solution may be provided by the average causal effect (ACE) of treatment allocation that is obtained by taking expectations over the whole population:

$$\text{ACE} = E[T^{Z=1}] - E[T^{Z=0}],$$

where ACE provides the ITT estimator (i.e. ACE=ITT) since the expectations are estimable from the two treatment arms by the randomization principle. Under ideal conditions, estimating efficacy of active treatment would involve comparing survival times for subjects complying with their active treatment assignment to those randomized to receive placebo (never receive active treatment), i.e. comparing outcomes

$$T[Z=1, A(t)=1] \qquad \text{with} \qquad T[Z=0, A(t)=0] \quad \forall \quad t > 0.$$

But because $A(t)$ is a post-randomization variable (cannot be assigned or controlled by study design), we can observe only one survival time $T = T[Z=z, A(t)=a]$, where z and a are the assigned treatment and the particular treatment actually received respectively. Modelling assumptions are required to help identify the efficacy parameter from the observed data. Robins and Tsiatis (1991) defined the potential treatment-free survival time $W = T[Z, A(t)=0]$, at time $t > 0$, as the survival time from enrollment to the study if the control treatment had

been given (or active treatment withheld) throughout the trial. This treatment-free survival time (placebo prognosis) is linked to the observed survival time and to the observed treatment received through a parametric Causal Accelerated Life Model (CALM):

$$W_i(\varphi) = \int_0^{T_i} \exp\left[\varphi A_i(s)\right] ds, \tag{2.6}$$

where φ measures the treatment's efficacy. The CALM model (2.6) provides a relation between the (potential) placebo prognosis W to the observed event time T_i if subject i were to receive active treatment according to the observed process $A_i(t)$. The quantity $\exp(\varphi)$ can be interpreted as the relative increase/decrease in survival if a subject was always on control treatment compared to if always on active treatment. We will use the alternative $\exp(-\varphi)$ which provides an intuitive interpretation of causal survival time ratio (Sterne and Tilling, 2002). We note that provided the assumption of no treatment access to the control holds then $T_i[Z_i=0]=W_i$: we would then observe W_i in the placebo arm in the presence of no censoring. On the other hand in the active treatment arm we only observe W_i for uncensored subjects who would never receive active treatment. Besides censoring as a defining characteristic of survival data, a second level of truncation often occurs at the end of a study when a trial is ended. However, such an administrative censoring at the end of follow-up may introduce bias (in estimating $W_i(\varphi)$) and White et al. (2002) suggested a user-specified recensoring time (e.g length of study duration) to address this phenomenon.

The CALM model given by Equation (2.6) is also referred to as a *rank preserving structural nested failure time model*(Robins, 1994, 1998c; Robins and Tsiatis, 1991). It is structural in the sense that the parameter φ provides a well-defined causal estimate and is not just an associational interpretation derived from the ordinary accelerated failure time model. The ability to adjust for time-varying confounding makes the implicit sequential modelling at every stage be nested to the previous stage (Hernán et al., 2005; Korhonen et al., 1999; Lok et al., 2004; Robins, 2000), i.e. the CALM model (2.6) can be re-expressed as

$$\begin{aligned} W_i(\varphi) &= \textstyle\int_0^{T_i} \exp\left[\varphi A_i(s)\right] ds = \int_0^{D_i} \exp(\varphi) ds + \int_{D_i}^{T_i} ds \\ &= D_i \exp(\varphi) + (T_i - D_i) = T_i - D_i\left[1 - \exp(\varphi)\right], \end{aligned} \tag{2.7}$$

where D_i is the time unit subject i would have spent if it were on active treatment: we observe $T_i = W_i$ in the placebo arm because $D_i = 0$, but in the treatment arm W_i is observed only if $D_i = 0$ (in the absence of censoring). The quantity $[1 - \varphi]^{-1}$ can be interpreted as the fractional increase/decrease in survival time if subject i were always on active treatment as opposed to never being on active treatment (Robins and Tsiatis, 1991). For example, $\varphi = 0.5$ implies the remaining lifetime is doubled if always on active treatment and $\varphi = -1$ implies it is halved if always treated as compared to never treated. Finally, the model is rank-preserving in the sense that order of failure time is maintained among individuals when they receive the same treatment, for example, if subject i would die before subject j had they both been untreated (i.e. $T_i^{a=0} < T_j^{a=0}$), then subject i would also die before subject j if they had been treated (i,e. $T_i^{a=1} < T_j^{a=1}$). This is what Robins (2008) refers to as identical effect for the two subjects under *time* 0 *intervention*.

2.6.2 CHARM: Causal Hazard ratio Adjustment Regression Models

This method was proposed by White et al. (2004) to address partial noncompliance by viewing survival outcome as a sequence of binary outcomes to provide an 'approximate' overall hazard ratio estimate which is adjusted for compliance. We note that this approach initially ignores survival time by considering the outcome as binary.

We define the following random variables for survival data to evaluate compliance with active treatment (i.e. compliance with active if offered) while ignoring compliance with placebo. Let Y represent survival status, for example, death (principal outcome of interest) so that $Y = 1(0)$ if participant is dead (alive) by the end of the interval or study as the case may be. Let $Z = 1(0)$ represent treatment or placebo arm while $A = 1(0)$ represent observed treatment (placebo) receipt. Let $S = 1(0)$ represent a participant's latent (potential) compliance behaviour. Further let α represent the probability of noncompliance, i.e. $\alpha = \Pr(S = 0)$. Given this setup, we can observe S as equal to A in the treatment arm but we cannot observe the latent compliance S in the placebo arm.

Assuming some form of compliance so that $\alpha > 0$, let π_Z^0 and π_Z^1 respectively represent the probabilities of death among noncompliers and compliers in the Z arm, i.e.

$$\pi_Z^0 = \Pr(Y = 1|Z = z, S = 0) \quad \text{and} \quad \pi_Z^1 = \Pr(Y = 1|Z = z, S = 1),$$

where π_Z^0 and π_Z^1 may be assumed equal by randomization principle. Then the overall probability of death in arm Z may be given by

$$\pi_Z = (1 - \alpha)\pi_z^1 + \alpha\pi_z^0. \tag{2.8}$$

We can use the above probabilities to define RR_{ITT} and RR_{CACE} as quantities based on the risk ratio scale where RR_{ITT} is simply the ratio of π_1 and π_0 while CACE may be considered as the ratio of π_1^1 and π_0^1.

However, we note that while we can directly estimate $\text{RR}_{\text{ITT}}, \alpha, \pi_1^0, \pi_1^1$ and π_1, we may not be able to directly estimate $\text{RR}_{\text{CACE}}, \pi_0^0$ and π_0^1 because S is an unobservable (latent) variable. For survival data, we can resolve this difficulty by using the exclusion restriction (ER) which assumes that the risk of death does not depend on the arm of randomization given the treatment received, i.e. $Y \perp Z|A$. The ER assumption then implies that

$$\pi_0^0 = \pi_1^0. \tag{2.9}$$

We may use the ER assumption to find CACE estimates. Following White et al. (2004), we can subdivide survival data into small intervals and assume subjects on active treatment may only stop treatment at the start of that interval. A sufficiently small interval may allow us to assume constant hazard rates. Then by considering death as a sequence of binary outcomes and using T to denote the interval in which death occurred, White et al. obtained an ITT risk ratio in interval i as

$$\text{ITT}_i = \frac{\Pr(T = i|T \geq i, Z = 1)}{\Pr(T = i|T \geq i, Z = 0)}, \ i = 1, \ldots, I. \tag{2.10}$$

Using the counterfactuals framework, we can evaluate treatment effects by considering

participants in the placebo arm. Now let the random variable S (the compliance-type) define the interval in which stopping (death/censoring) occurs, or $I+1$ if no stopping occurs. For participants on the control arm, S represents the counterfactual stopping interval, i.e. the interval in which stopping *would have occurred* had they been allocated to the treatment arm. White et al. (2004) then use an *extended exclusion restriction* assumption to obtain CACE estimate in each interval i. With the extended exclusion restriction, we assume that randomized allocation has no effect on any survivors to the start of interval i who would stop treatment by the start of interval i, i.e.

$$\begin{aligned} \Pr(T &= i | T \geq i, S = s, Z = 1) \\ &= \Pr(T = i | T \geq i, S = s, Z = 0) \ \forall \ i = 1, \ldots, I, \ s \leq i. \end{aligned} \tag{2.11}$$

We observe that the extended exclusion restriction assumption (2.11) mimics a version of the Markov property that the present is conditionally independent of the past. They then considered the CACE in interval i as the risk ratio among those who survive to interval i and who would not stop treatment before the end of interval i:

$$\text{CACE}_i = \frac{\Pr(T = i | T \geq i, S > i, Z = 1)}{\Pr(T = i | T \geq i, S > i, Z = 0)}. \tag{2.12}$$

Both exclusion restriction (2.9) and extended exclusion restriction (2.11) assumptions may enable us use the proportion of non-compliers to estimate the proportion of compliers used in obtaining CACE estimate. White et al. then showed that

$$\text{CACE}_i \approx \frac{\text{ITT}_i(1 - \theta_i)}{1 - \theta_i \ \text{ITT}_i}, \tag{2.13}$$

where $\theta_i = \Pr(A_i = 0 | T = \ frm[o] - - i, Z = 1)$ is the probability that a participant in the active treatment arm who experienced event (e.g. died) in interval i had previously stopped treatment, i.e. was non-complier by the time of her death.

Assuming short intervals in addition to EER assumption, White et al. (2004) showed that we can extend the CACE estimate given by equation (2.12) above for discrete time case to continuous (pooled) hazard rates at time t. By weighting the ITT(t) estimates from (2.10)

we then CACE estimate for proportional hazards as

$$\text{CACE}_{\text{PH}}(\text{t}) = \frac{\text{ITT}(t)(1-\theta(t))}{1-\theta(t)\ \text{ITT}(t)}, \tag{2.14}$$

where $\theta = \Pr\left(A=0|Y=1, Z=1\right)$, is the noncompliance probability among participants in the treatment arm who died. The CACE_{PH} method is essentially a time-adjusted ITT estimate that fits a Cox's PH model (HR_{ITT}) to provide a Causal Hazard ratio Adjustment Regression model (CHARM) where the linear predictor in the control arm is zero and in the active arm is a function of time, for example,

$$\text{CHARM} \equiv \text{h(t)} = \text{h}_0(t)\exp[\beta_0 Z + \beta_1 f(t) Z], \tag{2.15}$$

where f(t) can be specified as a combination of linear/quadratic function of time to event.

White et al. (2004) proposed two ways to obtain a constant CHARM estimate (2.15) above:

(i) either allowing both HR_{ITT} and odds of noncompliance θ to vary over time or

(ii) assuming time invariance for both HR_{ITT} and θ so as to allow a simple CHARM approximation of (2.14) by

$$\text{CHARM} \equiv \widehat{\text{CACE}}_{\text{PH}} = \frac{\widehat{\text{HR}}_{\text{ITT}}(1-\hat{\theta})}{1-\hat{\theta}\ \widehat{\text{HR}}_{\text{ITT}}}, \tag{2.16}$$

where $\theta = \Pr\left(A=0|Z=1\right)$ represent the proportion of noncompliers randomized to the treatment arm who experienced event of interest (e.g death).

Assuming a single death at a time is experienced, estimation procedure (ii) may be implemented by estimating individual θ_is for the dead non-compliers and using a logistic regression model with these values to examine evidence of change or trend in the θ_is (White et al., 2004). For example,

$$\text{logit}\ \theta(t) = \delta_0 + \delta_1 t, \tag{2.17}$$

would represent a linear trend of noncompliance. Here the outcome would be 1 if a participant is a noncomplier by the time of death and 0 if a complier by the time of her death, i.e.

$$\theta = \begin{cases} 1; & A = 0|Y = 1 \\ 0; & A = 1|Y = 1. \end{cases}$$

In our analysis we will implement the complex procedure (i) above using the Stata command `adjhr` (White, 2002) which provides an 'approximate' constant (overall) CHARM estimate as a hazard ratio adjusted for compliance. We will also compare this with a simple CHARM estimate obtained using a two-stage regression strategy: logistic regression to estimate θ which is then substituted in Equation (2.16) above.

2.7 Modelling noncompliance in two treatment arms

Introduction

All the specialist methods considered above adjusted for noncompliance in one (active) treatment arm while ignoring compliance in the placebo arm. The decision to ignore compliance data for the control group is an exercise in precision-bias tradeoff (Sommer and Zeger, 1991). Imperfect compliance with placebo allocation results in possible noncompliance in two arms. Such scenarios present further challenges given that even ITT estimation here produce biased efficacy estimates even under homogeneous treatment assumption (Aalen, 1998; Baker and Kramer, 2005; Robins, 1998a). For example, a double-blind placebo-controlled randomized clinical trial with active treatment may present noncompliance at two levels (Jin and Rubin, 2008): first simple noncompliance with assignment in either arm (active/placebo) and second differential noncompliance with placebo in the presence of possible imperfect blinding, e.g. due to adverse side effects of the active treatment which may unconceal blinding. The Esprit study is a candidate for such a scenario where bleeding would reveal a victim's treatment allocation (active) and possible differential noncompliance. This presents further challenge/s.

2.7.1 Brief literature review

In their pioneer work, Efron and Feldman (1991) used compliance as a covariate in a regression adjustment. However, their method has been criticized for the implicit strong assumption of comparability in compliance between the active treatment and placebo arms (Albert and DeMets, 1994). In the presence of selectivity effects, many methods have been developed to account for noncompliance in more than one treatment arm. For example, Robins (1989a, 1994, 1999) introduced the structural mean models (SMM) whose parameter provide well-defined efficacy estimates as functions of expected potential outcomes for the population of subjects on treatment. In addition to exclusion restriction, the *placebo link assumption* employed here posits that potential response to no active treatment is equal to the response under control conditions, i.e. the assumption permits use of randomization in the estimation procedures hence inducing exchangeability. Fischer-Lapp and Goetghebeur (1999, 2004) applied SMMs on a placebo-controlled double-blind trial of patients with mild hypertension in the United Kingdom to assess the effect of baseline predictors on reduction of blood pressure.

While structural mean models were developed for continuous outcomes data, Vansteelandt and Goetghebeur (2003) developed the generalized structural mean models as an extension to handle non-linear average treatment effects. When those randomized to placebo have no access to active treatment (no defiers assumption holds), the estimator from structural mean models is equivalent to the instrumental variable estimator (Goetghebeur and Fischer-Lapp, 1997; Robins, 1994). To relax the exclusion assumption in SMMs, Robins and Rotznitzky (2004) imposed additional parametric modelling restrictions to account for possibility of placebo having access to treatment. Robins (1994, 1997, 1998c) introduced the structural nested mean models which adjusts for noncompliance in time-to-event data. And Baker and Kramer (2005) used potential outcomes to construct maximum likelihood estimates for discrete-time survival outcomes under all-or-nothing switching of treatments where switching occurred immediately after randomization or at the start of the time period.

Principal stratification is another causal modelling framework whose flexibility allows

application to adjusting for noncompliance in more than one treatment arm. While it is possible to perform pairwise efficacy comparisons in the presence of more than one active treatment, a joint analysis is likely to provide additional analytical insights (Cheng and Small, 2006). However, trials involving multiple treatments with possible noncompliance are likely to present complex identification problems (Long et al., 2010; Roy et al., 2008). For example, principal stratification of a three-armed trial ($Z, A \in \{1, 2, 3\}$) produces a total of 27 principal strata defined by the set of 3^3 possible combinations. But all that we observe about the principal strata is the value of $A(z)$ corresponding to the treatment a that is actually received. The rest remain unobserved which results in identification problem that can only be addressed with additional assumptions, a solution that may limit generalization of results.

Most studies in the health sciences report results in terms of causal point estimates premised on implicit assumptions. As a result, the validity/credibility of these estimates are dependent on the accuracy of such assumptions. Manski (2007) refers to this phenomenon as the as *the law of decreasing credibility* which posits that the measure of credibility of inference is inversely proportional to the strength of underlying assumptions. In practice, relaxing the assumptions may widen credibility of the results but at the price of making point identification impossible, i.e. the estimates cannot be uniquely identified from the probability distribution of observable outcomes (Cai et al., 2007). However, the estimates may be partially identifiable under mild assumptions which permit construction of upper and lower bounds of the identification region of parameters (Manski, 1990, 1995, 2003). For example, while Imbens and Angrist (1994) showed that relaxing the monotonicity assumption by allowing defiers breaks down point identification for a two-armed trial, Balke and Pearl (1997) used linear programming to provide the 'tightest' possible bounds on the average treatment effects. Allowing defiers is likely to complicate identification more in the presence of two/more active treatments. Using method-of-moments approach under two sets of assumptions obtained by decomposing the monotonicity assumption, Cheng and Small (2006) derived sharp bounds for principal effects for a binary outcome in a three-armed trial and constructed confidence intervals for the corresponding identification regions of parameters

with fixed probability. For a control (0) and two active treatments (1 and 2), the first part of their monotonicity assumption posited that a subject in the control had no access to either of the two active treatments and no treatment switches was permitted among those assigned to the two active treatments. In addition, their second part, '*extended monotonicity*', assumed similarity in compliance behaviour among those assigned to active treatment where a subject complying with treatment 2 would have complied with treatment 1. We note that this is a plausible assumption if treatment 1 is equally effective as treatment 2 but with possibly more tolerable/same side-effects.

While Roy et al. (2008) proposed a principal stratification framework for two-active treatment trials (no placebo) using baseline covariates to address identification problem, Long et al. (2010) recently proposed a likelihood-based extension of the Cheng and Small (2006) model to provide point causal estimands for a three-armed trial and general (continuous or discrete) outcomes. Using Bayesian methods, they model the arm-specific compliances directly while treating the principal compliance status as missing. This approach relaxes the *extended monotonicity assumption* of Cheng and Small (2006) described above by not making any assumptions about compliance behaviours for those subjects assigned to active treatment. On the other hand, Fischer et al. (2011) also recently proposed a structural mean modelling approach using baseline covariates predictive of compliance in individual arm to obtain compliance-adjusted efficacy in a randomized-controlled trial comparing two active treatments.

The next section provide a comprehensive review of the Roy et al. (2008) method which uses principal stratification framework to adjust for noncompliance in two arms for a two-active treatment trials with binary outcomes. Our objective will be to apply this model to survival data using Bayesian methods (Chapter 6) and also to use simulation studies to evaluate its performance in terms of bias and 95% credible intervals (Chapter 8).

2.7.2 Principal stratification for two-active treatment arms

Adopting notation from previous chapter for two-armed trial, let $Z \in \{0, 1\}$ denote a randomization indicator: $Z = 1$ indicate randomization to the new treatment (e.g. HRT treatment) and $Z = 0$ indicates randomization to control/placebo. In the present case 1 ($Z = 1$) will represent randomization to new treatment and 0 ($Z = 0$) randomization to standard treatment. Let $A \in \{0, 1\}$ denote compliance with assigned treatment and define $Y \in \{0, 1\}$ to be the outcome of interest (e.g. death). We note that each subject has two potential compliance levels A_0 and A_1 (compliance with standard and new treatment respectively) and two potential outcomes Y_0 and Y_1 (outcome under standard and new treatment respectively). But the observed compliance and outcomes are respectively given by $A = ZA_1 + (1 - Z)A_0$ and $Y = ZY_1 + (1 - Z)Y_0$.

Analysis under this formulation utilizes baseline covariates X to modify the standard assumptions for causal modelling outlined above with two additional assumptions (d and e):

(a) The stable unit-treatment value assumption (SUTVA),

(b) Randomization: $Z \perp \{Y_0, Y_1, A_0, A_1, X\}$ and

(c) The exclusion restriction: $\Pr(Y_1|A_Z, X) = \Pr(Y_0|A_Z, X)$,

(d) Treatment access restriction: which posits no treatment switches among subjects.

(e) Monotonicity: $\Pr(A_1 = 1|A_0 = 1, X) \geq \Pr(A_1 = 1|A_0 = 0, X)$, i.e. the probability of compliance with treatment assigned by $Z = 1$ is higher among those who would comply with treatment assigned by $Z = 0$, compared to those who would not.

The monotonicity assumption helps tighten the bounds of causal effects (Cheng and Small, 2006; Roy et al., 2008). This version of monotonicity assumption is applicable to Esprit study because there was no preference to one treatment over the other, i.e compliance with HRT treatment would be more prevalent among those who would comply with placebo. In our simulations, we reflect this assumption through a pre-specified positive correlation (sensitivity parameter ϕ) between A_0 and A_1.

Each subject is assumed to belong to one of four basic principal strata defined by unique combinations of (A_0, A_1) where the principal strata comprise the set $\mathcal{S} = \{(0,0),(1,0),(0,1),(1,1)\}$. The interest (next section) will be to seek the joint distributions $[(Y_0, Y_1)|S = s]\ \forall\ s \in \mathcal{S}$ which provides principal effects of interest for each stratum.

2.7.3 Causal model linking two marginal compliance models

We consider two active treatments A and B which we denote as 0 and 1 respectively. To predict compliance to treatment allocation for each arm separately given a selected set of predictors of compliance $\mathrm{x}_0 = 1$ and $\mathrm{x}_1, \ldots, \mathrm{x}_n$, we use the logistic models

$$\text{logit}\left[\mu_j(\mathrm{x})\right] = \left(\sum_{i=0}^{n} \gamma_{ji}\mathrm{x}_i\right),\ j = 0, 1, \tag{2.18}$$

where $\mu_j(\mathrm{x})$ is the probability of compliance with allocated treatment j given set of covariates X: the estimated probabilities of complying with arm-specific treatment allocation may then be obtained using

$$\hat{\mu}_j(\mathrm{x}) = \left[1 + \exp\left(-\sum_{i=0}^{n} \hat{\gamma}_{ji}\mathrm{x}_i\right)\right]^{-1},\ j = 0, 1, \tag{2.19}$$

where γ represent the log odds ratio estimates of compliance.

An important issue is how the two compliance behaviours are correlated. Following Roy et al. (2008), we define a non-negative sensitivity parameter ϕ which is related to the correlation ρ between compliances to treatment allocation (0/1) and is assumed positive. It can then be shown that, if $\hat{\mu}_0(\mathrm{x}) > \hat{\mu}_1(\mathrm{x})$ then

$$\phi = \rho\sqrt{\frac{\bar{\hat{\mu}}_0(\mathrm{x})[1 - \bar{\hat{\mu}}_1(\mathrm{x})]}{\bar{\hat{\mu}}_1(\mathrm{x})[1 - \bar{\hat{\mu}}_0(\mathrm{x})]}}. \tag{2.20}$$

The joint probability distribution of compliance to treatment 0 and compliance to treatment 1 is then a function of the arm-specific marginal compliance probabilities and ϕ and

can therefore be estimated for a given value of ϕ. We however note that, ϕ is unknown in general. Specifically if $U(\mathrm{x}) = \min\{1, \frac{\hat{\mu}_1(\mathrm{x})}{\hat{\mu}_0(\mathrm{x})}\}$ then Roy et al. (2008) showed that the joint probabilities are given by:

$$\begin{aligned}
\hat{\mu}_{11}(\mathrm{x}) &= \Pr(A_0 = 1|X)P(A_1 = 1|A_0 = 1, X) \\
&= \hat{\mu}_0(\mathrm{x})\hat{\mu}_1(\mathrm{x}) + \phi\hat{\mu}_0(x)[U(\mathrm{x}) - \hat{\mu}_1(\mathrm{x})], \\
\hat{\mu}_{01}(\mathrm{x}) &= \Pr(A_0 = 0|X)P(A_1 = 1|A_0 = 0, X) \\
&= \hat{\mu}_1(\mathrm{x}) - \hat{\mu}_0(\mathrm{x})\hat{\mu}_1(\mathrm{x}) - \phi\hat{\mu}_0(\mathrm{x})[U(\mathrm{x}) - \hat{\mu}_1(\mathrm{x})], \\
\hat{\mu}_{10}(\mathrm{x}) &= \Pr(A_0 = 1|X)P(A_1 = 0|A_0 = 1, X) \\
&= \hat{\mu}_0(\mathrm{x}) - \hat{\mu}_0(\mathrm{x})\hat{\mu}_1(\mathrm{x}) - \phi\hat{\mu}_0(\mathrm{x})[U(\mathrm{x}) - \hat{\mu}_1(\mathrm{x})], \\
\hat{\mu}_{00}(\mathrm{x}) &= \Pr(A_0 = 0|X)P(A_1 = 0|A_0 = 0, X) \\
&= 1 - \hat{\mu}_0(\mathrm{x}) - \hat{\mu}_1(\mathrm{x}) + \hat{\mu}_0(\mathrm{x})\hat{\mu}_1(\mathrm{x}) + \phi\hat{\mu}_0(\mathrm{x})[U(\mathrm{x}) - \hat{\mu}_1(\mathrm{x})],
\end{aligned} \tag{2.21}$$

where X is the set of covariates predictive of compliance in both arms and $A_1(A_0)$ is an indicator of compliance to HRT treatment (placebo) and $\hat{\mu}_{ij}(\mathrm{x})$ denote the probability of being in the compliance subgroup ij as illustrated in Table (2.1).

Table 2.1: Compliance proportions for each stratum

		Comply treat 1 (A_1)	
		Yes (1)	No (0)
Comply	Yes (1)	$\mu_{11}(\mathrm{x})$	$\mu_{10}(\mathrm{x})$
treat 0 (A_0)	No (0)	$\mu_{01}(\mathrm{x})$	$\mu_{00}(\mathrm{x})$

Following principal stratification framework (Frangakis and Rubin, 2002), the possible values of A_0 and A_1 define a stratification factor S for the population of patients. For a defined outcome variable Y (mortality/reinfarction for Esprit study), let Y_0 and Y_1 refer to potential outcomes under treatment A and treatment B respectively. The four possible realizations of (Y_0, Y_1) at each level of S (for example $\varpi_{11} = \Pr[Y_0 = 1, Y_1 = 1]$) are as shown in Table (2.2).

The expression for stratum $S = (0, 0)$ differs from the others because of the exclusion restriction: $\varpi_{10}(0, 0) = \varpi_{01}(0, 0) = 0$. The joint distribution of potential outcomes

Table 2.2: Probabilities for joint distribution of potential outcomes per stratum: $[Y_0, Y_1|S]$.

			(Y_0, Y_1)			
A_0	A_1	S	(0,0)	(0,1)	(1,0)	(1,1)
0	0	(0,0)	$\varpi_{00}(0)$	0	0	$\varpi_{11}(0)$
1	0	(1,0)	$\varpi_{00}(1)$	$\varpi_{01}(1)$	$\varpi_{10}(1)$	$\varpi_{11}(1)$
0	1	(0,1)	$\varpi_{00}(2)$	$\varpi_{01}(2)$	$\varpi_{10}(2)$	$\varpi_{11}(2)$
1	1	(1,1)	$\varpi_{00}(3)$	$\varpi_{01}(3)$	$\varpi_{10}(3)$	$\varpi_{11}(3)$

(Y_0, Y_1) for each stratum $S, f(Y_0, Y_1|S, \varpi)$ is assumed to be multinomial with probabilities $\varpi(S) = \Pr(Y_0 = y_0, Y_1 = y_1|S = s)$.

A crucial identification assumption for the Roy et al. (2008) model posits that $Y \perp X|S, Z$, where X represents the set of covariates predicting compliance. This assumption underscores an integral component of the method which involves selecting suitable predictors of compliance. In the next Chapter (Chapter 3), we attempt to address this by considering how predictors of compliance should be selected by a review of general literature on model selection for prediction and apply it to select plausible separate predictors of compliance for HRT treatment and placebo arms for the Esprit study in Chapter 5. Although untestable, a comparison of results using different sets of predictors may be used to test this assumption (Chapter 6).

After reparameterizing in terms of π (probability of experiencing event, e.g. death or myocardial reinfarction) and $\beta = f(\gamma, \phi))$ (log odds ratio of compliance for specified sensitivity), Roy et al. (2008) showed that the observed-data likelihood is

$$L(\pi, \beta|Y, A, Z, X) = \sum_{s=0}^{3} [\pi_Z^{S=s}]^Y [1 - \pi_Z^{S=s}]^{1-Y} \Pr(S = s|X, \beta) G(s, A, Z), \tag{2.22}$$

where $\pi_Z^{S=s}$ is the probability that observed $Y = 1$, given $S = s$ and allocation to arm Z, and

$$\begin{aligned} G(s, A, Z) &= I(s = 0)\{1 - A\} + I(s = 1)\{A(1 - Z) + (1 - A)Z\} \\ &\quad + I(s = 2)\{AZ + (1 - A)(1 - Z)\} + I(s = 3)A. \end{aligned}$$

where $\pi_Z^{S=s}$ is the probability that observed $Y=1$, given $S=s$ and allocation to arm Z, and

Now Y and A are observed values; e.g. $A=A_1$ if allocated to active treatment and $A=A_0$ if allocated to placebo. We can decompose the expression (2.22) for each stratum:

$$\begin{aligned}
L(\pi,\beta|Y=1,A=1,Z=1,X) &= \pi_1^{s=(0,1)}\Pr\left(S=(0,1)|X,\beta\right)\\
&+ \pi_1^{s=,(1,1)}\Pr\left(S=(1,1)|X,\beta\right),\\
L(\pi,\beta|Y=0,A=1,Z=1,X) &= [1-\pi_1^{s=(0,1)}]\Pr\left(S=(0,1)|X,\beta\right)\\
&+ [1-\pi_1^{s=(1,1)}]\Pr\left(S=(1,1)|X,\beta\right),\\
L(\pi,\beta|Y=1,A=0,Z=1,X) &= \pi_1^{s=(0,0)}\Pr\left(S=(0,0)|X,\beta\right)\\
&+ \pi_1^{s=(1,0)}\Pr\left(S=(1,0)|X,\beta\right),\\
L(\pi,\beta|Y=0,A=0,Z=1,X) &= [1-\pi_1^{s=(0,0)}]\Pr\left(S=(0,0)|X,\beta\right)\\
&+[1-\pi_1^{s=(1,0)}]\Pr\left(S=(1,0)|X,\beta\right),\\
L(\pi,\beta|Y=1,A=1,Z=0,X) &= \pi_0^{s=(1,0)}\Pr\left(S=(1,0)|X,\beta\right)\\
&+ \pi_0^{s=(1,1)}\Pr\left(S=(1,1)|X,\beta\right),\\
L(\pi,\beta|Y=0,A=1,Z=0,X) &= [1-\pi_0^{s=(1,0)}]\Pr\left(S=(1,0)|X,\beta\right)\\
&+ [1-\pi_0^{s=(1,1)}]\Pr\left(S=(1,1)|X,\beta\right),\\
L(\pi,\beta|Y=1,A=0,Z=0,X) &= \pi_0^{s=(0,0)}\Pr\left(S=(0,0)|X,\beta\right)\\
&+ \pi_0^{s=(0,1)}\Pr\left(S=(0,1)|X,\beta\right),\\
L(\pi,\beta|Y=0,A=0,Z=0,X) &= [1-\pi_0^{s=(0,0)}]\Pr\left(S=(0,0)|X,\beta\right)\\
&+ [1-\pi_0^{s=(0,1)}]\Pr\left(S=(0,1)|X,\beta\right).
\end{aligned} \tag{2.23}$$

By the exclusion restriction $\pi_1^{s=(0,0)}=\pi_0^{s=(0,0)}$, i.e.the risk of experiencing event of interest (e.g. death) is independent of the arm of allocation among the people who would comply with neither allocation. Writing

$$\pi_1=\pi_1^{s=(0,1)},\pi_2=\pi_1^{s=(1,1)},\pi_3=\pi_1^{s=(0,0)},\pi_4=\pi_1^{s=(1,0)},$$

and

$$\pi_5 = \pi_0^{s=(1,0)}, \pi_6 = \pi_0^{s=(1,1)}, \pi_7 = \pi_0^{s=(0,1)},$$

we obtain 7 parameters captured by π from the likelihoods above using logistic models:

$$\begin{aligned}
\Pr[Y=1|A=1,Z=1] &= \pi_1\mu_{01} + \pi_2\mu_{11},\\
\Pr[Y=1|A=0,Z=1] &= \pi_3\mu_{00} + \pi_4\mu_{10},\\
\Pr[Y=1|A=1,Z=0] &= \pi_5\mu_{10} + \pi_6\mu_{11},\\
\Pr[Y=1|A=0,Z=0] &= \pi_3\mu_{00} + \pi_7\mu_{01}.
\end{aligned} \tag{2.24}$$

We then obtain the stratum-specific relative risks of death as

$$\begin{aligned}
\tau_{11} &= \frac{\pi_1^{s=(1,1)}}{\pi_0^{s=(1,1)}} = \frac{\hat{\pi}_2}{\hat{\pi}_6},\\
\tau_{01} &= \frac{\pi_1^{s=(0,1)}}{\pi_0^{s=(0,1)}} = \frac{\hat{\pi}_1}{\hat{\pi}_7},\\
\tau_{10} &= \frac{\pi_1^{s=(1,0)}}{\pi_0^{s=(1,0)}} = \frac{\hat{\pi}_4}{\hat{\pi}_5}.
\end{aligned} \tag{2.25}$$

The τ_{ij} provides principal (causal) effects in terms of causal risk ratios obtained as means of posterior relative risks of event (death or reinfarction) for each stratum defined by compliance type:

(i) τ_{11}: risk of event due to compliance with HRT treatment relative to placebo among the subgroup of patients who would comply with either treatment allocation - $S=(1,1)$,

(ii) τ_{01}: risk of event due to compliance with HRT treatment only among women who would comply if allocated to it - $S=(0,1)$ and

(iii) τ_{10}: risk of event due to compliance with placebo treatment only among the subgroup who would comply if allocated to it - $S=(1,0)$.

The parameters (2.25) can be estimated using Bayesian methods with suitable priors, for example, $\pi \sim U(0,1)$ for uninformative priors.

To extend the methods which adjust for noncompliance in one treatment arm to adjusting for noncompliance in two treatment arms, we will apply principal stratification using the Roy et al. (2008) model reviewed above for survival data but which was originally proposed for binary data. Using arm-specific compliance predictors (see Chapter 3), we develop a causal model linking the two marginal models that allows us to estimate causal effects in each stratum. By assuming all-or-nothing compliance to treatment allocation, we apply the method to the Esprit data in Chapter 6 and apply statistically designed simulation studies in a Bayesian setup to evaluate the performance of the method in terms of bias and 95% credible intervals in Chapter 8.

Developing arm-specific compliance models given by Equations (2.18) and (2.19) is an exercise in model selection, i.e. selecting plausible predictors of compliance with treatment assignment for each trial arm. In the next Chapter (3) we consider model selection techniques to obtain prediction models of compliance for each treatment arm, a challenge which has not been addressed much in the literature and practice. The next chapter provides a review of model selection techniques.

Chapter 3

Model Selection for Prediction

3.1 Introduction

In this chapter we review some of the methods for model selection with a focus on adopting the methods to build prediction models for compliance with treatment allocation. We start with a brief literature review on model selection then discuss the strengths and limitations of stepwise regression procedures. For completeness, the next section briefly outlines the methods of principal components and partial least squares regression. The next section presents a review of penalized regression techniques providing the bridge estimation as a Bayesian generalization of best-subset, ridge and Lasso regressions. The final section provides a summary of measures for model validation which enable us evaluate the performance of selected prediction models.

3.2 Prediction models

Prediction is one of the principal objectives of statistical modelling besides estimation and testing hypothesis. The purpose of a prediction model is to provide valid and reliable outcomes for new observations. Statistical prediction involves selecting a subset of variables that are predictive of a specified outcome. Developing an effective and useful prediction model is

often accomplished after the twin processes of model selection and model assessment. Model selection is the process of choosing a suitable subset of predictors from an original set while model assessment on the other hand is the evaluation of models' stability when applied to independent data. The main aims of model selection include accurate predictions and easy interpretation (Harrell, 2001; Hastie et al., 2009). Accurate prediction is likely to provide regression coefficients with no/minimal bias, narrow confidence intervals for the parameter estimates and lower prediction error. A model composed of fewer, meaningful and clinically most relevant predictors is easily interpretable. With a stable model, small changes in the data do not result in large changes in either the subset of predictors used, the associated coefficients, or the predictions made from it. While accurate prediction is an integral component of a model's internal validity, model stability is a measure of external validity that allows generalization/application of the model to new/independent data. Although both components are desirable for effective prediction, internal validity is a prerequisite for external validity (Steyerberg, 2009), i.e. it makes less statistical sense to assess a model that is poorly selected. In general, validation lends credence to prediction models. For the present work we adopt the general definition of validation by Harrell (2001) as a measure of performance of predictive model on independent/new data.

Most clinical trials often record a number of prognostic and demographic baseline variables. Such variables might sometimes be used to adjust ITT comparisons for random imbalances but they may also provide an insight into the propensity of subjects to comply with treatment allocation. However, not all variables may predict compliance hence utilizing all of them may add little or no information in analysis. Predicting compliance is essentially a study in model selection. In general, prediction models use empirical data from a sample of patients drawn from a larger population. But data from such sample are only important to the extent that they are representative of the underlying population (Altman and Royston, 2000). Analysis of a representative empirical data is not only capable of revealing patterns in the population but also models derived from it can provide accurate predictions for new subjects from this population.

However, a major risk in extrapolating information from a sample to a population is the failure of predictions to generalize to new subjects outside the sample (Steyerberg, 2009). This challenge may manifest itself in overfitted prediction models where the data concentrates in capturing specific trends and idiosyncrasies of the sample. Overfitting the data is the major cause of unreliable predictive models; it occurs when noise variables are retained as potential predictors (Harrell, 2001). Overfitting is what statisticians often refer to as the curse of dimensionality, i.e. fitting statistical model with too many parameters which is likely to produce inflated variance for the regression function (Hastie et al., 2009; Izenman, 2008). Overfitting implies that there is overoptimism about a model's performance in new subjects. Overfitting in prediction models can be defined as fitting a statistical model with too many degrees of freedom in the modelling process (Steyerberg, 2009). Overfitting during this process arise from two sources of variation: parameter uncertainty and model uncertainty. While model uncertainty refers to level of complexity/specification of structure of model (e.g. number of predictors), parameter uncertainty refers to potential instability of the coefficients. The two processes acting together impact on the number of degrees of freedom of a model arising from the fact that on the one hand some degrees of freedom are used in estimation of the coefficients in a regression model (address parameter uncertainty) and on the other more degrees of freedom are used in searching for the optimal model structure, i.e. addressing model uncertainty. For example, this may arise from the predictor selection procedures.

Developing reliable and valid prediction models may be viewed as striking an informed balance between curiosity and skepticism. On one hand, we use science to make discoveries and advance knowledge while on the other hand we must subject such discoveries to stringent tests like validation in order to eliminate any possibility of being fooled by chance (Babyak, 2004). So given the important role of model validation in the development of reliable predictive models, how do we construct internally and externally valid models to predict treatment compliance?

Stepwise variable selection procedures and best-subset selection are some of the most commonly used (and abused) model selection techniques (Draper and Smith, 1998; Harrell, 2001).

Both methods use classical statistics like t and F tests for judging the significance of individual predictors and comparing two/more models respectively. But these tests are premised on the assumption that the set of predictors and the models are pre-specified. However, the truth is that the set of predictors or models are chosen adaptively according to algorithms specific to each selection procedure. Results from such analysis are prone to 'testimation' (estimation after multiple testing) bias and such bias affects the variable selection process itself due to model and parameter uncertainty (Chatfield, 1995; Hesterberg et al., 2008). In general, the resulting regression coefficients are biased upwards in absolute value which may then lead to misleading conclusions about the magnitude of effects.

While unbiased estimates are desirable, in the presence of many covariates this is obtained at the price of large variances. Generally models with many covariates have small bias but large variance and conversely models with few covariates have large bias but small variance (Draper and Smith, 1998). Better predictions can be obtained by carefully balancing these two extremes, i.e. trading a little bias for large reduction of variance. This bias-variance tradeoff is mostly achieved by penalized (shrinkage) regression techniques. Shrinkage means constraining the coefficients during estimation with the objective of increasing reliability to ensure internal validity. Here shrinkage regression is a solution to overfitting which draws estimated regression coefficients to less extreme values. Shrinkage estimation can be viewed as a generalized technique of regression towards the mean (Wright and London, 2009).

The first section of this chapter provides a review of the classical stepwise selection and best subset regression together with their merits and limitations. This is followed by a review of two types of multivariable regression methods: principal component regression and partial least squares regression with a summary of their advantages and disadvantages. The next section reviews modern selection techniques: ridge regression and Lasso (Least Absolute Selection and Shrinkage Operator) methods, their merits and demerits and how they relate to each other. Finally we look at model validation techniques and review two measures of performance (calibration slope and discrimination/concordance c-statistic) to quantify optimism in the selected prediction models. We use the terms covariates and predictors interchangeably.

3.3 Stepwise and best subset regression

Stepwise variable selection is the most commonly used model selection technique (Harrell, 2001; Hastie et al., 2009). As a selection procedure, stepwise is implemented in three versions. Forward selection begins with an empty model consisting of the intercept only. We then add variables sequentially to the model until a predefined stopping rule is satisfied. At each step of the selection process, we add the variable whose inclusion results in the best fit (i.e. greatest increase in the summary measure). Some of the most commonly used summary measures include R^2, adjusted R^2, residual sum of squares and deviance (Draper and Smith, 1998). A predefined significance level is typically used as a stopping criteria so that only statistically significant variables are added to the model. Backward elimination procedures on the other hand begin with a saturated (full) model composed of all candidate predictor variables. Using a predefined stopping rule, the procedure sequentially removes variables which contribute least to the model fit. Stepwise selection is a variation combining both forward and backward selection algorithms: at each step of the variable selection process, after a variable has been added to the model, variables are allowed to be eliminated from the model. For instance, if the p-value of a given predictor is above a specified threshold, it is eliminated from the model. The iterative process is ended when a pre-specified stopping criteria is satisfied.

Stepwise variable selection procedures produce nested sequences of models. The inherent collinearity can cause predictors to compete hence making the selection of 'important' predictors arbitrary. The competition and accompanying (potential) arbitrariness in selection procedures often results in use of greedy algorithms (Hastie et al., 2009; Hesterberg et al., 2008). Such model selection process is prone to make the best change at each individual step independent of future effects. This can produce unstable models where relatively small changes in the data is likely to cause one variable to be selected instead of another, after which subsequent choices may be completely different. Best-subset selection is an attempt to address this limitation by considering all subsets of variables of each size only limiting itself

to a maximum number of best predictor subsets (Furnival and Wilson, 1974). Given p variables, best-subset selection finds the subset of size $k \in \{0, 1, 2, \ldots, p\}$ that provides smallest residual sum of squares. A distinct advantage over stepwise procedures is the fact that with best-subset regression, the best set of two predictors need not include the predictor that was best when considered in isolation. However, because it considers a much greater number of possible models, biases in inference are even larger (Draper and Smith, 1998).

Although overfitting often results from too many predictors, using very few variables may fail to reveal the true underlying structure of a prediction model from inadequate information. Generally a saturated model often outperforms reduced models. When using stepwise selection procedures, Harrell (2001) suggested the need for less stringent stopping rules like Akaike's information criterion (AIC) to decide candidate variables to retain or discard. When using backward elimination selection, Steyerberg et al. (2000) proposed using a p-value of 0.5 to allow deletion of some variables. In general, backward elimination performs better than forward stepwise selection procedures in the presence of multicollinearity (Mantel, 1970). Moreover, backward elimination initially allows examination of the full model which has the correct standard errors and p-values. Later we consider use of backward elimination procedures to obtain reduced models. Stepwise selection procedures are implemented in most statistical softwares. Best-subset regression may be implemented using the `leaps` package in R[1] software (Ihaka and Gentleman, 1996).

3.3.1 Limitations of stepwise selection procedures

While it may be objective to consider a subset and not individual potential predictors, the best-subset selection method unlike stepwise methods fails in reducing dimension by selecting more predictors. Moreover, the best subsets selection method can compare only the models with the same number of predictors (Draper and Smith, 1998) hence restricting the number of models that can be compared. The discrete process (variables either retained or discarded)

[1] http://www.r-project.org/

inherent in the best-subset selection often exhibits high variance (Hastie et al., 2009) resulting no reduction of the prediction error in the full model.

Despite their wide application in practice, stepwise selection procedures are associated with many limitations (Austin and Tu, 2004; Wang et al., 2004). The principal drawbacks of stepwise multiple regression include bias in parameter estimation, inconsistencies in results from different model selection algorithms, and multiple hypothesis testing before estimation. Although parsimony may be a desirable statistical practice, reliance on a single best model may lead to loss of information from excluded predictors. Even smaller models with few predictors is no guarantee of exclusion of noise variables. Excluding important predictors can be very costly, for example, Steyerberg et al. (1999) demonstrated that excluding a true predictor is worse than including a noise variable.

The stepwise selection procedures are based on test of hypothesis of individual parameters. The process that produces the 'final' model fails to account for the inherent multiple testing. In addition to 'testimation' bias, the resulting standard errors are also invalid because stepwise procedures fail to fully account for the search process (Harrell, 2001). Analysis of the 'final' model from stepwise selection procedures assumes that the selected predictors were pre-specified which is not true since the predictors were selected adaptively according the selection algorithm. Consequently comparison of any models produces biased results because the analysis erroneously assumes the two models were fixed in advance. Specifically, variance of the regression coefficients calculated as if the selection were pre-specified will underestimate standard errors and p-values in the resulting model (Harrell, 2001). The problem is even acute for small data sets where stepwise procedures have limited power to select prognostically important predictors which in turn lead to lower predictive ability (Chatfield, 1995).

The final model derived from stepwise selection procedures are dependent upon the correlation between individual predictors (Draper and Smith, 1998). This single model is not guaranteed to be the best among candidate models and interpretation using such a model includes only those predictors entered in that final model while ignoring other predictors not selected. Besides correlation, the final model depends on the order of entry/exit of predic-

tors into the model. Such a dependency is likely to result in inconsistencies among different model selection algorithms. In general, excluding potential important predictors on account of little/no correlation may lead to suboptimal decisions and limited predictability. The next section reviews multivariate regression techniques which attempt to address the problem of many predictors with high degree of correlation.

3.3.2 Principal component and partial least squares regressions

Ordinary least squares (OLS) estimation of regression coefficients in the presence of more predictors than number of observations and/or high degree of near collinearity among the predictors performs poorly by producing very unstable estimates and very poor prediction accuracy (Draper and Smith, 1998). Principal component regression (PCR) and partial least squares regression (PLSR) provide a solution to both challenges by using linear combinations of predictors instead of individual predictors (Vigneau et al., 1997). Strictly speaking PLSR is a generalization of PCR.

The difference between PCR and PLSR lies in the different ways they construct new predictor variables (components) as linear combinations of the original predictor. While PCR creates components to explain the observed variability in the predictor X variables only, PLSR creates components to explain variability in both the predictor X and response Y variable. As a results PLSR becomes PCR if it ignores the response during the process of creating components to explain observed variability. In constructing the principal components of X, the PLSR algorithm iteratively maximizes the strength of the relation of successive pairs of X and Y component scores by maximizing the covariance of each X-score with the Y variables. Because of its general strategy, PLSR is sometimes called Projection to Latent Structures in the natural sciences (Abdi, 2010). The distinct advantage of both methods lies in their use of linear combinations which leads to models that are able to fit the response variable with fewer components. We however note that whether or not this reduction ultimately translates into a more parsimonious model, in terms of its practical use,

depends on the context. While PCR is a popular technique among social scientists, PLSR enjoys large popularity among the natural sciences particularly in chemommetrics (computational chemistry) where it is heavily used in chemical analysis following developments in spectroscopy (many highly correlated predictors) since the 1970s (Mevik and Wehrens, 2007).

PLSR is a generalized multivariate statistical technique with ability to model multiple predictors as well as multiple responses, handle multicollinearity among predictors and help make stronger predictions by creating independent components/latent variables directly on the basis of cross-products involving the response variable/s. Some of its limitations include greater difficulty of interpreting the loadings of the independent latent variables (which are based on $X-Y$ cross-product relations, not based as in common factor analysis on covariances among the manifest predictors). The mix of advantages and disadvantages makes PLSR more appealing as a predictive technique and not as an interpretive technique tool.

Theoretically PLSR should have an advantage over PCR but in most situations in practice both methods achieve similar prediction accuracies (Wold et al., 2001). While PLSR usually needs fewer latent variables than PCR (i.e. with the same number of latent variables) and will cover more of the variation in the response Y, PCR will cover more of the variances in predictor/s X. Frank and Friedman (1993) showed that both PCR and PLSR behave very similar to ridge regression (next section). Although Hastie et al. (2009) showed that both PCR and PLSR behave as shrinkage methods (next section), in some cases PLSR seems to increase the variance of individual regression coefficients, an observation which may mean that PLSR is not always better than PCR. PLSR can be implemented as a regression model to predict one or more responses from a set of one or more predictors using `pls` package in R software (Mevik and Wehrens, 2007).

3.4 Penalized regression techniques

Biased estimates may be more preferable than unbiased estimates in multivariate situations to make better predictions. This phenomenon is what is commonly referred to as Stein (1981) paradox owing to the fact that it appears to negate the principle of unbiased estimation. Shrinkage is the principle of reducing/penalizing the regression coefficients to improve quality of predictions (Steyerberg, 2009). In the presence of multicollinearity, techniques that penalize regression coefficients can help improve ordinary least squares (OLS) estimates in terms of both prediction and interpretation (Draper and Smith, 1998). Ridge regression is a suitable alternative to OLS in the presence of multicollinearity among predictors is of primary concern (Hoerl and Kennard, 1970). While OLS estimation in the presence of multicollinearity is likely to produce poorly determined coefficients where extreme positive coefficients on one variable can be canceled by similarly large negative coefficient on related variables, ridge regression addresses this problem by imposing a size constraint on the coefficients. In effect ridge regression includes all candidate predictors, but with considerably smaller (shrunk) coefficients compared to those from OLS (Draper and Smith, 1998; Miller, 2002).

By continuously penalizing the coefficients, ridge regression achieves better prediction performance through a bias-variance trade-off by minimizing the residual sum of squares subject to the sum of the squares of the coefficients bounded by a constant i.e. $\sum_{j=1}^{p} \beta_j^2 \leq t$, where t is the constant. The ridge regression minimizes a penalized OLS:

$$\hat{\beta}_{\text{Ridge}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \lambda \geq 0 \tag{3.1}$$

where λ is a positive scalar called penalty and is often chosen by cross-validation in practice. We note that $\lambda = 0$ corresponds to OLS regression. Bias would increase with an increase in λ. Equation (3.1) can be equivalently represented in matrix form as:

$$\hat{\beta}_{\text{Ridge}} = [X'X + \lambda I]^{-1} X'Y.$$

The motivation of ridge regression lies in the addition of a positive constant (λI) to the diagonal of the $X'X$ matrix before inverting it which makes the problem nonsingular even if $X'X$ is not of full rank. The variables used in ridge regression are standardized (mean=0, variance=1) to make the penalty invariant to the scale of the original data. No penalty is applied to the intercept because to do so would make the procedure dependent on the origin chosen for the response. However, standardizing variables makes the resulting ridge parameters difficult to interpret between models because the parameters are not on natural scale. While the ridge parameter may not be the most useful for understanding the amount of shrinkage performed, we can use the effective degrees of freedom (df_λ) to measure the impact of the penalty, where

$$\mathrm{df}_\lambda = \mathrm{trace}\left[X\left(X'X + \lambda I\right)^{-1} X'\right]. \tag{3.2}$$

We note from the definition in Equation (3.2) above that for a model fitted with p variables, $\mathrm{df}_0 = p$ (OLS model) and $\mathrm{df}_\lambda \to 0$ as $\lambda \to \infty$ (null model). In general, ridge regression can do better than OLS on the full model provided we do not over-shrink (add more bias).

Although ridge regression produces shrunken regression estimates, it fails to reduce the number of predictors since it always keeps all the predictors in the model. Comparatively Breiman (1996) showed that while best-subset selection may produce a sparse model (fewer predictors) than ridge regression, the former model is extremely variable (unstable) because of its inherent discreteness unlike the later model which is more likely to be stable owing to its continuous penalization. However, Harrell (2001) point out that ridge regression may not be suitable for categorical predictors where the amount of shrinkage is dependent on the choice of reference cell when creating dummy variables.

Tibshirani (1996) introduced the Least Absolute Shrinkage and Selection Operator (Lasso) as an alternative to ridge regression with the distinct advantage that it accomplishes the twin tasks of model selection and coefficient shrinkage for finite values of the penalty parameter λ. The Lasso minimizes the penalized OLS subject to the sum of the absolute values of the

parameters bounded by a constant (i.e. $\sum_{j=1}^{p} |\beta_j| \leq t$):

$$\hat{\beta}_{\text{Lasso}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \lambda \geq 0. \tag{3.3}$$

3.4.1 Variable selection using the Lasso

Asymptotic considerations of the Lasso leads to an interesting extension of using Lasso as a variable selection method. Frank and Friedman (1993) showed that if the Lagrange term $\lambda \sum_j |\beta_j|$ in expression (3.3) is replaced by $\lambda \sum_j |\beta_j|^q$ where $q \to 0+$, then the minimization will be identical to optimal variable (best-subset) selection. They however noted the difficulty in obtaining a unique solution of the function from the many local minima (up to $2^p - 1$), i.e. no closed form hence nonlinear solution.

In practice, variable selection using Lasso is a two-step procedure: in the first step, the Lasso algorithm identifies/selects the important predictors (non-zero parameter estimates), and step two applies ordinary linear regression to the selected predictors, i.e. if the matrix $\tilde{X}$ denotes the matrix containing only those columns j of X for which the lasso estimate $\hat{\beta}_j \neq 0$, then the variable selection estimate is defined as

$$\hat{\beta}_{\text{Lasso}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y.$$

The steps above suggest a connection between Lasso and the ordinary variable selection. But Öjelund et al. (2001) pointed out the difference between Lasso and Lasso as a variable selection method is that in variable selection only the parameters that turn out to be zero are penalized (set to zero) while the others are estimated without any penalization. The number of non-zero parameters will be the same for the two methods under a uniform shrinkage that uses the same value of λ. However, in practice we often obtain fewer predictors when Lasso is used as a variable selection method because we often deliberately select a larger value of λ for Lasso.

3.4.2 Bayesian generalization of ridge and Lasso regressions

We observe that ridge regression and the Lasso differ only in the penalty form imposed on the regression coefficients. Specifically ridge regression uses the squared L_2 norm of parameters while Lasso uses the L_1 norm. The bridge regularization (Frank and Friedman, 1993) is a generalization with best-subset, ridge regression and Lasso as its special cases:

$$\hat{\beta}_{\text{Bridge}} = \underset{\beta}{\overset{\text{argmin}}{}} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \quad q \geq 0, \tag{3.4}$$

so that

$$\begin{aligned}
&L_0 \text{ norm penalty}: && L_0(\beta) = \textstyle\sum_{j=1}^{p} I(\beta_j \neq 0) && \Rightarrow \text{ Best-subset,} \\
&L_1 \text{ norm penalty}: && L_1(\beta) = \textstyle\sum_{j=1}^{p} |\beta_j| && \Rightarrow \text{ Lasso, and} \\
&L_2 \text{ norm penalty}: && L_2(\beta) = \textstyle\sum_{j=1}^{p} \beta_j^2 && \Rightarrow \text{ Ridge regression.}
\end{aligned}$$

In theory the L_0 penalty is the ideal choice for variable selection because it directly penalizes the size of model (design matrix is orthogonal). A choice of $q \in (0, 1]$ regularly gives estimates that are 0 for some of the β_js. In particular, Lasso as a model selection method that penalizes the number of parameters can be seen as limiting cases of bridge estimation as $q \to 0$ because $|\beta_j|^q \to 0, 1$ accordingly as $\beta_j = 0$ or $\beta_j \neq 0$ when $q \to 0$ (Clarke et al., 2009).

From a Bayesian perspective, the bridge equation (3.4) provides Bayesian estimates where we can think of $|\beta_j|^q$ as the log-prior density for β_j (i.e. the best-subset selection) whereas ridge regression and Lasso can be considered as Bayes estimates with different priors: the independent double exponential (Laplace) distribution is the prior corresponding to Lasso ($q = 1$) and the ridge regression is the posterior mode (and mean, its Gaussian) due to its quadratic formulation (Hastie et al., 2009). Here $q = 0, 1, 2$ respectively corresponds to best-subset selection where the penalty λ simply counts the number of nonzero parameters, the Lasso and ridge regression. The Lasso ($q = 1$) provide the smallest q such that the constraint

region is convex hence making the optimization problem less difficult, i.e. the function has only one minimum (Knight and Fu, 2000; Kyung et al., 2010; Park and Casella, 2008), an observation that may be supportive of its twin roles in coefficient shrinkage and model selection.

3.4.3 Penalized maximum likelihood estimation

While ordinary least squares estimation is adequate for linear models with continuous outcomes, optimal fit to data in generalized linear models is obtained by maximizing the log likelihood. However, in the presence of many predictors, maximizing the log likelihood often results in fitting noise and parameter estimates that are unstable (Moons et al., 2004). By trusting too much in the trends of such data, the maximum likelihood estimation (MLE) may produce spurious effects not reflective of the true data. Penalized maximum likelihood estimation (PMLE) is a generalization of the ridge regression method used to obtain more stable parameters for linear regression models (Draper and Smith, 1998). Minimizing the PMLE instead of the MLE is another bias-variance tradeoff technique that can be used to produce stable models in the presence of many predictors:

$$\log L - 0.5\lambda \sum (s_i \beta_i)^2, \tag{3.5}$$

where L is the model's MLE adjusted (shrunk) by a penalty factor λ, β estimates regression coefficients and s_i is a scaling/shrinkage factor for each β_i to make $s_i\beta_i$ unitless. A major advantage of PMLE is the direct adjustment (during the model fitting) for overoptimism of the estimated regression coefficients and predictive accuracy measures (Harrell, 2001; van Houwelingen, 2001; Verweij and van Houwelingen, 1994).

Choosing the optimal penalty factor λ is a challenge in PMLE. While various methods including sample-splitting and cross-validation are often used to estimate the penalty factor, Harrell (2001) pointed out that maximizing a modified AIC method is a more efficient method. With r degrees of freedom (the number of fitted predictors), the AIC is traditionally used to

penalize the maximum log likelihood of a fitted model such that the $\log L$ in Equation (3.5) is often replaced by $\log L - r$, where L is the maximum likelihood of the fitted model. We can translate the AIC to the model χ^2, i.e., the difference in log likelihood between the model with and without predictors) scale so that

$$\chi^2_{\text{LR}} - 2r, \tag{3.6}$$

where χ^2_{LR} is the likelihood ratio χ^2 of the penalized model, but ignoring the penalty function.

Equation (3.5) defines AIC for ordinary maximum likelihood estimation. To find the optimum penalty factor for PMLE, Harrell (2001) proposed using a modified AIC defined as

$$\chi^2_{\text{LR}} - 2 \times \text{df}_\lambda, \tag{3.7}$$

where χ^2_{LR} is the likelihood ratio χ^2 of the penalized model and df_λ is the effective degrees of freedom, i.e. degrees of freedom after penalizing the fitted predictors with penalty λ. We note that the the higher the number of predictors, the higher the degrees of freedom and the more likely the model is overfitted. Ordinarily, the degrees of freedom is equal to the number of predictors, but penalization ensures that the degrees of freedom effectively used in PMLE is lower than the actual number of predictors hence decreasing the potential for overfitting.

In practice, we use a trial and error process over a wide grid to find the optimal penalty factor. We obtain the optimal penalty factor λ_{opt} as the penalty that maximizes the modified AIC. Using equation (3.5), the fitted model is penalized with λ_{opt}. This penalty is also called the tuning parameter and its estimate decides the model's complexity of the prediction method used.

In matrix notation, equation (3.5) can be represented as $\log L - 0.5\lambda\beta' P\beta$, where P is the so-called penalty matrix often composed of covariances of the standardized coefficients. van Houwelingen and le Cessie (1990) showed that a uniform shrinkage factor s can be ob-

tained from equation (3.5) using the relation:

$$s = \frac{\text{model } \chi^2 - r}{\text{model } \chi^2}, \tag{3.8}$$

where as before r denotes the degrees of freedom (df) of the predictors in the model and χ^2 for the model is calculated on the log-likelihood scale. We observe that $s \to 0$ for larger numbers of predictors (r increases) or when the sample size is smaller (model χ^2 decreases). In general, penalization aims at an improved predictive performance in a new data set by balancing the fit to the data and the stability of the estimates.

For the present study, we will apply PMLE of the original model composed of predictors considered meaningful and use the modified AIC (3.7) where $\chi^2 > 2r$ at each selection iteration to obtain intermediate models for comparison of fewer predictors. We will then use relation (3.8) to calculate uniform shrinkages as measures of quality of prediction fit (see next section for overoptimism). Using a wide grid search over the $0 - 50$ range, we will obtain the penalized coefficients and report optimal penalty factors λ_{opt} for each model.

A model selection procedure is likely to produce competing models. So how do we choose from such a selection and what criteria do we use? The next section provides a brief overview of validation measures that can be used to evaluate the performance of selected prediction models.

3.5 Validation performance: optimism, calibration and discrimination

One of the negative consequences of overfitting is optimism which means that there is a too optimistic impression of model performance that may be achieved in new subjects from the underlying population. Optimism can be defined as the difference between true performance and apparent performance (Steyerberg, 2009). Here the true performance refers to model behaviour in the underlying population and apparent performance refers to the estimated performance in the sample.

Validation of statistical prediction models means the evaluation of their predictive ability. Following model development, we might be able to assess performance of prediction in new/independent patients, i.e. external validation (Harrell, 2001; Hastie et al., 2009; Steyerberg, 2009). However, if further data is unavailable, we might seek to use original sample of data for model development and validation. One approach is to randomly split the data into two portions: one for model development/training and the other for model validation. But such data splitting procedures are often considered suboptimal (Efron and Tibshirani, 1993) and may fail to demonstrate the adequacy of the model when subjected to external validation given the possibility to produce less representative subsamples for either model development and/or validation. Bootstrap is considered the most efficient validation procedure (Harrell, 2001). As a validation technique, bootstrap repeatedly analyses subsamples of a given data where each subsample is a random sample with replacement from the original sample. The fact that each bootstrap samples will be relatively different than the original data set mimics application to a different sample at every iteration hence providing reliable results (Anderson, 2005). Bootstrap validation allows calculation of predicted probabilities from a model which can be compared with the actually observed outcomes.

Calibration and discrimination statistics are among the most effective and commonly used measures of validation performance (Harrell et al., 1996; Steyerberg et al., 2004). Calibration (also called reliability) refers to how well the model predictions compare with the observed outcomes. Calibration is essentially a measure of bias that evaluates the agreement between observed and predicted probabilities. For example, if the average predicted proportion of compliers to HRT tablet allocations among similar group of women is 80% and the actual proportion complying is 80%, then the predictions are well calibrated. In measuring calibration, it is sufficient to focus on either an intercept or slope of a linear predictor and not both (Steyerberg et al., 2004). However, in practice calibration is often quantified by the calibration slope as originally proposed by Cox (1958b). The calibration slope can be obtained from the validation plot which is a plot of observed probabilities against the predicted probabilities. The line from the validation plot can be defined with an intercept α

and a slope β, where $\alpha = 0$ and $\beta = 1$ corresponds to a perfect calibration. The calibration slope lies between 0 and 1 and the bigger, the better calibrated the model under study.

There is a relationship between the calibration slope and penalty factor in penalized regression (Miller et al., 1993). For example, for a logistic model with the linear predictor as the only covariate, the calibration slope is the estimated regression coefficient β, i.e.

$$\text{logit}(\text{treatment compliance}) = \alpha + \beta \cdot \text{linear predictor}. \tag{3.9}$$

Copas (1983) and van Houwelingen and le Cessie (1990) demonstrated that the slope β of the linear predictor is identical to the uniform shrinkage factor s given by Equation (3.8) above.

Discrimination refers to the ability of the model to distinguish between subjects with positive or negative outcomes (e.g. the ability of a model to distinguish compliers with treatment allocation from non-compliers). Discrimination is commonly measured using the concordance (c)-statistic. For binary outcomes this statistic is identical to the area under the receiver operating characteristic curve (Harrell, 2001). Given a random pair of patients with different outcome values, the c-statistic can be interpreted as the likelihood of a patient with the desirable outcome (complier) to have a higher predicted probability for having that outcome than a patient without the outcome (e.g. non-complier). The c-statistic varies between 0.5 (random predictions) and 1.0 (perfect prediction) and the higher the better (Harrell et al., 1996; Miller et al., 1993). The concordance c-statistics can also be expressed in terms of the widely used Somers (1962) D_{xy} rank correlation which is a measure of the difference between concordance and discordance probabilities (Harrell, 2001):

$$D_{xy} = 2(c - 0.5), \tag{3.10}$$

where $D_{xy} = 0$ and 1 here implies random predictions and perfect discriminations respectively.

We will evaluate the predictive performance of our models using calibration slope, calculate discrimination's concordance c statistics from the reported D_{xy} value and the percentage

of optimism as implemented in the `Design` package in R software. However, it is worth noting that good (or even perfect) calibration and discriminative ability are not sufficient for a model to be declared clinically useful. Only a model's ability to provide useful additional information for clinical decision making makes application of a prediction model sensible. In our case the quality of compliance information for both treatment arms would be crucial to making relevant clinical decision.

Following the review, the next five chapters provide applications of the methods discussed thus far. Chapters 4, 5 and 6 provide analyses of the Esprit data. On the other hand Chapters 7 and 8 provide simulation studies which evaluate respectively the performance of specialist methods adjusting for noncompliance in one treatment arm and Roy et al. (2008) method of principal stratification adjusting for noncompliance in two treatment arms.

Chapter 4

Esprit Analysis I: Modelling Noncompliance Effect in One Arm

4.1 The Esprit study background and data

The ITT analysis of the oEStrogen in the Prevention of ReiInfarction Trial (Esprit) data was published in 2002 (Cherry et al., 2002). The aim of the Esprit study was to ascertain whether or not unopposed oestrogen reduces the risk of further cardiac events in postmenopausal women who survive a first myocardial infarction. Previous observational studies suggested that hormone replacement therapy (HRT) prevented cardiac events but there had been no randomized clinical trails on the same. Evidence from observational studies also showed that unopposed oestrogen increased the risk of endometrial cancer although the addition of progesterone reversed this effect. Additionally, there was evidence that oestrogen reduced risk to fractures. There was evidence of increased risk of breast cancer in long term users (> 5 years). Organizers of the Esprit trial argued that the benefits (in terms of cardiovascular deaths) of giving unopposed oestrogen to a group at high risk of second infarction could outweigh any increased risk of endometrial cancer. The study comprised a total of 1017 women aged between $50 - 69$ years who were recruited from 35 hospitals in England and Wales. 513 and 504 women were randomized to the treatment and placebo arms respectively and monitored over 24 months (2)-year period. The primary outcomes were reinfarction or

cardiac deaths and all-cause mortality. Although ITT analysis of the data has been previously published, the analysis took no account of compliance data.

This chapter provides new set of results for the Esprit trial taking account of compliance data as well as, for comparison, non-randomization based methods. Our analysis considers two outcomes: all-cause mortality and myocardial reinfarction or cardiac deaths. All the analysis in this Chapter are performed using Stata software (StataCorp, 2008) and the Stata codes are shown in Appendix page 261.

4.2 Methods

We use five classes of methods to evaluate HRT treatment. First we perform primary analyses by intention-to-treat (ITT) with and without covariates. The six covariates included here were agreed in the protocol prior to the study. They were chosen (before data inspection) as factors likely to have a relation with the primary endpoints. We considered age (in years) at entry, body mass index (BMI) as continuous variables. Smoking habit at recruitment (never, ex-smoker, intending to give up, or current smoker) was also included. We classified participants as smokers and non-smokers where the later group consisted those of who had never smoked and those who gave up smoking at least in the last 2 years. Other covariates considered were reported binary (yes/no) histories of hysterectomy, high blood pressure and diabetes. The ITT analysis compared the rates of non-fatal reinfarction or cardiac death in the 2 years after study entry, with observation time censored at reinfarction, death, or 24 months, whichever occurred first. We used the Cox proportional hazards model to obtain adjusted and unadjusted hazard ratios.

Secondly, we use per-protocol and as-treated methods to explore treatment efficacy although these methods are regarded as suboptimal as discussed earlier. We assume perfect compliance with placebo while considering compliance with HRT treatment and classify a woman as a complier per-protocol if she strictly complied with assignment without interruption up to the end of month 23 of the study or until the time of event (death/reinfarction).

As-treated analysis was a comparison of outcomes based on current treatment status, i.e. comparing those women on treatment to those who are not. We also perform simple regression adjustments to obtain 'naive' treatment effects by using a time-invariant and time-varying binary all-or-nothing noncompliance indicator as a covariate.

Thirdly we considered three specialist methods which attempt to provide unbiased HRT treatment effect estimates by adjusting for all-or-nothing and partial noncompliance to treatment assignment. For each of these methods we consider compliance with the active HRT treatment while ignoring compliance with placebo, i.e. we evaluate treatment effects among those women who actually took their HRT tablets given that they were randomized to the treatment arm. For exploration and in attempt to describe the data, we subdivided the data into 5 short intervals ($0-3$, $3-6$, $6-12$, $12-18$ and $18-24$ months) and assume constant piecewise hazard rates for each interval. This subdivision with utilize compliance information reported for respective intervals. We define compliance as actual taking of HRT tablets up to a day before experiencing event of interest (death/reinfarction) or end of study, whichever occurred first. The Esprit study did not permit treatment switches (from placebo to active) and had no information on whether patients intermittently stopping and starting of medication. Ignoring such scenarios together with our compliance definition may be considered plausible with respect to the exclusion restriction assumption in the absence of carryover effects.

We used all-or-nothing compliance for the C-Prophet method. Although the CALM method allows for multiple crossovers between treatment arms, our analysis for the present work considered a unidirectional crossover where patients who stopped taking their tablets became non-compliers and we utilized the time to stoppage of medication and a potential recensoring time specified at end of study (24 months). The C-Prophet, CHARM and CALM methods were implemented using the Stata commands `stcomply` (Kim and White, 2004), `adjhr` (White, 2002) and `strbee` (White et al., 2002) respectively (codes are shown on page 261). For comparison and completeness, we also obtained a simple CHARM estimate by using the treatment arm to estimate the counterfactual compliance in the control arm and then adjusting the ITT estimate with the proportion of potential non-compliers who experienced

the event of interest (White et al., 2004).

4.3 Results

There were 71 all-mortality deaths in total during the 2-year study period, i.e. 32 and 39 from treatment and placebo arms respectively. On the other hand there were a total of 123 myocardial reinfarction or cardiac deaths: 62 and 61 in the active treatment and placebo arms respectively. Most of the all-mortality deaths (reinfarctions) were registered during the early periods of the study with more than 64% (75%) events by month 10. While month 3 recorded the highest all-cause mortality rate at 15%, the highest myocardial reinfarction or cardiac deaths were recorded in the first month (23%). We analyse data for all-cause mortality outcome to explore individual intervals for the Esprit data.

Table 4.1: All-cause mortality: incidence rates and survival distribution per interval

Intvl.	Number alive at start of interval		Survival function			Incidence rates		
	Treat (Dead)	Plcbo (Dead)	Treat	Plcbo	Overall (Dead)	Treat	Plcbo	Overall
0-3	506 (8)	491 (14)	0.984	0.972	0.978 (22)	0.00532	0.00936	0.00728
3-6	500 (6)	485 (6)	0.972	0.960	0.967 (12)	0.00399	0.00411	0.00405
6-12	494 (6)	475 (10)	0.961	0.941	0.951 (16)	0.00202	0.00348	0.00274
12-18	487 (7)	466 (9)	0.948	0.923	0.935 (16)	0.00238	0.00320	0.00278
18-24	481 (5)	465 (0)	0.938	0.923	0.930 (5)	0.00172	0.00000	0.00088
0-24	513 (32)	504 (39)	0.938	0.923	0.930 (71)	0.00270	0.00341	0.00305

Table 4.1 provide the incidence rates and survival distribution for each of the 5 intervals for all-cause mortality. The first 3 months recorded the highest incidence rates (IR) compared to subsequent 6-months IR (Table 4.1) and even the overall 24 months IR. The first 3-months IR are about double (0.005 and 0.009 treatment and placebo respectively) compared to the next 3-months IR (0.004 and 0.004). In general, Table 4.1 shows that mortality fell (throughout) as the study progressed. In addition, the incidence rates in the treatment group

was lower that in the placebo group for all intervals except the last one, $18-24$ months.

Table 4.2: All-cause mortality: ITT hazard rates for each interval

Interval	Number dead		Haz. ratio	SE	P-value	95% CI
	T(%)	P(%)				
0-24	32 (6.2)	39 (7.7)	0.795	0.190	0.335	0.498, 1.268
0-3	8 (1.6)	14 (2.8)	0.559	0.248	0.189	0.234, 1.332
3-6	6 (1.2)	6 (1.2)	0.971	0.561	0.959	0.313, 3.010
6-12	6 (1.2)	10 (2.1)	0.580	0.300	0.292	0.211, 1.597
12-18	7 (1.4)	9 (1.9)	0.744	0.375	0.557	0.277, 1.997
18-24	5 (1.0)	0	5.08×10^{-15}	1.52×10^{-23}	0.991	0 -
12-23	12 (1.6)	9 (1.2)	1.277	0.563	0.580	0.538, 3.030

T≡treatment; P≡placebo

There were 22 all-cause mortality (ACM) events in interval 1 with treatment and placebo groups experiencing 8 and 14 deaths respectively. The last month of this interval (month 3) accounted for 11 (50%) of the total ACM in this interval. Overall, ACM during interval 1 accounted for almost 31% of the total number of deaths during the study period. Although the treatment effect is not statistically significant (HR= 0.56, p-value= 0.189, 95% CI:$0.234, 1.332$), the treatment appeared to be beneficial compared to placebo and comparatively reduces the risk of ACM by about 44% (Table 4.2).

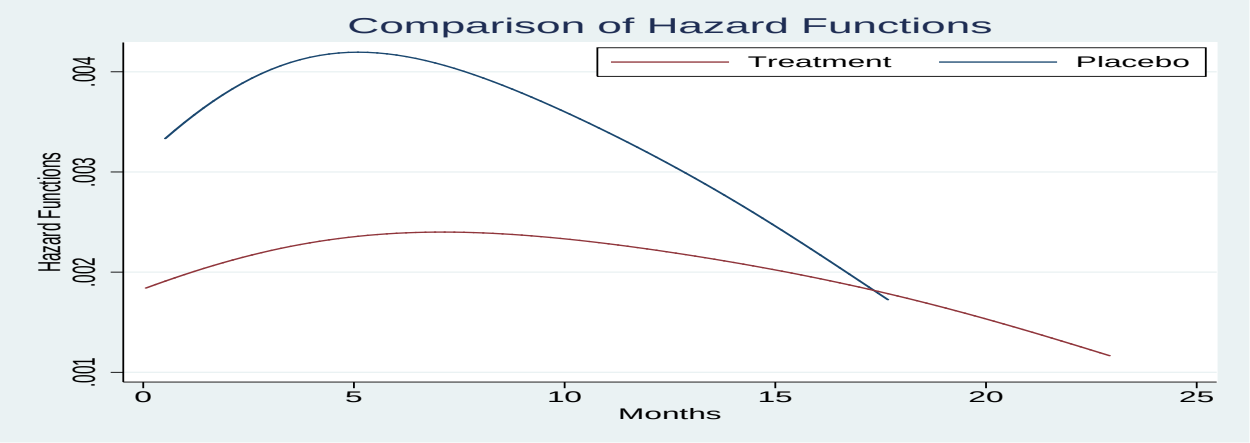
A total of 12 patients experienced ACM during the second interval. This constituted a 45% reduction from 22 ACMs in the previous interval. The 3-months IR for this interval appear similar at about 0.004 for both treatment and placebo groups as Table 4.1 shows. The equal number of ACMs (6 each) from both participants under treatment and placebo may be an apparent manifestation of a stable IR across the intervals. Half of the ACMs in interval 2 occurred during month 4. The 6 ACMs here were also evenly distributed between treatment and placebo groups. However, there are only 2 ACMs from the placebo group by the end of this interval. The piecewise hazard ratio for interval 2 is 0.97 (p-value= 0.959, 95% CI:$0.313, 3.010$). Hence despite being statistically insignificance, the treatment appears to marginally reduce the risk for ACM by about 3%.

There were 16 ACMs recorded in both intervals 3 and 4. Both intervals also recorded less ACM from the treatment group, i.e. 6 and 7 compared to 10 and 9 from treatment and placebo groups for respective intervals. The 6-months IR for both intervals almost remain the same at about 0.002 and 0.003 for treatment and placebo groups respectively. This similarity may be the source of stability and near-equal mortality rates at both intervals. However, although both the treatment effects are not statistically significant at both intervals, the hazard ratio is apparently higher at interval 4 than at interval 3, i.e. treatment has more than 40% chance to reduce the risk to ACM in interval 3 compared to a reduction of about 25% in interval 4.

There were a total of 5 ACMs in the last interval ($18-24$-months). All these ACMs were exclusively registered from the treatment arm. The 6-months IR for this interval is the lowest for the study at 0.002 for the treatment arm. 2 (40%) ACMs in this interval were registered in month 19, while months $21, 22$ and 23 each had one ACM. There were no ACMs registered in the 20^{th} and 24^{th} months. In attempt to utilize all the data, we pooled both intervals 4 and 5 into a single 12-months long interval that now has 21 ACMs. Hence we now consider a total of 4 intervals.

A comparison of the smoothed hazard functions (Figure 4.1) indicates relatively lower hazard rates for treatment compared to placebo. This is an indication of some form of benefit derivable from active treatment.

Figure 4.1: All-cause mortality: smoothed hazards for the 24 months



4.3.1 ITT analysis

The original analysis (Cherry et al., 2002) revealed statistically non-significant effects of HRT with regard to both all-cause mortality ($HR_{ACM} = 0.790$, p-value$= 0.340$, 95% CI: $050, 1.27$) and myocardial reinfarction or cardiac deaths ($HR_{MRCD} = 0.990$, p-value$= 0.970$, 95% CI: $070, 1.41$).

For both all-cause mortality and myocardial reinfarction or cardiac deaths, the ITT analysis considered the model below comprising 6 baseline covariates

$$\mathrm{h(t)} = \mathrm{h}_0(t) \exp\left(\begin{array}{c} \text{Treatment} + \text{Hysterectomy} + \text{Age} + \text{BMI} \\ +\text{Smoking} + \text{Blood Pressure} + \text{Diabetes} \end{array} \right). \tag{4.1}$$

The mean age at admission was 62.6 years, the youngest and oldest were about 50 and 70 years old respectively. The average BMI was about 26.7 kg/m^2 where the lowest and and largest BMI stood at 14.8 kg/m^2 and 51.3 kg/m^2 respectively. A total of 244 (24%) women had a history of hysterectomy.

Table 4.3: PH ITT analysis: all-cause mortality and myocardial reinfarction

	All-cause mortality (ACM)				Myocardial reinfarction (MRCD)			
Covariate	HR	SE	P-val	95% CI	HR	SE	P-val	95% CI
Treatment†	0.795	0.190	0.335	0.498, 1.268	0.993	0.179	0.967	0.697, 1.414
Treatment‡	0.805	0.195	0.369	0.501, 1.293	0.995	0.181	0.978	0.696, 1.422
Hysterectomy	0.436	0.156	0.021	0.216, 0.881	0.785	0.174	0.275	0.508, 1.213
Age (in yrs)	1.068	0.029	0.017	1.012, 1.127	1.024	0.020	0.227	0.985, 1.064
BMI	1.011	0.023	0.623	0.967, 1.058	1.027	0.018	0.123	0.993, 1.062
Smoking	1.113	0.280	0.671	0.680, 1.822	1.110	0.214	0.589	0.761, 1.619
B-Pressure	1.197	0.294	0.464	0.740, 1.937	1.587	0.300	0.015	1.096, 2.299
Diabetes	3.724	0.971	¶	2.234, 6.207	2.718	0.556	¶	1.820, 4.058

HR≡ hazard ratio; †no covariates;‡with covariates; ¶p-value < 0.001

Table 4.3 provide results from the analysis of model (4.1) for both all-cause mortality and myocardial reinfarction or cardiac deaths. For both outcomes, the HRT treatment is not significantly different from placebo, both with and without covariates. The hazard ratios (HR) for treatment are similar regardless of whether covariates are included or not. In general, the

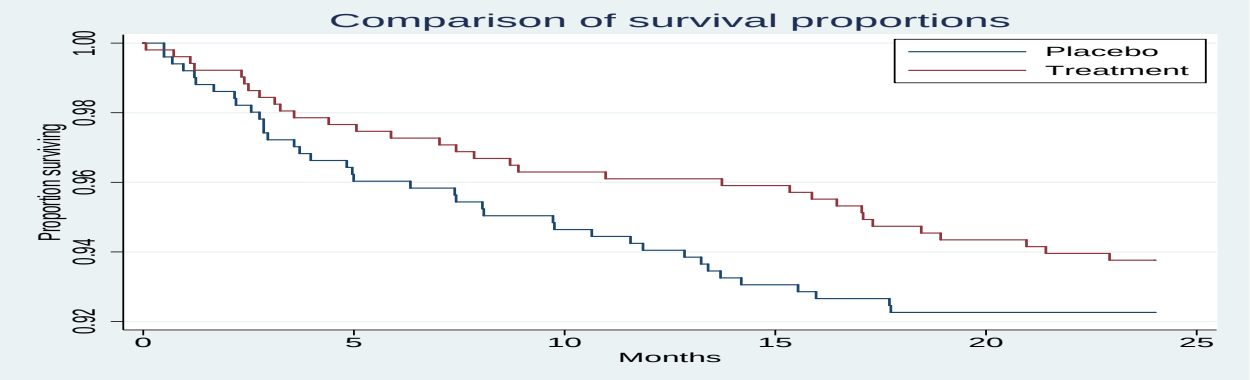
age of women at admission, histories of diabetes and hysterectomy are significantly related to survival from all-cause mortality (at 5% level of significance). On the other hand histories of blood pressure and diabetes were statistically significantly related to survival from reinfarction or cardiac death. Although statistically insignificant, the analysis shows benefit of active treatment in reducing the risk of all-cause mortality by about 20% (HR$=0.80$, p-value$=0.335$, 95% CI: $0.498, 1.268$). However, the analysis showed no difference in effects between the active treatment and placebo with regard to the risk of myocardial reinfarction or cardiac death (HR$=0.99$, p-value$=0.965$, 95% CI: $0.697, 1.414$). Also Table 4.4 shows we obtain similar ITT results for the Esprit data under AFT (Weibull on HR scale) analysis.

Table 4.4: AFT (Weibull) ITT analysis: all-cause mortality and myocardial reinfarction

	All-cause mortality (ACM)				Myocardial reinfarction (MRCD)			
Covariate	HR	SE	P-val	95% CI	HR	SE	P-val	95% CI
Treatment†	0.794	0.189	0.334	0.498, 1.268	0.994	0.179	0.973	0.698, 1.415
Treatment‡	0.804	0.195	0.367	0.500, 1.292	0.995	0.181	0.977	0.696, 1.422
Hysterectomy	0.432	0.155	0.019	0.214, 0.873	0.782	0.174	0.268	0.506, 1.209
Age (in yrs)	1.068	0.029	0.017	1.012, 1.127	1.028	0.018	0.115	0.993, 1.063
BMI	1.011	0.023	0.623	0.967, 1.058	1.029	0.018	0.098	0.995, 1.065
Smoking	1.094	0.277	0.723	0.665, 1.799	1.112	0.214	0.582	0.762, 1.622
B-Pressure	1.201	0.295	0.457	0.741, 1.944	1.591	0.301	0.014	1.099, 2.304
Diabetes	3.759	0.982	¶	2.253, 6.271	2.747	0.561	¶	1.840, 4.100

¶p-value < 0.001;†no covariates; ‡with covariates;$p_{ACM} = 0.669$, $p_{MRCD} = 0.507$ (Equation 2.4)

Figure 4.2: All-cause mortality: overall survival proportions in the 2 treatment groups



Although there was no noticeable difference in survival proportions for myocardial reinfarction or cardiac death between women on HRT treatment compared to those on placebo, Figure 4.2 indicates a higher survival proportion among those women randomized to HRT treatment compared to those on placebo for all-cause mortality outcome.

Overall, there was no evidence to suggest that the model violated the proportional hazards assumptions for both all-cause mortality (χ^2 =!2.51, p-value= 0.113) and myocardial reinfarction or cardiac deaths (χ^2 =1.970, p-value= 0.161).

4.3.2 Per-protocol and as-treated analysis

Table 4.5: Number (percentage) of non-compliers since entry

Treat arm	Time (months)				
	3	6	12	18	24
Active	149 (29%)	122 (41%)	259 (51%)	279 (54%)	294 (57%)
Placebo	118 (23%)	134 (27%)	157 (31%)	171 (34%)	184 (37%)

Table 4.5 shows the rate of noncompliance in the two treatment arms for all-cause mortality. On average there was greater proportion of non-compliers among women randomized to active treatment arm compared to those assigned to the placebo arm (Figure 4.3).

Figure 4.3: Variation of noncompliance proportion with time

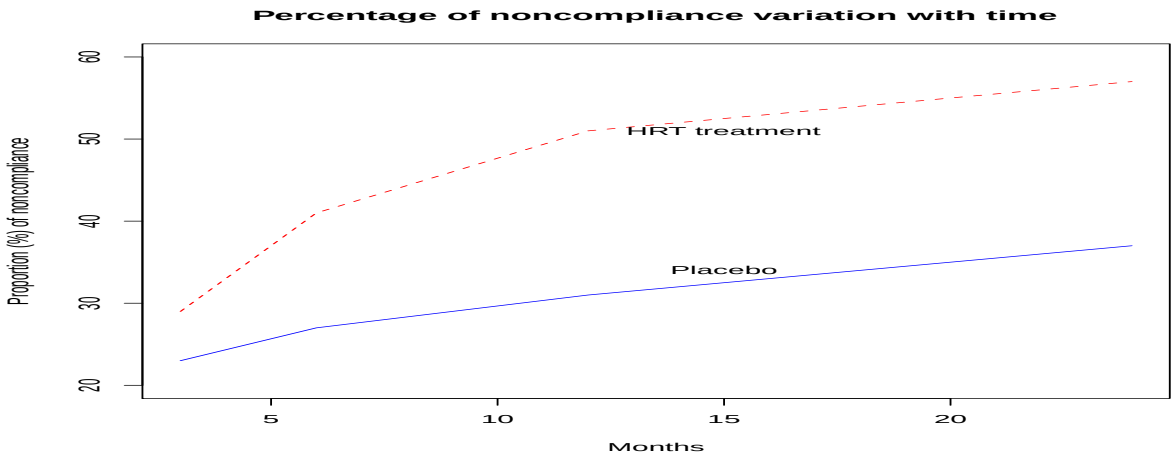


Table 4.6: Per-protocol and as-treated analyses

Analysis	HR	SE	p-value	95% CI
Per-protocol				
Mortality (ACM)	1.128	0.416	0.744	0.548, 2.322
Reinfarction (MRCD)	1.196	0.328	0.514	0.699, 2.045
As-treated				
Mortality (ACM)	0.819	0.251	0.515	0.449, 1.494
Reinfarction (MRCD)	0.891	0.203	0.613	0.571, 1.392

Table 4.6 provide results from per-protocol and as-treated analyses for both all-cause mortality and myocardial reinfarction outcomes. We classified compliers per-protocol as those women who complied with their HRT tablets assignment without interruptions until end of month 23 or until experiencing event, i.e. the per-protocol analysis censored all women who violated protocol (non-compliers in both arms). On the other hand as-treated method analyzed patients according to treatment they actually received regardless of their original treatment assignment, i.e. analysis provided a comparison between women on HRT treatment and those randomized to placebo together with non-compliers (since there were no treatment switches from placebo to active). The rate of compliance per-protocol was about 42% (217/513) and 63% (319/504) in the active treatment and placebo arms respectively.

Per-protocol analysis suggests harmful treatment effects for both outcomes: while HRT treatment increased the risk of death by about 13% ($\text{HR}_{\text{PP(ACM)}} = 1.128,\ 95\%$ CI: $0.548, 2.322$), it increased the risk of myocardial reinfarction by about 20% ($\text{HR}_{\text{AT(MRCD)}} = 1.196,\ 95\%$ CI: $0.699, 2.045$). As-treated analysis on the other hand contradicted per-protocol results by indicating marginal benefits of treatment: now taking HRT treatment reduced the risks of death and myocardial reinfarction by 18% and 11% respectively. Such a contradiction may be a reflection of suboptimal analysis using these two methods probably due to the fact that they are both prone to breaching the canonical principle of randomization.

4.3.3 Simple regression adjustments for noncompliance

We also performed regression adjustments of noncompliance by considering a binary (0/1) (i) fixed and (ii) time-varying all-or-nothing noncompliance indicator N using the models:

$$\text{(i)} \quad \mathrm{h}(t|Z,N) = \mathrm{h}_0(t)\exp[\beta_1 Z + \beta_2 N] \quad \text{and} \quad \text{(ii)} \quad \mathrm{h}(t|Z,N) = \mathrm{h}_0(t)\exp[\beta_1 Z + \beta_2 N(t)],$$

where $\exp(\hat{\beta}_1)$ provides a naive treatment effect estimate and $\exp(\hat{\beta}_2)$ estimates the effect of noncompliance (NCE) and time-varying noncompliance is fitted as a time-varying covariate.

Table 4.7: Simple regression adjustments with binary noncompliance

	All-cause mortality (ACM)					Myocardial reinfarction (MRCD)				
	$e^{\hat{\beta}_1}$	$e^{\hat{\beta}_2}$	SE	P-val	95% CI	$e^{\hat{\beta}_1}$	$e^{\hat{\beta}_2}$	SE	P-val	95% CI
(i)										
Trt†	0.709		0.173	0.160	0.440, 1.145	0.907		0.168	0.599	0.631, 1.304
NCE‡		1.673					1.493			
(ii)										
Trt	0.750		0.182	0.235	0.466, 1.206	0.931		0.170	0.696	0.651, 1.332
NCE		1.032					1.046			

$^{(i)}$Fixed noncompliance;$^{(ii)}$Time-varying noncompliance;†Trt≡treatment; ‡NCE≡effect of noncompliance

Table 4.7 provides results for the simple regression adjustments with noncompliance. Similar to ITT, the results show statistically insignificant treatment effects for both outcomes. However, the regression adjustments also indicated beneficial effects of treatment in reducing the risk of death by 29% and 25% when adjusting for fixed and time-varying noncompliance respectively. The results also showed statistically insignificant marginal treatment effects for MRCD outcome: treatment reduced risk of myocardial reinfarction by 9% and 7% respectively for fixed and time-varying noncompliance. Although the results showed essentially no effects of time-varying noncompliance for both outcomes, time-invariant noncompliance had relatively greater effects in increasing the risk of death (67%) compared to increasing risk of myocardial reinfarction (49%). Given that ordinarily we would expect noncompliance to double the risks, these result may be an indication that the effects of noncompliance are probably captured better under ACM outcome compared to MRCD.

4.3.4 C-Prophet analysis

The C-Prophet method adjusted for all-or-nothing compliance in the treatment arm for both ACM and MRCD where (as defined above) a woman was considered a complier if she took her allocated HRT tablets up to a day before an event (ACM/MRCD) or end of study, whichever came first. On average, there were about 43% women compliers to ACM event and 45% compliers to MRCD events in the active treatment arm: these are the women who actually took HRT tablets up to a day before ACM/MRCD event or end of study.

Table 4.8: C-Prophet results for all-cause mortality and myocardial reinfarction

All-cause mortality (ACM)		Myocardial reinfarction (MRCD)	
HR†	95% CI	HR†	95% CI
0.656	0.334, 1.597	0.988	0.568, 1.931

†effects adjusted for compliance

Table 4.8 provide results for C-Prophet analysis in terms of hazard ratios adjusted for compliance with HRT treatment. The results indicate an overall benefit of HRT treatment in reducing the risk for ACM by 34% ($HR_{ACM} = 0.66$). Also similar to ITT analysis, there was no overall reduction in on risk of MRCD from treatment compared to placebo ($HR_{MRCD} = 0.99$). We also note a relatively wider corresponding 95% confidence interval for the estimate of treatment effect adjusted for compliance for the myocardial reinfarction outcome compared to mortality. Kaplan-Meier (KM) plots (Figures 4.4 and 4.5) indicate that the latent observed survival probabilities would be relatively lower compared to the predicted survival probabilities in the placebo arm throughout the study period for both outcomes.

Figure 4.4: ACM: predicted vs observed KM plot

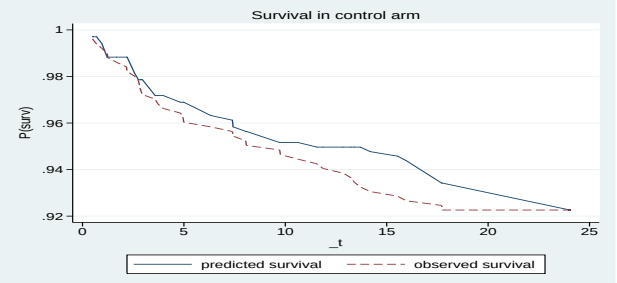
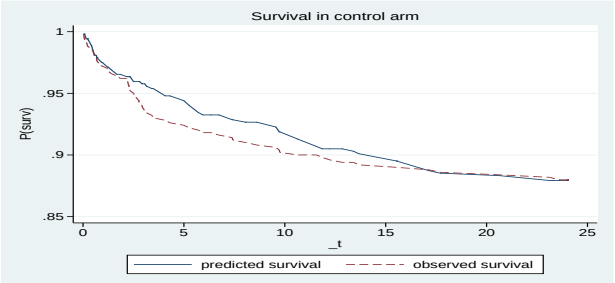


Figure 4.5: MRCD: predicted vs observed KM plot



4.3.5 CHARM analysis

Table 4.9: CHARM results for all-cause mortality and myocardial reinfarction

Analysis	All-cause mortality (ACM) $\hat{\beta}_0$ (95% CI)	$\exp(\hat{\beta}_0)$	SE	p-val.	Myocardial reinfarction (MRCD) $\hat{\beta}_0$ (95% CI)	$\exp(\hat{\beta}_0)$	SE	p-val.
ITT								
Treatment	-0.230 (-0.698, 0.237)	0.795	0.239	0.335	-0.007 (-0.361, 0.346)	0.993	0.180	0.967
Adjusted†								
Treatment	-0.417 (-1.193, 0.359)	0.659	0.396	0.292	-0.012 (-0.602, 0.577)	0.988	0.301	0.951

†for crossovers, $\exp(\hat{\beta}_0) \equiv$adjusted hazard ratio (treatment effect) estimate - Equation 2.15

Table 4.9 provide results for CHARM analysis in terms of hazard ratios. Ignoring crossovers, the ITT results under CHARM are similar to those of obtained above (Section 4.3.1). Specifically, although the results show that the effects of HRT tablets were statistically insignificant, the treatment policy would reduce the risk of death by about 20% ($HR_{ACM} = 0.80$) compared to placebo while there was essentially no difference in risk of myocardial reinfarction ($HR_{MRCD} = 0.99$) between those women randomized to HRT treatment and placebo.

On average there were about 58% (293/513) and 55% (281/513) crossovers from the active treatment arm for ACM and MRCD outcomes respectively. The first part of the analysis using logistic model of noncompliance revealed statistically insignificant log odds of noncompliance: 0.125 (p-value= 0.724) for ACM outcome and -0.392 (p-value= 0.130) for MRCD outcome. The p-values from the test suggests that the odds do not vary with time. It may then be considered meaningful to estimate constant hazard ratios. Adjusting for unidirectional crossovers among non-compliers from active treatment to placebo arm, CHARM estimate indicates a substantial treatment benefit which would reduce the risk of death by 34% compared to placebo. On the other hand adjusting for crossovers indicated no effects on the risk of myocardial reinfarction which remained essentially unchanged. However,

adjusting for noncompliance (crossovers) apparently introduced more variability as evident in increase in the standard error (SE) and relatively wider 95% confidence intervals (CI) for the CHARM estimates adjusted for crossovers compared to SE for the ITT estimates. This may be attributable to the fact that `adjhr` command evaluates the SE and corresponding 95% CI for the CHARM estimate conditional on odds of noncompliance.

Simple CHARM approximation

Table 4.10: Distribution of events in the active treatment arm

Interval (months)	1 (0-3)		2 (3-6)		3 (6-12)		4 (12-23)		Overall (0-23)	
Events (active arm)	C	N	C	N	C	N	C	N	C	N
ACM # (%)	6 (2.7)	2 (0.7)	4 (1.9)	2 (0.7)	2 (1.0)	4 (1.4)	4 (1.9)	8 (2.8)	16 (7.3)	16 (5.5)
MRCD # (%)	19 (8.2)	3 (1.1)	10 (4.7)	3 (1.1)	5 (2.5)	9 (3.3)	3 (1.6)	10 (3.8)	37 (16.0)	25 (8.9)

C≡Compliers;N≡Non-compliers;ACM≡All-cause mortality; MRCD≡Myocardial reinfarction.

Table 4.10 shows distribution (numbers and proportion in each complier/non-complier classification) of both events of interest for the active treatment arm. Overall, there were more MRCDs than ACMs in the active treatment arm, i.e. a total of 62 (37 compliers and 25 non-compliers) compared to 32 ACMs (16 for both compliers and non-compliers) in the treatment arm. Analysis of the logistic model of compliance (2.17) produced the estimate $\delta_{1(\text{ACM})} = 0.080$ (p-value = 0.129, 95% CI: $-0.023, 0.184$) for ACM and $\delta_{1(\text{MRCD})} = 0.180$ (p-value$= 0.001$, 95% CI: $0.075, 0.284$) for MRCD. Although it may not be a sufficient test, as an illustration/exploration we may consider the null hypothesis $H_0 : \delta_1 = 0$ true for ACM but not MRCD given an apparently implied significant time trend/influence for the MRCD outcome.

Overall, we may then estimate the proportion of non-compliers randomized to treatment who experienced ACM event as $\hat{\theta}_{\text{ACM}} = 16/32 = 0.5$. We use Equation (2.16) to approximate the overall simple CHARM estimate for ACM as

$$\text{CHARM}_{\text{ACM}} = \frac{0.795 \times (1 - 16/32)}{1 - (16/32 \times 0.795)} = 0.660,$$

a result which indicates that on average, taking HRT tablets reduces the risk of ACM by 34%. We observe that the simple CHARM estimates approximated by Equation (2.16) agrees with that obtained using the `adjhr` command in Stata.

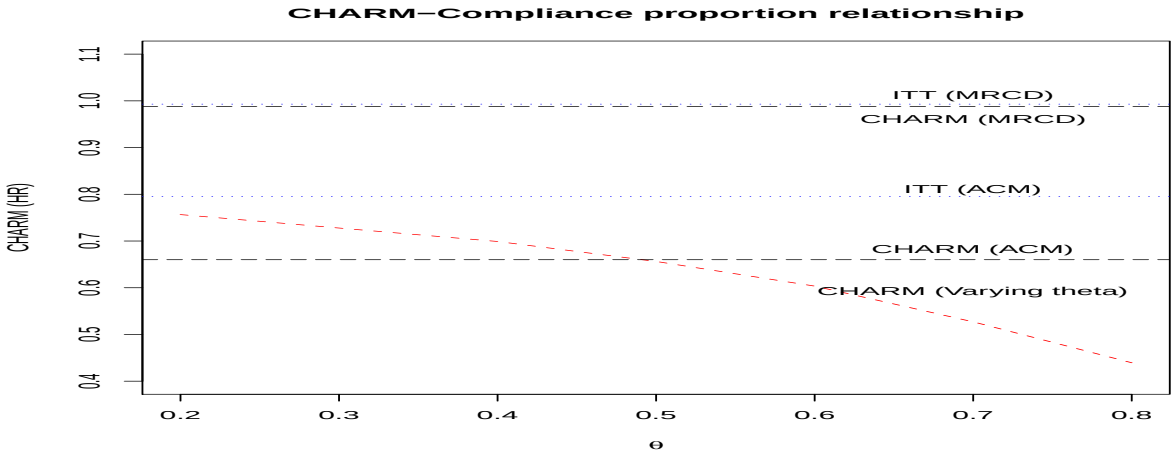
Also although the time trend apparently influences treatment effects for the MRCD outcome, for completeness and comparison we can calculate its CHARM ($\hat{\theta}_{\text{MRCD}} = 25/62$) estimate as

$$\text{CHARM}_{\text{MRCD}} = \frac{0.993 \times (1 - 25/62)}{1 - (25/62 \times 0.993)} = 0.988,$$

which indicates no difference between HRT treatment and placebo for MRCD outcome, a result that agrees with others above.

In general, for a given hazard ratio, the CHARM estimates decreases with increase in $\hat{\theta}$ values for both ACM and MRCD. Figure 4.6 shows the general variation of simple CHARM estimates with compliance proportion (θ) for ACM outcome.

Figure 4.6: Simple CHARM-compliance proportion relationship



4.3.6 CALM analysis

Table 4.11: CALM results for all-cause mortality and myocardial reinfarction

	All-cause mortality (ACM)				Myocardial reinfarction (MRCD)			
Analysis	$\hat{\varphi}$	$\exp(\hat{\varphi})$	$\widehat{\text{HR}}^{\S}$	95% CI ($\hat{\varphi}$)	$\hat{\varphi}$	$\exp(\hat{\varphi})$	$\widehat{\text{HR}}^{\S}$	95%CI($\hat{\varphi}$)
ITT								
best	-0.567	0.567	0.684	-1.541, 0.358	-0.022	0.978	0.989	-1.205, 0.892
Adjusted†								
best	-1.123	0.325	0.472	-2.100, 0.760	-0.009	0.991	0.995	-2.423, 1.548
Adjusted‡								
best	-1.024	0.359	0.504	-2.408, 0.760	-0.020	0.980	0.990	-2.236, 1.548
lower bd*	-1.200	0.301	0.448	-3.000, 0.600	-0.100	0.905	0.951	-2.500, 1.500
upper bd*	-0.600	0.549	0.669	-2.400, 1.200	0.300	1.350	1.164	-2.100, 1.900

†crossovers;‡crossovers+recensoring;*bound;$^{\S}\widehat{\text{HR}} \cong \exp(\hat{\varphi})^{p}$:$\text{p}_{\text{ACM}} = 0.669$,$\text{p}_{\text{MRCD}} = 0.507$ (Weibull, Eqn. 2.4)

For the CALM method, we considered a unidirectional crossover where those women who stopped taking their HRT tablets were classified as non-compliers but no women cross in the opposite direction. The analysis used the duration in months to stoppage of tablet taking and specified the time of the end of study (24 months) as the potential recensoring time. Table 4.11 provide results in terms of the acceleration parameter φ and equivalent hazard ratios (assuming Weibull model) for the CALM method for both outcomes, all-cause mortality (ACM) and myocardial reinfarction (MRCD). For interest, besides ITT we also present results that adjust for crossovers only but not recensoring. CALM estimates are obtained by grid search over a range of values of φ and computing the test statistic $Z(\varphi)$ and $Z(\varphi) - \varphi$ graph may be used to aid in visual assessment of the 'best' grid choice (White et al., 2002).

The ITT estimates under CALM were $\hat{\varphi}_{\text{ITT(ACM)}} = -0.567$ and $\hat{\varphi}_{\text{ITT(MRCD)}} = -0.022$ for ACM and MRCD outcomes respectively. These results imply that for ACM outcome, lifetime was used up only 0.567 times as fast when on treatment compared to when on placebo or equivalently taking HRT treatment increased survival time 1.8-fold ($e^{0.567}$) compared to taking placebo. On the other hand, there was essentially no difference in survival times ($e^{0.022} = 1.02$) between women on HRT tablets and those on placebo for MRCD outcome, results that agree with ITT estimates of the hazard ratio obtained above (Section 4.3.1). The default grid search range $(-1, 1)$ under CALM procedures for both outcomes produced no lower 95% confidence interval limits for the corresponding ITT estimates, i.e possibly indicating too narrow grid. Also graphs of $Z(\varphi)$ against ψ using coarser grid searches (e.g 0.01 steps) showed evidence of non-monotonicity, i.e. a possible indication of non-unique ITT Z-statistic estimate). The above ITT estimates were obtained by exploring grid search in the $(-3, 2)$ range taking 0.6 steps for ACM outcome and $(-2.5, 2)$ range taking 0.4 steps for MRCD outcome. The resulting $Z(\varphi)$ vs φ graphs (Figures 4.7 and 4.8) showed no evidence of non-monotonicity for both outcomes, indicating that the procedure captured the point estimate $Z(0)$ for the ITT Z-statistic. Our next interest is to assess the effect of adjusting for noncompliance due to treatment crossovers (and possibly recensoring).

Figure 4.7: ACM: ITT under CALM

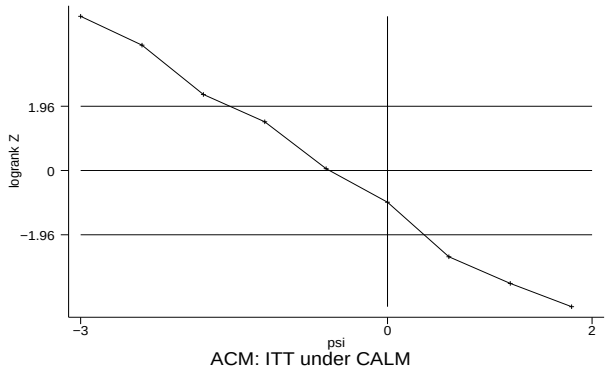
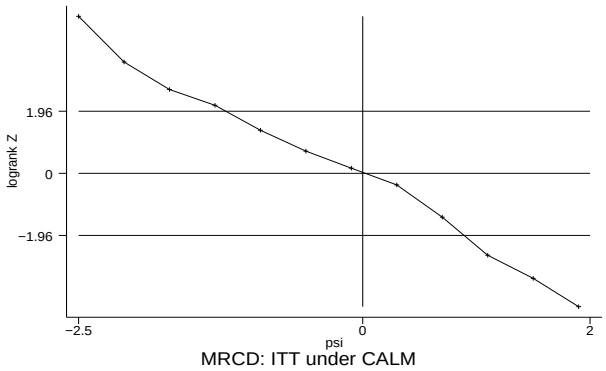


Figure 4.8: MRCD: ITT under CALM



Ordinarily it is plausible to expect adjusting for both crossovers (non-compliers switching from active to placebo) and recensoring to increase the estimated HRT treatment effect compared to placebo. On average there were about 58% (293/513) crossovers for

the ACM outcome and 55% (281/513) for the MRCD outcome. Adjusting for crossovers (CO) using the above coarse grid produced $\hat{\psi}_{\text{ACM(CO)}} = -1.140$ (95% CI: $-1.994, 0.843$) and $\hat{\varphi}_{\text{MRCD(CO)}} = -0.009$ (95% CI: $-2.423, 1.548$), i.e. adjusting for noncompliance due to crossovers from active to placebo arms resulted in a larger absolute value for the HRT efficacy estimate compared to ITT, albeit with relatively wider corresponding 95% confidence intervals.

Adjusting for both crossovers and censoring (recensoring at 24 months), the CALM method's best grid search within the $(-2.5, 2)$ and $(-3, 2)$ ranges, which contain the 95% confidence intervals reported above, and taking 0.6 and 0.4 steps for ACM and MRCD outcomes respectively produced $\hat{\varphi}_{\text{ACM}} = -1.024$, (95% CI: $-2.408, 0.760$) and $\hat{\varphi}_{\text{MRCD}} = -0.020$, (95% CI: $-2.236, 1.548$). The results imply a further improvement in efficacy estimates: compared to placebo, taking HRT treatment increased survival time 2.8-fold ($e^{1.02}$) when the outcome considered was ACM. But similar to previous results, the twin adjustments indicated no difference in survival time ($e^{0.020} = 1.02$) between those taking HRT tablets and placebo for the MRCD outcome. Using the above ranges, the grid search produced the CALM point estimate $\hat{\varphi}$ as lying between -1.2 and -0.6 for ACM and between -0.1 and 0.3 for MRCD. We observe that the results from the twin adjustments retained the upper confidence limits which is an indication of unchanged p-values under CALM procedure (White et al., 2002). Also now the respective graphs of $Z(\varphi)$ against φ for ACM outcome (Figure 4.9) and MRCD outcome (Figure 4.10) indicate no evidence of nonmonotonicity, i.e. no lack of fit.

Figure 4.9: ACM: $Z(\varphi)$ vs φ graph (Esprit)

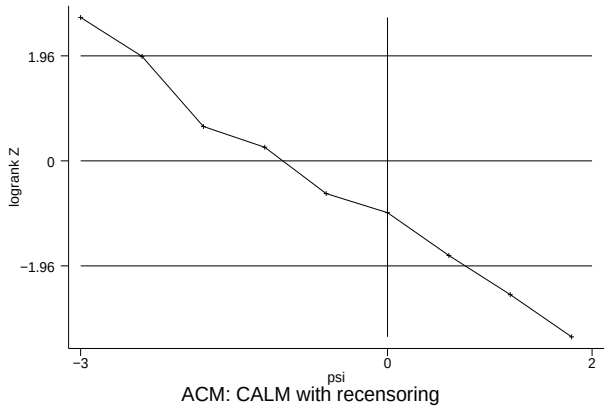
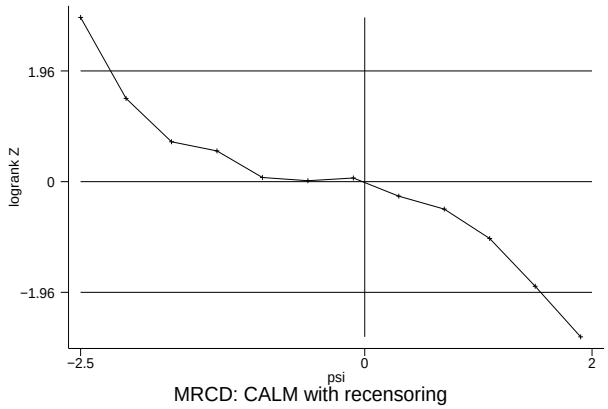


Figure 4.10: MRCD: $Z(\varphi)$ vs φ graph (Esprit)



4.4 Conclusion and discussion

Overall the effects of active treatment were not statistically significantly different from placebo for the Esprit study. There were a total of 71 deaths over the two years of the Esprit study: 32 and 39 from the treatment and control arms respectively. An equal number (16) of women died among both compliers and noncompliers in the treatment arm. Although ITT analysis revealed that active HRT treatment is not statistically different from placebo, the treatment suggests positive benefits in reducing the risk of death by about 20%. The study duration of two years may have been too short to reveal HRT effects on myocardial reinfarction given that such significant effects may only be demonstrated after considerably long period of medication, e.g. 5 years (Chilvers et al., 2003). ITT analysis ignored any compliance information, the results may be considered conservative due to dilution effects by noncompliers.

On average a higher proportion of women complied with their placebo treatment (63%) compared to those who complied with their HRT treatment (42%). Compliance rates for both arms decreased as the study progressed (Figure 4.3). While per-protocol analysis suggested HRT treatment would reduce risk for both mortality and reinfarction outcomes, as-treated analysis showed marginal risk reduction due to HRT treatment. This contradictory results may be attributable to the loss of randomization ideals by both per-protocol and as-treated analysis: the per-protocol analysis censored all women non-compliers in both arms. Such an analysis may fail to account for systematic difference between compliers and noncompliers. For example, the noncompliers would have stopped medication as a result of undesirable (adverse) side effects like bleeding hence the analysis is likely to produce biased results because the comparison is not between like with like. On the other hand, as-treated analysis simply compared those women who received treatment to those who did not where the later group comprised both women randomized to placebo and non-compliers. In effect, such an analysis considered the non-compliers in the treatment arm as belonging to the placebo (assumes they essentially received no treatment). As a result the analysis has the potential to introduce

selection bias due to a violation of the randomization principle. In general, both per-protocol and as-treated methods classify participants according to treatment received and analysis based on such direct measures of treatment efficacy are prone to bias (e.g. selection bias). The two methods are widely and justifiably maligned as flawed (Lee et al., 1991; Lui, 2011). In fact according to Joffe and Brensinger (2003) such invalid estimates of effect may essentially explain nothing in terms of treatment efficacy.

Results from C-Prophet analysis suggested that on average complying with HRT medication would reduce the risk of death by about 35% compared to placebo. Provided the exclusion restriction holds the C-Prophet allows for causal estimation in the subpopulation of compliers. However, a key challenge for inference on C-Prophet stems from the fact that the all-or-nothing treatment indicator is unobserved in the placebo arm. Furthermore, C-Prophet estimation is premised on the exclusion restriction assumption which is often restrictive and may be violated in most trials. Also just like the other specialists methods considered for this work, the application of C-Prophet estimation is limited to covariate-free models. The exclusion restriction assumption may be considered as reasonably satisfied in the Esprit data because it was a double-blind trial. However, participants were likely to know their treatment hence likelihood of breaching the assumption. The no-defiers assumption may be considered satisfactory for the Esprit data because the participants only had access to the allocation they were randomized to, i.e. no access to alternative treatment (or no switching).

In general, the CHARM estimates for ACM were similar to the estimates obtained from C-Prophet analysis: compliance with HRT treatment compared to placebo reduced risk of death by 34%. For a time-invariant hazard ratio and odds of noncompliance, simple weighting of the ITT hazard ratio estimate using the proportion of noncompliance who experience event of interest in the treatment arm provided an accurate CHARM estimate. Expectedly, the simple CHARM approximations for both ACM and MRCD indicated sensitivity to proportion of compliance (Figure 4.6): higher proportion of compliers indicated more reduction in risk of experiencing both outcomes. The extended exclusion restriction assumption (see Equation 2.11) is key to CHARM estimation and may be considered plausible for the

Esprit data since we assume that a participant's past treatment had no effect on her present risk to death. However, in practice the assumption is likely to be violated if the treatment were to induce adverse side effects (e.g. bleeding) that can potentially affect compliance. Such a breach may be addressed by allowing lagging of effects into successive intervals (White et al., 2004), i.e. allowing treatment effect to last past the date of treatment stoppage. This would however require use of more advanced statistical techniques such as time series in addition to making other assumptions. The low rate of compliance (43%) in the Esprit study may generally limit generalization the findings, i.e. questionable external validity. Also despite its appeal, the all-or-nothing compliance analysis may fail to account for the more prevalent partial form of compliance. For example, noncompliance with chronic medication in practice is rarely an all-or-nothing phenomenon (Sheiner and Rubin, 1995). We may extend the model (possible with additional plausible assumptions) to account for partial compliance which is more of the norm than the exception in real life. This may allow us to relax the assumption so that participants are only assumed to stop treatment at the beginning of the interval.

Assuming a Weibull model for the Esprit data, on average the ITT under the CALM procedures performed as well as both the CHARM and C-Prophet methods (on the hazard ratio scale). By adjusting for both crossovers and censoring the CALM results indicated that compared to placebo, HRT tablets reduced risk for all-cause mortality by half (50%). Also twin adjustments of both crossovers and censoring indicated no change in the p-values by retaining the upper confidence limits under CALM procedure.

The proportional hazard assumption is key to validity of the Cox's model which also treats covariates multiplicatively. For the Esprit data there was no evidence to suggest a violation of the proportional hazards assumptions for both outcomes: all-cause mortality ($\chi^2 = 2.51$,p-value$= 0.113$) and myocardial reinfarction or cardiac deaths ($\chi^2 = 1.970$,p-value$= 0.161$). But other flexible techniques could also be used, for example the Aalen (1989) additive model may be used to explore the additive effects of covariates. In addition we could simultaneously consider death from both reinfarction and cardiac attack. With multiple failures models, we could use competing risk analysis to explore the individual causes of death simultaneously. This

would enable us examine their respective effects on the survival rates among compliers. All the covariates considered were assumed to be time invariant but we could extend the Cox PH model to include time-dependent covariates that would account for amount of tablet dosage.

In general, all the the three specialist methods (C-Prophet, CALM and CHARM) were applicable to the Esprit study. Overall, analysis of the Esprit data suggested that results on possible effects of HRT treatment vary between methods. Performance of both C-Prophet and CHARM methods were similar: compliance with HRT treatment would produce a reduction of about 35% in risk of death compared to placebo. Although limited compliance information would make all-or-nothing noncompliance classification plausible for C-Prophet analysis, the CHARM method provided more flexibility in adjusting for possible crossovers. Analysis using the CALM method suggested that compliance with HRT treatment compared to placebo would reduce the risk of death by half (50%). As a result we recommend the CALM method which allows twin adjustments of crossovers (partial noncompliance) and recensoring.

In Chapter 7 we will apply statistically designed simulation studies to evaluate the performance of all the specialist methods in terms of bias, root mean squared error and 95% confidence intervals coverage.

Chapter 5

Esprit Analysis II: Predicting Arm-specific Compliance

The objective in this chapter is to apply methods developed in Chapter 3 to obtain plausible predictors of compliance to treatment allocation up to 23 months in each treatment arm of the Esprit study from the set of recorded baseline covariates to construct marginal (arm-specific) compliance prediction models. We then validate the selected prediction models using performance calibration and discrimination indices. For both the active treatment and placebo arms, we consider a binary all-or-nothing compliance variable, i.e. women compliers who took their HRT tablets and compliers who stuck with their placebo allocation without interruptions. All the analysis in this Chapter (model selection and validation of selected prediction models) are performed using the R Software (R Development Core Team, 2008) - annotated codes are presented in Appendix on page 263.

5.1 Selecting predictors

Before performing any formal modelling, we choose the following 9 potential predictors of compliance to treatment allocation on the basis of plausibility:

1. History of hysterectomy: the risk of womb cancer from HRT may influence noncompliance among those with wombs. On the other hand women without wombs may have no such fears making them more compliant with their treatment. Also since unopposed HRT may induce endometrial bleeding, it may become less acceptable to those with wombs.

2. Smoking: smokers may have different attitudes to risk than non-smokers which in turn may influence compliance to treatment, e.g. smokers may be more compliant in taking the tablets with no fear of any risks. We use binary smoking status: never smokers and others.

3. Social class: professionals and skilled women are likely to be more educated which may influence their attitudes towards HRT. We use three ordinal categories: professional, skilled and others.

4. Age: attitudes towards HRT and possible risks may be different between older and younger women. While older people may be used to some of the postmenopausal symptoms, younger women may be more compliant with treatment hoping to relieve some of the symptoms. Also older people might be less willing to tolerate side effects like bleeding.

5. Cerebrovascular risks (CVD) was defined as a single predictor combining any history of angina, blood pressure or stroke at baseline. A patient with a history of any of these risk factors for future infarctions may feel more vulnerable than other patients and this may affect their attitude to tablet taking.

6. Diabetes: diabetics often manage their condition/lives with intensive intervention on lifestyle and risk factors compared to non-diabetics. While vulnerability for those with history of the above risk factors may encourage them to take HRT tablets hoping to benefit from any protective effects, they might not wish to increase their burden of medications. Hence diabetic women may have a different attitude to risk.

7. Fractures. Given the known effect of HRT to strengthen bone density, knowledge of a history of fracture may make a patient more inclined to take the tablet.

8. Alcohol: knowledge of risk of breast cancer among those taking alcohol and HRT may result in different compliance attitude to tablet between those who drink and those who do not.

9. HRT history: unpleasant or pleasant experience/s from previous HRT use may influence decision to take the tablets again.

We used enhanced ordinary bootstrap[1] resampling method (Davison and Hinkley, 1997; Efron, 1979, 1983, 1986) by taking 200 bootstrap samples to estimate the bias due to overfitting and optimism for the models predicting compliance. We then use the resulting bias/overfitting-corrected estimates of predictive accuracy to investigate the validity of fitted models and evaluate predictive performance in terms of percentage of optimism, calibration slope and concordance c-statistics using Somers (1962) D_{xy} statistics (see Equation 3.10, Chapter 3) for each individual arm in five models:

(i) Original model with all the 9 predictors without any selection:

$$\begin{aligned}\text{logit}(\mu) = \; & \beta_0 + \beta_1 \text{Hysterectomy} + \beta_2 \text{Smoking status} + \beta_3 \text{Social-class} + \\ & \beta_4 \text{Age} + \beta_5 \text{CVD Risks} + \beta_6 \text{Diabetes} + \beta_7 \text{Fracture} + \beta_8 \text{Alcohol} + \beta_9 \text{HRT},\end{aligned}$$

where μ is the probability of compliance with treatment (placebo/HRT) allocation and histories of hysterectomy, risks, diabetes, fracture, alcohol together with smoking status are taken as binary 0/1 predictors.

(ii) Reduced model obtained from (i) above by stepwise backward elimination procedures using AIC stopping rule and 0.10 significance level for a variable to be retained in a model.

(iii) Model fitted with the retained predictors in reduced model (ii) above but the predictors assumed pre-specified (following suggestion by Harrell et al. 1996).

[1]Implemented in `Design` package in R

(iv) Intermediate model composed of 6 variables constructed using penalized maximum likelihood estimation with modified AIC ($\chi^2 > 2\text{df}$) to find the penalty factor as defined in the previous chapter (Equation 3.7).

(v) Lasso[2] (Least Absolute Shrinkage and Selection Operator) model selection from the original (i) above.

We also considered penalized maximum likelihood estimation regression versions of the original and intermediate models above following discussion in Chapter 3 (Section 3.4). We calculated a uniform shrinkage s using Equation (3.8). The penalized coefficients are obtained from a wide grid search over the $0 - 50$ range and we report optimal penalty factors λ_{opt} for each model. Finally we compare the predictive performance of the above 5 models in terms of calibration, discrimination and level of optimism for the the selected prediction model.

5.2 Results

A higher rate of noncompliance to treatment was recorded among patients randomized to the active treatment arm (63%) compared to those randomized to placebo (42%). The association between treatment and compliance to allocation was highly significant (OR$= 0.43, 95\%$ CI $0.33, 0.55$). Smoking status, history of diabetes and of HRT use had opposing effects in predicting compliance in the individual arms for both the original and penalized models. Women with histories of hysterectomy and cerebrovascular risks risks were more likely to comply with either treatment allocation but association was stronger in the active arm. Women with history of taking alcohol were more likely to comply if allocated placebo.

Results from Table (5.1) show the simple expected (uniform) shrinkage estimate for the original model predicting compliance to HRT and placebo using 9 predictors as 0.76 and 0.47 respectively. This implies the models for predicting compliance to HRT would be less overfitted (24%) compared compared to those predicting compliance to placebo (53%). Penalizing

[2]Implemented using `glmpath` package in R

Table 5.1: Original (9 predictors) model results: log odds ratio of predicting compliance (SE)

Predictors	Coefficients (log odds ratio)					
	Original		Penalized		Lasso	
	Plcbo	Active	Plcbo	Active	Plcbo	Active
Hysterectomy	0.368	1.135	0.163	0.873	0.112	0.903
	(0.244)	(0.210)	(0.159)	(0.182)	(0.240)	(0.207)
Smoking	-0.608	0.303	-0.318	0.244	-0.360	0.158
	(0.203)	(0.196)	(0.144)	(0.172)	(0.193)	(0.193)
Sclass_1 (Skilled)	0.144	0.328	0.064	0.205	0.000	0.000
	(0.219)	(0.237)	(0.149)	(0.194)		
Sclass_2 (Professional)	-0.353	-0.469	-0.158	-0.290	0.000	0.000
	(0.391)	(0.385)	(0.259)	(0.316)		
Age‡	-0.011	-0.025	-0.005	-0.022	0.000	-0.007
	(0.020)	(0.019)	(0.016)	(0.017)		(0.019)
CVD Risks	-0.191	-0.380	-0.075	-0.293	0.000	-0.194
	(0.195)	(0.193)	(0.142)	(0.170)		(0.188)
Diabetes	-0.096	0.223	-0.046	0.139	0.000	0.000
	(0.272)	(0.267)	(0.167)	(0.217)		
Fracture	0.109	0.011	0.054	0.008	0.000	0.000
	(0.244)	(0.274)	(0.160)	(0.222)		
Alcohol	0.467	0.126	0.240	0.088	0.265	0.000
	(0.207)	(0.194)	(0.146)	(0.171)	(0.202)	
HRT	0.220	-0.336	0.092	-0.195	0.000	0.000
	(0.337)	(0.299)	(0.179)	(0.234)		
Shrinkage (s^*)	0.471	0.760	0.635	0.804		
λ_{opt}			43.00	13.00		
df_λ			5.220	7.740		

‡ log OR per year

Sclass≡social class

Plcbo≡placebo

*s calculated using Equation (3.8) for $r(\text{df})=9$

the coefficients resulted in substantial reduction/shrinkage of the log odds ratio estimates (Table 5.1, columns $4-5$). Penalization improved the original models by reducing the level of overfitting when predicting compliance to both HRT and placebo: models predicting HRT compliance are now overfitted by 20% and those predicting placebo by 36%. Although penalization produced fewer effective degrees of freedom while predicting compliance to placebo ($df_\lambda = 5.22$) compared to predicting compliance to HRT ($df_\lambda = 7.74$), the tradeoff was a larger optimal penalty factor: $\lambda_{opt} = 43$ and 13 respectively. These results indicate better prediction for compliance to HRT compared to compliance to placebo, i.e. the predictive model for placebo was far more overoptimistic compared to HRT prediction.

Table 5.2: Intermediate (6 predictors) model results

	Coefficients (log odds)					
Predictors	Original		Penalized		Lasso	
	Plcbo	Active	Plcbo	Active	Plcbo	Active
Hysterectomy	0.384	1.099	0.237	0.904	0.117	0.904
Smoking	-0.607	0.293	-0.424	0.249	-0.367	0.168
Age	-0.011	-0.019	-0.007	-0.019	0.000	-0.009
CVD Risks	-0.190	-0.343	-0.112	-0.288	0.000	-0.203
Fracture	0.104	0.007	0.072	0.007	0.000	0.000
Alcohol	0.490	0.095	0.334	0.077	0.269	0.000
Shrinkage (s^*)	0.655	0.840	0.724	0.862		
λ_{opt}			20.00	10.00		
df_λ			4.300	5.030		

*s calculated using Equation (3.8) for $r(df)=6$

Analysis of the intermediate model comprising 6 predictors (Table 5.2) using all the three criteria produced improved prediction of compliance to both HRT and placebo. In predicting compliance to HRT and placebo the original models would be overfitted by 16% and 34% respectively. There was relatively less penalization of the coefficients from the intermediate model compared to the original model - small improvement (20% to 16%) in overfitting for

predicting compliance to HRT. However, there is significant improvement in predicting compliance to placebo with the optimal penalty now reduced to only twice (20) that for HRT compliance (10) while it was more than three times (43 : 13) larger for the full model (Table 5.1). Furthermore, the effective degrees of freedom are now almost similar (5.3 and 4) for predictions of compliance to HRT and placebo.

Table 5.3: Selected predictors (log-odds ratios) of compliance for 3 criteria

Stepwise	Intermediate		Lasso
<u>Active arm</u>			
Hysterec† (1.088)	Hysterec (0.904)	Smoking (0.249)	Hysterec (0.903)
Smoking (0.331)	Age (−0.019)	CVD Risks (−0.288)	Smoking (0.158)
CVD Risks (−0.355)	Fracture (0.007)	Alcohol (0.077)	CVD Risks (−0.194)
			Age (−0.007)
<u>Placebo arm</u>			
Hysterec (0.349)	Hysterec (0.237)	Smoking (−0.424)	Hysterec (0.112)
Smoking (−0.552)	Age (−0.007)	CVD Risks (−0.112)	Smoking (−0.360)
Alcohol (0.494)	Fracture (0.072)	Alcohol (0.334)	Alcohol (0.265)

†Hysterectomy

The Lasso method in general selected similar predictors of compliance to both HRT and placebo as those from the stepwise backward elimination but with significantly reduced/shrunk estimates (Table 5.3). Specifically, both Lasso and the stepwise selection procedures selected the same predictors of compliance to placebo. But Lasso method also selected age as a marginal additional predictor of compliance in the active treatment arm. The coefficients (log odds) produced by the Lasso model were more severely penalized compared to those from the intermediate model.

5.3 Validation performance of selected models

We used enhanced bootstrap on all aspects of models development (selection and estimation procedures) to revalidate on samples taken with replacement from the whole sample and apply on the 5 models above: the original composed of 9 predictors, the reduced model of 3 predictors, same reduced model with 3 predictors assumed pre-specified, the intermediate of 6 predictors and the Lasso model. The reduced model was obtained from the original model using stepwise backward elimination procedures using AIC stopping rule and 0.10 significance level for retaining a predictor in a model. The variables selected for the reduced model were consistently (90%) selected across bootstraps resamples. These were the same predictors deemed important by the backward elimination algorithm.

The original model consisting of 9 predictors produced better predictions of compliance to HRT than placebo (Table 5.4). While predicting compliance to HRT and placebo, the original models would be overfitted by 18% and 33% respectively. In addition, these models would be optimistic by 6% and 9% respectively in predicting compliance to HRT and placebo. The reduced model would also perform relatively better in predicting compliance to HRT compared to predicting placebo - specifically the reduced model predicting compliance to HRT would perform better at distinguishing compliers from non-compliers (concordance $c = 0.620$) the reduced model predicting compliance to placebo ($c = 0.566$). Reduced models predicting compliance to placebo would be more optimistic (8%) than those predicting compliance to HRT (6%). Predictions for compliance to HRT using the reduced model would be equally well calibrated (slope $= 0.83$) compared to predictions from the original model (slope $= 0.82$).

As expected, the model with 3 predictors assumed pre-specified performed best in terms of both calibration and discrimination among the 5 models considered in predicting compliance to both HRT and placebo. These models also produced least optimistic fits for predicting compliance to both arms. Specifically predictions of compliance to both HRT and placebo using the 3 predictors assumed pre-specified were almost perfectly calibrated (0.96 and 0.95) and least optimistic (1% and 2%).

Table 5.4: Validation performance: calibration,concordance and optimism

Model		Calibration slope	Optimism (%)	Concordance $c^{\ddagger}$-statistics
(i) Original				
Active		0.818	6.1	0.639
Placebo		0.671	8.6	0.573
(ii) Reduced				
Active	(hyst+smk+CVD)	0.827	5.8	0.620
Placebo	(hyst+smk+alc)	0.667	8.1	0.566
(iii) Reduced†				
Active	(hyst+smk+CVD)	0.961	1.4	0.642
Placebo	(hyst+smk+alc)	0.950	2.0	0.597
(iv) Intermediate (6 predictors)				
Active		0.879	4.1	0.636
Placebo		0.766	6.0	0.580
(v) Lasso				
Active	(hyst+smk+age+CVD)	0.935	2.3	0.647
Placebo	(hyst+smk+alc)	0.925	2.3	0.595

hyst≡hysterectomy

smk≡ smoking status

alc≡alcohol

CVD≡cerebrovascular disease

†model assumed pre-specified

‡c calculated from D_{xy} (see Equation 3.10)

Validation of the model with 6 predictors produced good performance with intermediate measures between the original models of 9 predictors and the Lasso models (see below). We observe that predictions of compliance to both HRT and placebo using the intermediate model was equally good as predictions of compliance to HRT using the the reduced model. For example predictions of compliance to placebo using the intermediate model was equally optimistic (6%) as predictions of compliance to HRT using the reduced model. Overall, models from the intermediate model provided significantly improved predictions of compliance to both HRT and placebo in terms of calibration and optimism without affecting the capability to distinguish (discriminate) between compliers and non-compliers.

Besides the reduced model fitted with predictors assumed pre-specified, Lasso models produced the best calibrated and discriminative models predictive of compliance to both HRT and placebo (Table 5.4, lower panel). Predictions of compliance to both HRT and placebo using the Lasso models were the least optimistic (2%) and almost perfectly calibrated (slope= 0.93). Although the performance of Lasso models produced substantially shrunk and improved estimates compared to the stepwise backward elimination method, the severe penalization (considerably many coefficients shrunk exactly to zero) have the potential to exclude potential predictors (waste of data).

In the next chapter we link the two marginal models for each treatment arm into a casual model that provides principal effects for each stratum. Specifically, using the intermediate model we choose common penalized predictors of compliance for both HRT treatment and placebo (smoking status, history of hysterectomy, CVD risks, alcohol, fracture and age at admission) and join the resulting two marginal models using an association model from which we obtain relative risks for each stratum.

Chapter 6

Esprit Analysis III: Modelling Effects of Noncompliance in Two Arms

6.1 Introduction

In this chapter we use the selected arm-specific predictors of compliance obtained in the previous Chapter 5 and apply Roy et al. (2008) method of principal stratification framework as outlined in Section 2.7 to develop causal models linking the two marginal models in order to obtain principal effects for each stratum. Applied to the Esprit data, one marginal model estimates arm-specific probability to comply with HRT treatment in the presence of six baseline (penalized) covariates selected in Chapter 5 above which are predictive of compliance while the other model estimates probability of compliance to placebo allocation using the same set of predictors of compliance. We considered all-or-nothing compliance to treatment allocation up to 23 months in the Esprit study. We used the intermediate model comprising 6 potential predictors of compliance considered for each arm including five binary (0/1) variables: histories of hysterectomy, smoking, cerebrovascular risks (CVD-angina or blood pressure or stroke), fractures and alcohol. We also included age of patient at admission as a continuous predictor variable. The Stata and WinBUGS codes are shown in Appendix on page 265.

6.2 Linking two marginal compliance models

In the Esprit data there is compliance data for both the active HRT treatment and placebo. Here placebo and HRT tablets are considered as treatment 0 and treatment 1 respectively. Owing to poor compliance rates with HRT assignment for the Esprit data, we assume the intermediate model comprising smoking status, history of hysterectomy, CVD risks, alcohol, fracture and age at admission as common predictors of compliance for both HRT treatment and placebo. We use logistic models to predict compliance with both treatments separately:

$$\text{logit}\left[\mu_j(\mathrm{x})\right] = \left(\sum_{i=0}^{6} \gamma_{ji}\mathrm{x}_i\right),\ j = 0, 1$$

where $\mu_j(\mathrm{x})$ is the probability of compliance with treatment allocation j given the selected predictors $\mathrm{x}_1, \ldots, \mathrm{x}_6$. We estimated probabilities of complying with HRT treatment/placebo allocations using:

$$\hat{\mu}_j(\mathrm{x}) = \left[1 + \exp\left(-\sum_{i=0}^{6} \hat{\gamma}_{ji}\mathrm{x}_i\right)\right]^{-1},\ j = 0, 1,$$

where γ estimates the log odds ratio of compliance estimated from the prediction model as outlined before in Chapter 5.

6.2.1 Fitting the model

We estimated the causal risk ratio parameters τ_{ij} (Equation 2.25) in a Bayesian setting using WinBUGS[1] (Spiegelhalter et al., 2004) within Stata (Thompson et al., 2006) software. We used non-informative priors $N(0, 10^6)$, i.e. normal distributions with mean zero and large variance for all log odds ratio parameters γ_j for potential predictors of compliance. We specified uniform $(0, 1)$ priors for the $\pi_Z^{\mathcal{S}=\mathrm{s}}$ $(\pi_i, i=1, \ldots, 7)$ parameters, $z=0, 1$, $\mathcal{S}$=(0,0),(1,0),(0,1),(1,1) and set the sensitivity parameter $\phi=0, 0.2, 0.5$ and 0.8. The choice of ϕ were motivated by the need to explore all possible compliance behaviours including conditional independence

[1] `BUGS Project. BUGS: Bayesian Inference Using Gibbs Sampling (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/)`

($\phi=0$) and almost-perfect correlation ($\phi=0.8$). We ran three chains: null starting values for chain one, mean and median values from a trial run for chains two and three respectively. For convergence assessment, we ran simulation for $101,000$ iterations for each of the three chains and excluded the first $1,000$ as burn-in. Posterior median relative risks (minimizing linear loss function) provided Bayesian point estimates for each stratum.

First we use the prediction models to estimate log odds ratio of compliance to both HRT (active) treatment and placebo for both ACM and MRCD outcomes separately. Next we estimate the mean proportion of compliance $\bar{\hat{\mu}}_0(\mathrm{x})$ and $\bar{\hat{\mu}}_1(\mathrm{x})$ for all subjects. We then estimate stratum-specific joint compliance proportions under conditional independence ($\phi=0$, see Table 6.1) and for three other different values of sensitivity ($\phi = 0.2, 0.5, 0.8$), i.e. to explore possible effects of ϕ. Finally we estimate causal effects in terms of posterior mean relative risks of death for each stratum and compare them for the 4 values of ϕ listed above.

6.3 Results

Results from Table (6.1) shows that the posterior median log odds ratios of compliance to treatment were generally similar for both ACM and MRCD outcomes. Under conditional independence ($\phi=0$), patients with histories of hysterectomy were more likely to comply with their treatment allocation if randomized to either placebo or HRT tablet. While smokers were more likely to comply with treatment under HRT allocation, they were less likely to comply under placebo. On the other hand women with history of alcohol use were more likely to comply under placebo allocation only. On average, the results are comparable to those from the previous chapter (Table 5.2).

The plot (Figure 6.1) of estimated compliance probabilities $\hat{\mu}_{ij}$ at different values of sensitivity parameter ϕ depicts how the joint compliance behaviour pattern/trend depends on choice of the ϕ. We observe that the proportion of women who would comply with one treatment allocation but not the other ($S=(0,1)$) and ($S=(1,0)$), decreased with increase in the value of ϕ.

Table 6.1: Median estimates (mean 95% CI) of the posterior distribution of compliance model parameters for ACM and MRCD outcomes when $\phi=0$ (conditional independence)

Parameter	Placebo arm posterior median log odds ratio ($\hat{\gamma}_0$)		Active treatment arm posterior median log odds ratio ($\hat{\gamma}_1$)	
	ACM	MRCD	ACM	MRCD
Intercept	0.837 (0.385, 1.309)	0.737 (0.350, 1.127)	-0.602 (-1.068,-0.113)	-0.656 (-1.076, -0.258)
Hysterectomy	0.472 (0.001, 0.929)	0.507 (0.059, 0.970)	0.996 (0.607, 1.401)	1.079 (0.682, 1.486)
Smoking	-0.568 (-0.983,-0.160)	-0.545 (-0.923, -0.172)	0.227 (-0.188, 0.632)	0.315 (-0.054, 0.063)
Age‡	0.001 (-0.036, 0.038)	-0.008 (-0.046, 0.008)	−0.027 (-0.061,0.007)	-0.012 (-0.048, 0.022)
CVD Risks	-0.188 (-0.560,0.171)	-0.275 (-0.629, 0.087)	-0.282 (-0.650, 0.071)	-0.271 (-0.0631, 0.078)
Fracture	0.043 (-0.408,0.487)	0.088 (-0.357, 0.547)	0.073 (-0.455,0.580)	-0.050 (-0.593, 0.470)
Alcohol	0.460 (0.071, 0.854)	0.481 (0.120, 0.863)	-0.052 (-0.417,0.312)	0.135 (-0.237, 0.495)

‡ log OR per year

Figure 6.1: Compliance behaviour pattern for each stratum.

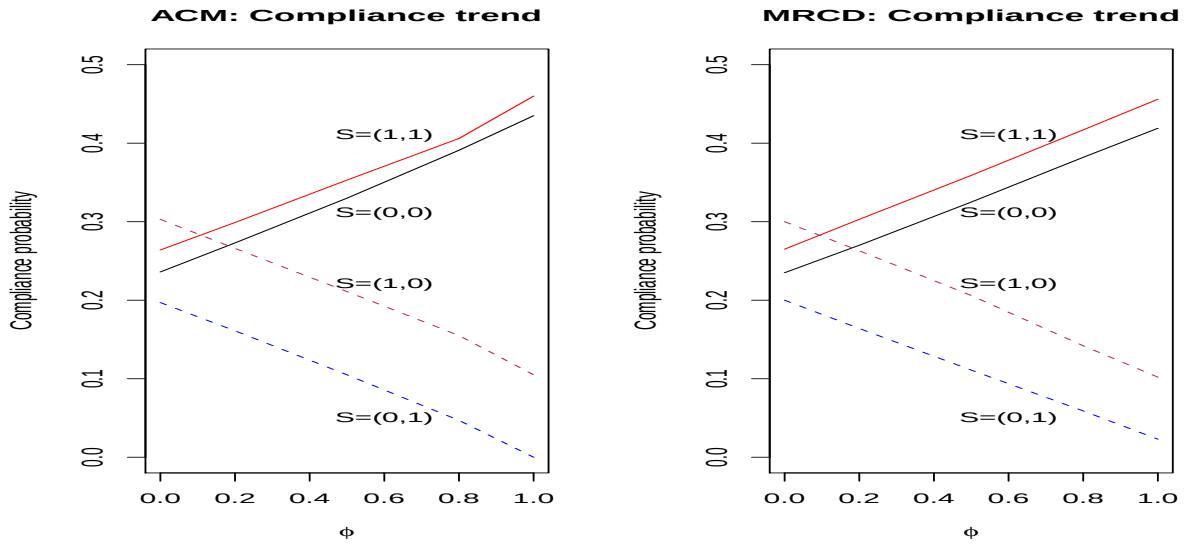


Figure (6.1) shows that the proportion of patients who would have the same compliance status under either treatment allocation increases as ϕ increased, i.e. the estimated proportion of women in stratum 3 ($S=(1,1)$) and in stratum 0 ($S=(0,0)$). The proportion of subjects who would actually comply with either treatment allocation ($S=(1,1)$) would be generally higher than those who would not comply with either allocation ($S=(0,0)$). The proportion of patients who would comply with HRT tablets only ($S=(0,1)$) would be relatively lower compared to the proportion that would comply with placebo allocation only ($S=(1,0)$).

On average, the estimated median probabilities of compliance was higher among those patients allocated to placebo for both ACM (MRCD) outcomes ($\bar{\hat{\mu}}_0(\mathrm{x})=0.567$ (0.565)) compared to those on HRT tablets ($\bar{\hat{\mu}}_1(\mathrm{x})=0.461$ (0.470)), i.e. the ratio $U(\mathrm{x})=\min\{1,\frac{\bar{\hat{\mu}}_1(\mathrm{x})}{\bar{\hat{\mu}}_0(\mathrm{x})}\}=0.795$ (0.810). We note the likelihood that a higher prevalence (proportion) of placebo compliance compared to HRT may be a limitation of the model to effectively evaluate active HRT efficacy.

Table 6.2: Median compliance proportion per principal stratum for different values of ϕ

Type (stratum)	Outcome: All-cause mortality					Outcome: Reinfarction				
	ϕ					ϕ				
	0	0.2	0.5	0.8	1	0	0.2	0.5	0.8	1
3 $(1,1)$	0.264	0.296	0.353	0.406	0.460	0.266	0.303	0.359	0.417	0.456
2 $(0,1)$	0.197	0.165	0.105	0.047	0	0.200	0.164	0.111	0.059	0.023
1 $(1,0)$	0.303	0.262	0.211	0.155	0.105	0.300	0.263	0.206	0.142	0.102
0 $(0,0)$	0.236	0.277	0.330	0.391	0.435	0.235	0.270	0.325	0.382	0.419

For the 4 strata at moderate value of sensitivity parameter $\phi=0.5$ (Table 6.2), the group with the highest prevalence was patients who would comply with either treatment ($\bar{\hat{\mu}}_{11}=0.353$ (0.359)) and the group with the lowest prevalence is those who would only comply with HRT tablets ($\bar{\hat{\mu}}_{01}=0.105$ (0.111)). The median probabilities of compliance among those patients who would only comply with placebo and those who would not comply with either treatment allocation were $\bar{\hat{\mu}}_{10}=0.211$ (0.206) and $\bar{\hat{\mu}}_{00}=0.330$ (0.325) respectively.

Table 6.3: Causal risk ratio estimates (means of median posterior relative risks) for ACM and MRCD (mean 95% CI) for each stratum for different values of ϕ: (a) All-cause mortality (b) Myocardial reinfarction.

	Compliance with both HRT and placebo		Compliance with HRT only		Compliance with placebo only	
ϕ	$\hat{\pi}_2(\pi_1^{s=(1,1)})$	$\hat{\pi}_6(\pi_0^{s=(1,1)})$	$\hat{\pi}_1(\pi_1^{s=(0,1)})$	$\hat{\pi}_7(\pi_0^{s=(0,1)})$	$\hat{\pi}_4(\pi_1^{s=(1,0)})$	$\hat{\pi}_5(\pi_0^{s=(1,0)})$
(a)						
0	0.010	0.049	0.085	0.081	0.016	0.015
	$\tau_{11}=0.229$	(0.008,2.833)	$\tau_{01}=1.039$	(0.345, 3.554)	$\tau_{10}=1.031$	(0.030,37.439)
0.2	0.011	0.044	0.100	0.097	0.019	0.016
	$\tau_{11}=0.271$	(0.009,2.691)	$\tau_{01}=1.017$	(0.259,3.969)	$\tau_{10}=1.152$	(0.032,36.779)
0.5	0.014	0.039	0.137	0.140	0.024	0.018
	$\tau_{11}=0.385$	(0.014,2.945)	$\tau_{01}=0.962$	(0.126,5.020)	$\tau_{10}=1.289$	(0.036,45.188)
0.8	0.024	0.037	0.206	0.291	0.033	0.017
	$\tau_{11}=0.659$	(0.030,2.702)	$\tau_{01}=0.725$	(0.032,6.652)	$\tau_{10}=1.875$	(0.042,69.652)
(b)						
0	0.010	0.025	0.108	0.144	0.093	0.126
	$\tau_{11}=0.094$	(0.003,2.466)	$\tau_{01}=0.746$	(0.324,1.408)	$\tau_{10}=0.748$	(0.310,15.170)
0.2	0.010	0.126	0.130	0.170	0.104	0.124
	$\tau_{11}=0.089$	(0.003,2.154)	$\tau_{01}=0.766$	(0.329,1.570)	$\tau_{10}=0.832$	(0.331,16.610)
0.5	0.010	0.125	0.182	0.198	0.115	0.124
	$\tau_{11}=0.088$	(0.003,2.107)	$\tau_{01}=0.941$	(0.388,3.102)	$\tau_{10}=0.920$	(0.275,17.930)
0.8	0.027	0.124	0.196	0.105	0.085	0.125
	$\tau_{11}=0.225$	(0.038,4.607)	$\tau_{01}=1.726$	(0.390,42.210)	$\tau_{10}=0.706$	(0.101,14.210)

Table 6.3 provide causal risk ratio estimates (Bayesian principal effects) obtained from mean posterior median relative risks for each stratum and corresponding mean 95% credible intervals for different values of sensitivity parameter ϕ. Here a posterior relative risk τ was obtained as the ratio of two probabilities of experiencing an event due to compliance with one treatment allocation relative to another in a stratum. Most of results (not all) show posterior median relative risks of less than one for all values of ϕ which indicates lower risks for ACM and MRCD for those women randomized to HRT who would be highly compliant with their treatment allocation. Of primary interest is the quantity $\tau_{11} = [\pi_1^{S=(1,1)}][\pi_0^{S=(1,1)}]^{-1}$, i.e. the posterior relative risk for mortality (or reinfarction) due to HRT treatment compared to placebo for the subgroup which would comply with either intervention (type 3). The results shows that the mean 95% credible intervals widened with increase in ϕ values, an indication of less correlation between HRT treatment and placebo compliances. Overall, the results indicated that HRT tablets reduced risks for myocardial reinfarction more than the reduction in risks for all-cause mortality.

For a moderate correlation value ($\phi = 0.5$), the results suggest that compliance with HRT treatment compared to taking placebo among those who would comply with either treatment reduced the risk for all-cause mortality by 61%, i.e. causal risk ratio $\tau_{11} = 0.385$, 95% CI : $0.014, 2.945$. Also for this value of sensitivity parameter, compliance with HRT tablets only would marginally reduce the risk of death by about 4%, i.e. causal risk ratio $\tau_{01} = 0.962, 95\%$ CI : $0.126, 5.020$. On the other hand, compliance with placebo treatment only would be harmful for all values of ϕ, for example, while compliance with placebo would increase the risk for death by 29% ($\tau_{10} = 1.289, 95\%$ CI : $0.036, 45.188$) when $\phi = 0.5$, the risk would increase substantially by 86% ($\tau_{10} = 1.875, 95\%$ CI : $0.042, 69.652$) when $\phi = 0.8$.

In general although the results suggest that compliance with HRT treatment compared to placebo would significantly reduce risks for both death and reinfarction among those who would comply with either treatment, the estimates are very different from those obtained earlier using specialist methods for one active arm and even ITT (see Chapter 4: Tables 4.3, 4.8,4.9 and 4.11). This seems likely to be due to a failure of the method as applied to Esprit data as may be deduced from the implausible causal risk ratio estimates

obtained above especially for the very low risks (τ_{11}) for reinfarction due to compliance with HRT treatment compared to placebo among the highly compliant for all ϕ values. We may also discern lack of precision and/or failure in the method manifested in wider ('abnormal') mean 95% credible intervals corresponding to the causal risk ratio estimates of efficacy due to compliance with placebo treatment only (τ_{10}) for both mortality and reinfarction outcomes. Furthermore, we observe that these 95% credible intervals became wider as the sensitivity parameter ϕ increased.

The size of causal (principal) effects varied according to value of sensitivity parameter. Risks for both death and reinfarction due to compliance with HRT tablets relative to placebo for those who would comply with either treatment increased with increase in the value of sensitivity parameter ϕ. Conversely, results show decrease in risks for both death and reinfarction due to compliance with HRT treatment only as the value of ϕ increases. As expected the risk for death due to compliance with placebo only was higher regardless of the level of sensitivity ϕ.

Table 6.4: Comparison of results from specialist methods and Roy et al. (2008) method

	Specialist methods Hazard ratio (HR)			Roy et al. (2008) method $\text{RR}^{\ddagger} = \omega\hat{\tau}_{11} + (1-\omega)\hat{\tau}_{01}$			
				ϕ			
Outcome	CALM†	CHARM	C-Prophet	0	0.2	0.5	0.8
Mortality	0.504	0.660	0.656	0.575	0.532	0.571	0.666
Reinfarction	0.990	0.988	0.988	0.374	0.327	0.289	0.411

†HR Weibull model (Table 4.4 & Equation 2.4); ‡Relative risk weighted using proportions in Table 6.2

Roy et al. (2008) argued that the two subgroups (types 3 and 2) who would comply with either treatment and only active HRT treatment respectively may be of interest. Table 6.4 provide a comparison of results from the specialist methods (discussed in Chapter 4) which considered compliance in the active treatment arm only while ignoring placebo compliance and the posterior relative risks from the Roy et al. method which accounts for placebo compli-

ance among the subgroup who would comply with active HRT treatment for different values of sensitivity parameterϕ. To obtain an overall efficacy estimate for comparison with specialist methods, we weighted the posterior relative risks (RR) with the proportion of compliance for each stratum provided in Table 6.2: RR$=\omega\hat{\tau}_{11} + (1-\omega)\hat{\tau}_{01}$, where $\hat{\omega} = \hat{\mu}_{11}/(\hat{\mu}_{11} + \hat{\mu}_{01})$.

The hazard ratio estimates from both C-Prophet and CHARM methods were comparable with the causal risk ratios from the Roy et al. method at high value of the sensitivity parameter ($\phi = 0.8$) for all-cause mortality outcome. This may be an indication of (high) correlation between compliances with placebo and HRT treatment. The CALM method also produced comparable results to Roy et al. method at low value of the sensitivity parameter ($\phi = 0.2$) for all-cause mortality outcome: compliance with HRT treatment relative to placebo would reduce risk of death by 47%. But in general results from the Roy et al. method were very different compared to those obtained by the specialist methods, i.e, the results varied with change in the values of sensitivity parameter ϕ. The apparent failure of the Roy et al. method may be attributed to the (strong/untestable) parametric assumption of the association model linking the two marginal models.

Overall, the CALM method performed 'best' among the specialist methods. Specifically, by adjusting for both crossovers and censoring the CALM results suggest compliance with HRT treatment would reduce risk for mortality by half (HR$= 0.5$) among those women who would comply with HRT medication compared to those on placebo. The variation in efficacy estimates from the Roy et al. method may be an indication of differences in compliance behaviour in the two arms. Perhaps the Roy et al. (2008) model captured the underlying correlation between such compliance behaviours in respective arms using the sensitivity parameter ϕ, hence making the results dependent on it. But we do not know the value of ϕ and so the method should not be recommended for the Esprit data.

As outlined in Chapter 2 (Section 2.7), application of the Roy et al. (2008) method is premised on plausibility of the the crucial (but untestable) assumption that the outcome is independent of the set of covariates predictive of treatment compliance given a compliance type and treatment assignment. Hence the task of selecting suitable predictors of treatment

Table 6.5: Sensitivity analysis of the Roy et al. (2008) method: Relative risk (mean 95% CI)

	All-cause mortality			Myocardial reinfarction		
ϕ	τ_{11}	τ_{01}	τ_{10}	τ_{11}	τ_{01}	τ_{10}
A						
0	0.256	1.036	1.041	0.062	0.733	0.851
	(0.040,3.750)	(0.101,3.359)	(0.299,41.897)	(0.017,2.368)	(0.296,1.383)	(0.474,19.696)
0.2	0.243	1.007	1.128	0.057	0.744	0.863
	(0.043,3.245)	(0.080,3.525)	(0.227,37.919)	(0.016,2.889)	(0.283,1.472)	(0.449,21.909)
0.5	0.363	1.001	1.164	0.053	0.910	0.854
	(0.053,2.847)	(0.056,5.962)	(0.078,41.729)	(0.014,2.889)	(0.295,3.180)	(0.264,23.516)
0.8	0.623	0.709	1.741	0.208	1.753	0.471
	(0.209,3.251)	(0.045,6.932)	(0.011,71.002)	(0.129,3.494)	(0.277,43.438)	(0.027,15.675)
B						
0	0.299	1.039	1.031	0.094	0.746	0.748
	(0.008,2.833)	(0.345,3.554)	(0.030,37.439)	(0.003,2.466)	(0.324,1.408)	(0.310,15.170)
0.2	0.271	1.017	1.152	0.089	0.766	0.832
	(0.009,2.691)	(0.259,3.969)	(0.032,36.779)	(0.003,2.154)	(0.329,1.570)	(0.331,16.610)
0.5	0.385	0.962	1.289	0.088	0.941	0.920
	(0.014,2.945)	(0.126,5.020)	(0.036,45.188)	(0.003,2.107)	(0.388,3.103)	(0.275,17.930)
0.8	0.659	0.725	1.875	0.225	1.726	0.706
	(0.030,2.702)	(0.032,6.652)	(0.042,69.652)	(0.038,4.607)	(0.390,42.687	(0.101,14.210)
C						
0	0.265	1.057	1.034	0.076	0.737	0.830
	(0.041,2.419)	(0.109,2.557)	(0.197,31.790)	(0.020,2.880)	(0.291,1.439)	(0.442,17.672)
0.2	0.262	1.011	1.137	0.067	0.761	0.828
	(0.048,2.127)	(0.084,2.063)	(0.101,31.760)	(0.017,2.620)	(0.289,1.600)	(0.392,1.961)
0.5	0.376	1.006	1.147	0.054	0.951	0.733
	(0.065,2.516)	(0.059,4.932)	(0.032,39.478)	(0.014,2.520)	(0.321,3.566)	(0.175,3.392)
0.8	0.648	0.717	1.712	0.216	1.316	0.438
	(0.231,3.097)	(0.044,6.119)	(0.008,66.795)	(0.089,5.420)	(0.206,41.190)	(0.027,11.020)

Model comprising [A]3 (Lasso),[B]6 (intermediate) and [C]9 (all) predictors of compliance

compliance constitutes an integral part of ensuring plausibility of this assumption. A comparison of results for models composed of different sets of predictors may provide a form of sensitivity analysis for this assumption. Table 6.5 show results in terms of causal relative risks for models predicting compliance using three sets of predictors considered earlier in Chapter 3: $3, 6$ and 9 predictors from Lasso, intermediate and all plausible predictors respectively.

The causal risk ratio estimates using 3, 6 and 9 sets of predictors were comparable for all strata. In general, given a set of predictors, the results show same trend in principal effects with respect to change in magnitude of the sensitivity parameter ϕ for both outcomes (mortality and reinfarction). Specifically, the causal risk ratios suggest marginally lower risks for both death and reinfarction when using fewer covariates selected using the Lasso method of model selection compared to relative risks resulting from use of 6 and 9 predictors of compliance. This may be an indication that the advantages of classical model selection are transferable to the Roy et al. (2008) method via use of optimal marginal compliance models. However, we note that while selecting plausible predictors of compliance forms an integral component of the method, model selection only acts as an intermediate step that provides covariates for marginal compliance prediction models which are then joined into a causal model using the crucial but unknown sensitivity parameter.

Results from the sensitivity analysis above (Table 6.5) may be useful tool to demonstrate the phenomenon that causal (principal) effects are dependent on the choice of covariates predicting compliance. This may be an indication of failure (implausibility) of the Roy et al. (2008) method as applied to the Esprit data owing to inadequate compliance data. Also the method's defining assumption that the outcome is considered independent of the set of covariates predicting compliance for a given stratum may not be plausible for the Esprit data especially with regard to history of hysterectomy and cerebrovascular risks given the likelihood of association between these factors and treatment compliance and hence possible efficacy.

In Chapter 8 we will apply a statistically designed simulation study to evaluate the performance of the Roy et al. (2008) method in terms of bias and 95% credible intervals under both homogeneous and heterogenous treatment effects assumptions.

Chapter 7

Monte Carlo Study I: Performance of Statistical Methods for Analysing Survival Data in the Presence of Nonrandom Compliance

7.1 Introduction

The objective in this chapter is to evaluate six statistical methods used to analyse the Esprit data: three naive and three specialist methods. The next section provides a summary of statistical properties that will be used to compare the performance of estimators in this chapter and the next one. In section three we outline the aims of the simulations followed by a section that describes the simulations set-up. The next section presents details of the six methods of analyses followed by the results and the final section provide the conclusions.

7.2 Properties of statistical estimators

Statistical analysis in this chapter and next will involve comparing estimates from different (competing) statistical methods. In this section we review some standard statistical concepts that will help us make these comparisons. Most statistical inference involve use of point estimators. Since such estimators are random variables, their desirable properties can be

evaluated from the characteristics of their sampling distributions (Casella and Berger, 2001; Shervish, 1995). Some of the desirable properties of estimators commonly used for statistical inference include unbiasedness, mean square error and most efficient estimators.

7.2.1 Unbiased estimator

Bias is a prevalent phenomenon that can be introduced in a study at any state, e.g. during design or execution. It can mostly be controlled but very difficult to avoid all together. Bias may ordinarily be described as the amount of deviation from the truth. We note that the statistical meaning of bias as used here is narrower that defined earlier in Chapter 1 (Section 1.6). For a random sample from a population $X_1, X_2, \ldots, X_n$, let $\hat{\beta}(X_1, X_2, \ldots, X_n)$ be an estimate for the unknown parameter β. An estimator $\hat{\beta}$ is said to be unbiased for β if

$$E[\hat{\beta}(X_1, X_2, \ldots, X_n)] = \beta,$$

i.e. on average, the estimator is right (with zero bias): $E(\hat{\beta}) - \beta = 0$. Otherwise the estimator is said to be biased if $E(\hat{\beta}) \neq \beta$. Formally, if $\hat{\beta}$ is an estimator of a parameter β, then the bias of $\hat{\beta}$ is defined as

$$\text{Bias}(\hat{\beta}) = E(\hat{\beta}) - \beta.$$

The presence of bias will result in a parameter (e.g. mean hazard ratio or relative risk) consistently overestimated or underestimated which implies inaccurate measure of efficacy.

7.2.2 Variance and mean squared error

The precision of estimates are often provided in terms of variance and/or width of confidence intervals. A confidence interval provides a range of values around the mean that is likely to contain the estimate. On the other hand, variance is a measure of a description of how far individual values for a given dataset vary from the mean. Formally, a sample variance is

defined as the average of the squared deviation from the mean:

$$\text{Var}(X_i) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

where $\bar{X}$ is the mean of Xs and $\sqrt{\text{Var}(X_i)}$ is the standard deviation.

An estimate with a smaller variance or narrower confidence interval is considered comparatively more precise, i.e. the narrow confidence interval and/or smaller variance indicates less variability between individual values. We often consider an estimate as accurate if it is precise and unbiased. A general criterion for assessing how good an estimator $\hat{\beta}$ is for β that combines both its relative magnitude of bias and variance (precision) is the mean squared error (MSE) which is defined as the expected value of the square of the sampling error:

$$\begin{aligned}
\text{MSE}(\hat{\beta}) &= E(\hat{\beta}-\beta)^2 = E[\hat{\beta} - E(\hat{\beta}) + E(\hat{\beta}) - \beta]^2 \\
&= E[\hat{\beta} - E(\hat{\beta})]^2 + 2\underbrace{E\left[(\hat{\beta} - E(\hat{\beta}))(E(\hat{\beta}) - \beta)\right]}_{=0} + E[E(\hat{\beta}) - \beta]^2 \\
&= E[\hat{\beta} - E(\hat{\beta})]^2 + [E(\hat{\beta}) - \beta]^2 = \text{Var}(\hat{\beta}) + [E(\hat{\beta}) - \beta]^2 \\
&= \text{Var}(\hat{\beta}) + [\text{Bias}(\hat{\beta})]^2.
\end{aligned}$$

The MSE is a useful measure of overall accuracy (Collins et al., 2001): it is a sum of bias which measures how far off the estimator is (on average and variance that measures the variability of the estimator, i.e MSE($\hat{\beta}$)=Var($\hat{\beta}$) for an unbiased estimator. The MSE provide squared values but sometimes it may be meaningful to use root mean squared error (RMSE)= $\sqrt{\text{MSE}}$. We will use RMSE for our analyses to evaluate the performance (accuracy) of our estimators in terms of bias.

A key concept behind a confidence interval is coverage probability, i.e. the proportion of the time that the interval contains the true specified value of interest (Burton et al., 2006; Dodge, 2003). Simulation studies are often used to evaluate the performance of estimates from competing methods in terms of the coverage probability where a higher coverage probability is an indication that Type I error rate (see below) for testing a null hypothesis of no effect is properly controlled (Collins et al., 2001). Following Burton et al. (2006), our simulation will aim to show this using the nominal value of 95%.

7.2.3 Most efficient estimator

Making a choice between two unbiased estimators in terms of their respective variances (precision) is a challenge given the infinite possibility of other estimators (Casella and Berger, 2001; Kapadia et al., 2005; Panik, 2005). The Cramer-Rao lower bound provide a solution in which we choose the estimator with the least variance among the unbiased set, i.e. if the variance for some β is equal to this lower bound, then there is no estimator that is more precise. Specifically, if $\hat{\beta}$ is an unbiased estimator of β, then under some regularity conditions[1], the variance of $\hat{\beta}$ must satisfy the Cramér-Rao (CR) inequality

$$\mathrm{Var}(\hat{\beta}) \geq -E\left[\frac{\partial^2 \log \mathcal{L}}{\partial \beta^2}\right]^{-1} = CR(\beta, n), \tag{7.1}$$

where $\log \mathcal{L} = \sum_{i=1}^{n} \log f(x_i; \beta)$. If $E(\hat{\beta}) = \beta$ (unbiased) and strict equality holds in (7.1), then $\hat{\beta}$ is the most efficient (or minimum variance bound) estimator of β. In other words, the most efficient estimator is the unbiased estimator whose variance equals the right hand side of Equation (7.1).

7.2.4 Type I and Type II errors

In statistically testing hypothesis to evaluate any difference between two treatments, the possible errors include committing Type I error (false positive) and Type II (false negative). Type I error occurs when a researcher concludes (on the basis of experimentation) that an experimental treatment has a statistically significant effect over the comparator, when in fact it does not. Researchers often use p-values to estimate the probability of committing Type I error, for example $p < 0.05$ (also called 5% significance level) means 1 in 20 chance of committing Type I error.

On the other hand Type II error occurs when a researcher misses a significant effect

[1] (i) the sample space Ω is independent of β; (ii) the derivatives $\partial\mathcal{L}/\partial\beta, \partial^2\mathcal{L}/\partial\beta^2$ exist for all admissible β; and (iii) $E(\partial \log \mathcal{L}/\partial\beta) = 0$.

of the treatment on an experimental group. Power of a test is the probability of making the right decision when the null hypothesis is not correct, i.e. correctly concluding effect of experimental treatment. Low power and small sample sizes are main causes of possible failure to demonstrate effect if it truly exists (van Belle et al., 2004).

7.3 Aims of the simulations

We consider the methods dealing with noncompliance in the context of a randomized controlled trial comparing an active treatment and control in terms of survival, using simulation studies when non-compliers comply for part of their treatment period. The primary interest is to compare performance in terms of induced bias and root mean squared error of methods which assume all-or-nothing compliance (i.e. non-compliers do not take any of the assigned treatment) throughout the study to methods assuming partial compliance, especially when compliance is nonrandom. For comparison, these methods are also evaluated under random compliance.

An important feature of methods which allow partial compliance in the active arm is the assumption about duration of treatment effects. Such models assume that risks for the non-compliers in the treatment arm, once they cease treatment, instantly revert to their counterfactual 'no treatment' risk, thus assuming that any effect of treatment stops immediately one becomes a non-complier. To reflect this, our simulation model assumed that treatment effects stop immediately. In evaluating the results, it is also important to note the separate effect of heterogeneous risks within groups on the performance of the proportions hazards (PH) methods. Even with full compliance, application of a simple Cox PH model which assumes homogeneous risks is likely to produce biased estimates of the treatment causal effect in these circumstances. This is because the most frail people are selected out of the data in the early stages so that the marginal hazard ratio for survivors actually changes over time. The formula by Aalen (1998) shows the time relationship of the marginal hazard ratio, HR(t)

say, when within group variation in risk follows Gamma distribution:

$$\mathrm{HR(t)} = \mathrm{HR(0)} \left[\frac{1 + \upsilon\ \mathrm{H(t)}}{1 + \upsilon\ \mathrm{HR(0)}\ \mathrm{H(t)}} \right], \tag{7.2}$$

where H(t) is the cumulative baseline hazard to time t and υ the standardized within-group variance of the hazard rates. We will use this formula to predict the typical degree of bias that would arise by fitting a non-time dependent PH model even with full compliance. We then use this prediction to gauge the extent to which any bias in the compliance adjusted methods described above might be due to ignoring heterogeneity rather that a flaw in the compliance adjustment procedure.

7.4 Simulations design

The simulation study mimicked a two-armed randomized trial lasting 24 months with control and active treatments; there were 2000 replications for each scenario in order to ensure coverage lies within two standard errors of the nominal 95% coverage probability (Tang et al., 2005). Also to mimic Esprit data, each simulation assumed a sample size of 1000 with equal probability of being randomly assigned to either treatment arm.

Each individual had two potential hazard rates which do not vary with time: λ_{1i} and λ_{0i} depending on whether or not they were taking active treatment at that time. The hazard rates $\{\lambda_{0i}\}$, assumed constant over time, were generated from a Gamma distribution with shape parameter 2 and scale parameter 0.006 so as to have mean 0.012 and variance 7.2×10^{-5}. The simulation model considered events in each month separately. Time to event for individual i was assumed to have an exponential distribution with parameter λ_{0i} or λ_{1i} but the simulation model considered events in each month separately so that the parameter could be changed according to treatment changes. Whether someone died and, if so, the time to death was set in two steps: first, the probability of dying in a given month, e.g. in the control arm, was set equal to $1 - \exp(-\lambda_{0i})$ for an individual i, who had survived to

the beginning of the month; random numbers from the uniform distribution were used to decide if the individual actually died. Next, for those who died in a given month, the time, t, since the beginning of the month was generated using random numbers and the conditional distribution $\Pr[T<t|T<1]=[1-\exp(-\lambda t)]/[1-\exp(-\lambda)]$. All those who survived to 24 months were treated as censored data in the analyses.

We assigned to each individual i a probability of noncompliance, α_{1i}, at end of month one of follow-up; for members in the control group, these referred only to potential noncompliance if, counter to fact, they had been assigned to the active group. To achieve non-random noncompliance, the conditional distribution $\log \alpha_{1i}|\lambda_{0i}$ for an individual's hazard rates in the control arm was chosen using a random number algorithm for generating Normally-distributed correlated data (Kleijnen, 1974). A positive correlation corresponded to the assumption that more frail patients were less likely to comply with their treatment. The choices for the correlation with log λ_{1i} resulted in $\text{corr}(\lambda_{0i}, \alpha_{1i}) = \rho = 0.5$ in the main set and, for comparison purposes $\rho=0.2, 0.8$ and -0.5. The mean probability of noncompliance $\bar{\alpha}_1$ in month one was set at 0.05 and $\alpha_{1i} \sim$ log-normal with a SD of 0.03. All generated values were in the range $(0, 1)$. The proportion of new non-compliers, among those who had been previously compliant, was assumed to decrease with time. This was achieved by setting the probability of noncompliance α_{ki} in month $k(k>1)$ for those compliant in month $k-1$ as follows: $\alpha_{ki} = \alpha_{1i}$ for $k = 2, \ldots, 6$, $\alpha_{ki} = 0.4\alpha_{1i}$ for $k = 7, \ldots, 12$ and $\alpha_{ki} = 0.2\alpha_{1i}$ for $x = 13, \ldots, 18$. We assumed no new noncompliance during the months 19-24. For new non-compliers in month k, noncompliance was assumed to take effect at the end of month k. Once a patient became a non-complier they remain so until end of the study.

In separate simulations compliance was assumed to be random but the overall probability of noncompliance in any month was set with the same mean as before, e.g. for each person the probability of noncompliance in month 1 was binomial with mean $\bar{\alpha}_1$.

Our simulation models assumed a homogeneous treatment effect: that is $\lambda_{1i}=[\exp(\psi)]\lambda_{0i}$ where $\exp(\psi)$ is the true causal hazard ratio, which was assumed to be either 0.5 or 1 (null model). In the analyses, when a subject in the active treatment arm became a non-

complier, they instantly reverted to their risk λ_{0i}. Since the survival times are generated to have an exponential distribution, itself a special case of the Weibull, the relationship $\beta = -\log\varphi$ between the proportional hazard model parameter β and the accelerated failure model parameter φ should hold.

7.5 Methods for comparison

Six analysis methods were applied to each simulated dataset. The three randomization-based methods for noncompliance adjustment (d-f) are based on the structural models C-Prophet, CALM and CHARM respectively while the other non-structural models (a-c) may be considered as naive regression approaches. The structural models are to some extent defined in terms of partially unobserved variables while the regression models are defined in terms of fully observed variables.

7.5.1 Notation

The following notation is used for all the methods:

- $Z = 1(0)$: denotes randomization to active (control) arm.
- T_i: time-to-event, which may not be observed in either group due to censoring.
- W_i: potential time-to-event if subject never received active treatment; only observed in the control arm.

  For all-or-nothing compliance models:

- $U = 1(0)$: potential compliance if offered active treatment. It is observed in the active arm only.
- $V = 1(0)$: refers to observed compliance with active treatment and is fully observed. $V = U$ in active arm and $V = 0$ in control arm.

For time-dependent compliance models:

- $V(t) = 1(0)$: refers to observed compliance with active status at time t and is fully observed; $V(t) = 0$ in control arm.
- C: time to potential noncompliance if offered active treatment, assuming those who stop treatment remain non-compliant; observed for active arm only.

7.5.2 Methods

(a) Cox_{ITT}: this is the usual ITT Cox PH model that ignores noncompliance:

$$h(t|Z_i) = h_0(t)\exp(\beta Z_i), \tag{7.3}$$

where $\text{h}(t|Z_i)$ denotes the hazard rate for failure at time t given exposure, $h_0(t)$ is the baseline hazard. Here HR_{ITT} is estimated by $\exp(\hat{\beta})$.

(b) Cox_{Reg1}: This Cox proportional hazards model attempts to allow for all-or-nothing noncompliance through simple regression adjustment:

$$h(t|Z_i, N_i) = h_0(t)\exp[\beta_1 Z_i + \beta_2(1 - V_i)], \tag{7.4}$$

where V is fixed over time. The naive treatment effect estimator is $\exp(\hat{\beta}_1)$ and the effect of noncompliance is estimated by $\exp(\hat{\beta}_2)$. To apply this method to our partial compliance simulated data, in the active arm V was set to zero if there was any noncompliance.

(c) Cox_{Reg2}: This also has a regression adjustment but now V(t) is time dependent to allow for partial compliance:

$$h(t|Z_i, N_i) = h_0(t)\exp[\beta_1 Z_i + \beta_2(1 - V_i(t))]. \tag{7.5}$$

This model was fitted in Stata using the standard approach for time-varying covariates

in PH models. We note that methods (b) and (c) implicitly assume random noncompliance and are not expected to perform well when noncompliance is associated with underlying risk, i.e. nonrandom.

(d) C-Prophet: This structural causal proportional hazards model assumes all-or-nothing compliance and takes the following form:

$$h(t|Z_i = 1, U_i = u) = h(t|Z_i = 0, U_i = u) \exp(\psi_{0u}), \tag{7.6}$$

where U_i is potential all-or-nothing compliance with active treatment for subject i. Model (7.6) makes the exclusion restriction assumption (Imbens and Rubin, 1997): that is, there is no effect of assignment to active arm on survival in the subgroup $U_i = 0$ of non-compliers and that U_i is independent of randomization. These assumptions facilitate nonparametric identification of the survival distribution for potential compliers in the control arm, and hence estimation of the log hazard ratio ψ_0 for active versus control conditions in the subgroup ($U_i = 1$) of potential compliers. In our implementation, ψ_0 is equivalent to ψ in the simulation model. Censoring for this method is considered non-informative for the control arm as a whole, while in the active arm censoring is non-informative conditional on compliance. To apply the model to our partial compliance data, partial compliers were assumed to be completely non-compliant. The method was implemented in Stata using the `stcomply` command (Kim and White, 2004).

(e) CALM: This structural Causal Accelerated Life Model relates observed event times, T_i, to the potential event time W_i that would have occurred if, possibly contrary to fact, the subject had received no active treatment:

$$W_i(\varphi) = \int_0^{T_i} \exp[\varphi V_i(t)]\ dt, \tag{7.7}$$

where $V_i(t)$ is the time-varying indicator of active treatment receipt The exclusion restriction assumption here states that the baseline prognosis W is independent of randomization (White and Goetghebeur, 1998). The acceleration factor φ is the causal

effect of taking active treatment compared to control: for example, $\varphi < 0$ implies that taking continuous treatment would extend 'life' by a factor of $\exp(-\varphi)$. It is important to note that censoring that is non-informative on the T-scale may become informative on the W-scale. A key requirement in the CALM method is for the user to provide a potential recensoring time on the W-scale for all subjects. In our simulations which assume a fixed follow-up time with censoring at 24 months, we followed the recommendation by Walker et al. (2004) and set the recensoring time at the end of study (24 months). The CALM method is implemented in Stata using the `strbee` command (White et al., 2002) in a number of ways depending on which test is used to find φ: $\hat{\varphi}$ is the value for which a test of difference in W_i between the two arms yields a zero test statistic. Here we present results from use of the logrank test but results using an alternative parametric test procedure based on the Weibull distribution were also generated. Estimation was performed by grid search over a range of values of φ which we specified as -3 to 1.5 for the non-null model and -2 to 2 for the null model with step size of 0.01. Given that simulation event times were exponentially distributed, the causal hazard ratio was estimated by $\hat{\psi}_i = -\log \hat{\varphi}_i$.

(f) CHARM: White et al. (2004) used the term Complier Average Causal Effect (CACE) for the possibly time-dependent estimand of this model but it is not immediately clear how it relates to the causal hazard ratio in our simulation model. By viewing survival data as a series of binary outcomes in consecutive unit time intervals, they defined CACE in time interval j as the risk ratio among survivors to that interval who would not stop treatment before end of interval j, i.e.

$$\text{CACE}_j = \frac{\Pr[T = j | T > j-1, Z = 1, C > j]}{\Pr[T = j | T > j-1, Z = 0, C > j]}, \tag{7.8}$$

where C is the potentially unobserved time to noncompliance. Key to obtaining the CACE estimates (7.8) for each interval is the 'extended exclusion restriction' assumption that randomized allocation has no effect on any survivors to the start of interval j who would stop treatment by the end of that interval (White et al., 2004). Imple-

mentation of the method for continuous time requires two models to be fitted to the data; in the first, log $\mathrm{HR}_{\mathrm{ITT}}$ is modelled as a function of time and, in the second, the log odds of crossover among subjects who experience events in the treated arm must also be modelled as a function of time. Assuming short time intervals together with the extended exclusion restriction, White et al. (2004) showed that

$$\mathrm{CACE}_{\mathrm{PH}}(\mathrm{t}) = \frac{\mathrm{HR}_{\mathrm{ITT}}(t)[1-\theta(t)]}{1-\theta(t)\mathrm{HR}_{\mathrm{ITT}}(t)}, \tag{7.9}$$

where θ is the proportion of noncompliance probability among those participants who experience the event in the treatment arm. $\mathrm{CACE}_{\mathrm{PH}}$ is essentially a Causal Hazard ratio Adjustment Regression model (CHARM) with a linear predictor which is zero in the control arm and in the active arm is a function of time:

$$\mathrm{CHARM} \equiv h(t) = h_0(t)\exp[\beta_0 Z + \beta_1 f(t) Z], \tag{7.10}$$

where f(t) can be specified as a combination of linear/quadratic function of time to event. White et al. (2004) proposed two ways to estimate a constant CHARM:

(i) assume both ITT hazard ratio ($\mathrm{HR}_{\mathrm{ITT}}$) and odds of noncompliance θ are constant over time which allows a simple CHARM approximation by

$$\mathrm{CHARM} \equiv \widehat{\mathrm{CACE}}_{\mathrm{PH}} = \frac{\widehat{\mathrm{HR}}_{\mathrm{ITT}}(1-\hat{\theta})}{1-\hat{\theta}\ \widehat{\mathrm{HR}}_{\mathrm{ITT}}}. \tag{7.11}$$

(ii) a more complex procedure that allows both $\mathrm{HR}_{\mathrm{ITT}}$ and odds of noncompliance θ to vary over time.

We implement procedure (ii) above in our simulations by using the Stata command `adjhr` (White, 2002) which provide an 'approximate' overall (constant) CHARM estimate as hazard ratio adjusted for compliance.

Results from analyses of the simulations were summarized as follows. All methods were treated as giving an estimate $\hat{\psi}$ of ψ, the log causal HR in the simulation model even though naive regression models were not expected to perform well and the theoretical relationship between CHARM and ψ is not clear. We then calculated the mean of the estimators, $\bar{\hat{\psi}}$, and their corresponding root mean squared errors (RMSE) where $\text{RMSE}(\bar{\hat{\psi}}) = \sqrt{[\bar{\hat{\psi}} - \psi]^2 + \text{var}(\hat{\psi})}$. In the table we show mean effect on the HR scale calculated as $\exp(\bar{\hat{\psi}})$. We use a one-sided t-test with $\alpha = 0.05$ to test for bias with t statistic $\frac{\bar{\hat{\psi}} - \psi}{s/\sqrt{2000}}$, where s is the standard deviation of $\{\hat{\psi}_i\}$. Assuming that $s = 0.20$ or less, the simulation study was large enough to give 90% power to detect a bias of 0.01 or more on the ψ scale (i.e. $\bar{\hat{\psi}} - \psi$) for any statistical method. A non-significant test was taken as evidence of no important bias. Coverage was defined as the proportion of simulations where the 95% Wald confidence interval (or that generated by Stata's `strbee`, `adjhr` or `stcomply` procedures) contained the true parameter values. For simple regression adjustment models (7.4) and (7.5) we also report $\exp(\bar{\hat{\beta}}_2)$ (effect of noncompliance) to provide insight into how these models may mislead.

7.6 Results

The simulation allocated subjects almost equally in both arms: 503 and 497 for active and control arms respectively. On average there were 9% (91/1000) deaths in total: 44 and 47 in the placebo and active treatment arms respectively. The rate of compliance was relatively lower in the treatment arm (44%) compared to the placebo arm (61%). The proportion of non-compliers experiencing event of interest in the active arm (θ) decreased marginally with increase in the value of sensitivity parameter ϕ used, for example, while $\theta = 72\%$ for $\phi = 0.2$, this reduced to an average of 68% for $\phi = 0.8$. The simulations showed an average crossover of 57% (288/503) from the active to placebo arms (no switching from placebo to active treatment was permitted). The resulting correlation ρ between hazard rates and probability of compliance (used to induce non-randomness) was relatively smaller than the pre-specified value, e.g. a pre-specified $\rho = 0.5$, produced a value of 0.26. These summary results are

comparable to those of the Esprit data. We may then use the simulations to evaluate the performance of 5 methods (ITT, 2 simple regression adjustments, C-Prophet, CHARM and CALM) that were applied to the data in Chapter 4.

Using Equation (7.2) above, we estimated the degree of bias expected due to ignoring frailty in methods which are partly based on PH models as follows: we assumed that if a constant HR model is fitted to data when in fact the marginal HR(t) changes monotonically over time then we might expect that the constant model would produce an estimate somewhere near the middle of the range between HR(0) and HR(24). We therefore used (7.2) to calculate HR(12) with HR(0) $= \psi$. For the null model (7.2) gives HR(t) $= 1$ for all t and there is no bias. However when $\exp(\psi) = 0.5, \upsilon = 0.5$ and H(12) $= 0.144$, we get HR(12) $= 0.517$. In initial exploratory work we simulated full compliance trials under both homogeneous and heterogeneous risk models when $\exp(\psi) = 0.5$. The results gave $\exp(\bar{\hat{\psi}})$ as 0.500 and 0.515 respectively. These findings are in line with the above predictions, suggesting that our rough method of estimating bias due to ignoring variation in frailty alone was adequate.

Tables 7.1 and 7.2 show results for the null model ($\exp(\psi) = 1$) and non-null model ($\exp(\psi) = 0.5$) respectively under random and non-random compliance. Results from Table 7.1 show that the ITT method under both random and non-random compliance produced similar results - statistically insignificant bias. When noncompliance was random, the simple regression method adjusting for time-dependent noncompliance produced no bias (on the HR scale); however, the method performed less satisfactorily (bias $= -0.065$) when compliance was non-random. The regression method assuming all-or-nothing compliance performed worse under both random and nonrandom scenarios. One can see the reason for bias from the estimated effects of noncompliance in these models: in truth non-compliance had no effect but, for example, in the non-random scenarios, the estimate for all-or-nothing noncompliance implied that noncompliance decreased the risk of death by 20%, while in the time-dependent noncompliance model suggests that it increased the risk of death by 25%.

The specialist methods (C-Prophet and CALM) produced no bias under random compliance. While the CALM method was also unbiased under non-random compliance, the C-Prophet

Table 7.1: Performance of methods under random and non-random compliance when causal HR $\exp(\psi)=1$.

Model	$\widehat{\text{HR}}$ $(\exp(\bar{\hat{\psi}}))$	Non-comply* $(\exp(\bar{\hat{\beta}}_2))$	SE$(\hat{\psi}_i)$	RMSE$(\hat{\psi}_i)$	p-value†	95% CI Coverage
			Random compliance			
AON‡ compliance						
(a) Cox_{ITT}	0.998		0.132	0.132	0.566	0.946
(b) Cox_{Reg1}	1.144		0.145	0.198	< 0.001	0.838
Noncompliance		0.647				
(c) C-Prophet	1.009		0.186	0.186	0.031	0.893
Partial compliance						
(d) Cox_{Reg2}	0.997		0.146	0.146	0.325	0.951
Noncompliance		0.997				
(e) CALM	0.997		0.209	0.209	0.506	0.965
(f) CHARM	1.019		0.191	0.192	< 0.001	0.926
			Non-random compliance:		$\text{corr}(\lambda_{0i}, \alpha_i)=0.5$	
AON‡ compliance						
(a) Cox_{ITT}	0.999		0.128	0.128	0.836	0.952
(b) Cox_{Reg1}	1.072		0.142	0.159	< 0.001	0.923
Noncompliance		0.802				
(c) C-Prophet	1.012		0.189	0.190	0.005	0.878
Partial compliance						
(d) Cox_{Reg2}	0.935		0.142	0.157	< 0.001	0.957
Noncompliance		1.249				
(e) CALM	0.999		0.214	0.214	0.761	0.961
(f) CHARM	1.024		0.202	0.204	< 0.001	0.923

*Noncompliance effect - see Equations (7.4) and (7.5) ‡AON≡All-or-nothing;

†Testing significance of bias

method resulted in a small bias (0.012) that was statistically significant. The CHARM method produced bias that was statistically significant under both modes of compliance with the bias under random compliance (0.019) relatively smaller compared to bias under non-random compliance (0.024).

The SE and RMSE results show that there is a price to be paid for accounting for noncompliance using the specialists (C-Prophet, CALM and CHARM) methods. However, adjusting for noncompliance resulted in relatively more efficient estimates for the simple regression methods (Cox_{Reg1} and Cox_{Reg2}) and the ITT method (Cox_{ITT}). Among the specialist methods, the C-Prophet and CHARM methods produced more efficient estimates of treatment effects (smallest RMSE) under both random and non-random compliance modes compared to the CALM method. However, the CALM method produced the best (largest) 95% CI coverage rates for the corresponding estimates.

Table 7.2 shows results for the non-null scenario when the correlation between noncompliance and baseline log hazards is 0.5. As expected, the ITT method produced biased estimates of efficacy with a greater bias under the random (0.147) and non-random case (0.164). The bias was greater in the nonrandom case in all cases where noncompliance and baseline hazards were positively correlated but less when there was a negative association (see Table 7.3).

Simple regression adjustment with an all-or-nothing, binary noncompliance variable reduced but did not eliminate the bias under both modes of compliance, the bias being 0.085 and 0.047 respectively for random and non-random compliance. The reason for these biases can be seen in the biased estimates of the effect of noncompliance: while in reality it would double the risk, the estimated effects implied that noncompliance increased the risk of death by only 27% under random compliance and by 59% under non-random compliance. The latter estimate is closer to the truth and hence the corresponding bias is less. Regression adjustment with a time-dependent noncompliance variable failed to eliminate the bias when compliance was random although the bias was small (0.011) while under non-random compliance it was greater (-0.019). Again, one can see the reasons for the size and direction of these biases from the estimated effects of noncompliance: results from the time-dependent

Table 7.2: Performance of methods under random and non-random compliance simulated models when causal HR $\exp(\psi)=0.5$.

Model	$\widehat{\text{HR}}$ $(\exp(\bar{\hat{\psi}}))$	Non-comply* $(\exp(\bar{\hat{\beta}}_2))$	SE$(\hat{\psi}_i)$	RMSE$(\hat{\psi}_i)$	p-value†	95% CI coverage
			Random compliance			
AON‡ compliance						
(a) Cox_{ITT}	0.647		0.147	0.297	< 0.001	0.557
(b) Cox_{Reg1}	0.585		0.178	0.238	< 0.001	0.830
Noncompliance		1.273				
(c) C-Prophet	0.516		0.215	0.217	< 0.001	0.886
Partial compliance						
(d) Cox_{Reg2}	0.511		0.179	0.181	< 0.001	0.902
Noncomply		1.964				
(e) CALM	0.487		0.255	0.257	< 0.001	0.915
(f) CHARM	0.517		0.213	0.214	< 0.001	0.901
			Non-random compliance:		$\text{corr}(\lambda_{0i}, \alpha_i)=0.5$	
AON‡ compliance						
(a) Cox_{ITT}	0.664		0.144	0.318	< 0.001	0.495
(b) Cox_{Reg1}	0.547		0.177	0.198	< 0.001	0.917
Noncompliance		1.587				
(c) C-Prophet	0.518		0.202	0.203	< 0.001	0.890
Partial compliance						
(d) Cox_{Reg2}	0.481		0.173	0.177	< 0.001	0.937
Noncomply		2.454				
(e) CALM	0.487		0.271	0.273	< 0.001	0.946
(f) CHARM	0.516		0.215	0.216	< 0.001	0.919

*Noncompliance effect - see Equations (7.4) and (7.5)

‡AON≡All-or-nothing

†Testing significance of bias

noncompliance model suggest that the effect of noncompliance increased the risk of death by 96% (close to the truth) while they suggest noncompliance more than doubled the risk of death under the non-random mode of compliance.

All the three specialist methods (C-Prophet, CALM and CHARM) methods performed consistently under both compliance modes to produce small bias (on HR scale) that was statistically significant. Both the C-Prophet and CHARM methods produced similar results in terms of bias under both random (0.016 and 0.017) and non-random (0.018 and 0.016) compliances. The CALM method also produced similar bias (-0.013) under both random and non-random compliance. Although statistically significant, the biases under these three methods were similar to that predicted for PH models which ignore heterogeneity.

In terms of SE and RMSE, the C-Prophet and CHARM methods produced more efficient estimates of treatment effects under both random and non-random compliance modes compared to the CALM method. Although estimates from the C-Prophet and CHARM methods were equally efficient under random compliance, the C-Prophet method produced the most efficient results under non-random compliance. Although the corresponding 95% CI coverage rates for CHARM estimates were relatively better (large) compared to C-Prophet, overall the coverage rate for CALM method was best under both random and non-random compliance compared to other methods.

Adjusting for noncompliance under the specialists methods resulted in larger standard errors (SE) than for the ITT method but the later had the largest RMSE. It is interesting to note that in terms of RMSE, the non-specialist time-dependent regression methods performed best even under non-random compliance. Among the specialist methods, C-Prophet and CHARM were more efficient in terms of SE or RMSE than the CALM method.

Table 7.3 provides results for different correlations between hazards and noncompliance probabilities, again with $\exp(\psi) = 0.5$. The simple regression method adjusting for time-dependent compliance performed best to produce unbiased estimate at low correlation of $\rho = 0.2$. But the method produced larger bias with increase in correlations between the baseline haz-

Table 7.3: Effect of hazard-noncompliance probability correlation on performance of methods: $\exp(\hat{\psi})$ (RMSE) when $\exp(\psi) = 0.5$ (where $\rho^{\S}$ =corr (λ_{0i}, α_i))

Mean ρ	Cox_{ITT}	Cox_{Reg1}	C-Prophet	Cox_{Reg2}	CALM	CHARM
-0.5	0.632 (0.276)	0.621 (0.279)	0.513 (0.208)	0.544 (0.191)	0.485 (0.251)	0.519 (0.199)
0.2	0.656 (0.307)	0.571 (0.221)	0.518 (0.219)	0.500 (0.181)	0.488 (0.263)	0.517 (0.218)
0.5	0.664 (0.318)	0.547 (0.198)	0.518 (0.203)	0.481 (0.177)	0.487 (0.273)	0.516 (0.216)
0.8	0.674 (0.330)	0.521 (0.189)	0.520 (0.229)	0.460 (0.201)	0.488 (0.278)	0.518 (0.225)

§Set (attained) $\text{corr}(\lambda_{0i}, \alpha_i) : -0.5\ (-0.13), 0.2\ (0.11), 0.5\ (0.26), 0.8\ (0.44)$

ard and probability of noncompliance. In contrast, adjusting for all-or-nothing compliance resulted in smaller bias as correlations increased. The CALM method performed consistently to produce bias of similar magnitude for all correlation values considered. Both C-Prophet and CHARM methods produced similar (but biased) estimates of treatment effects; the size of the bias was fairly consistent regardless of the correlation and was of the order expected when within-group heterogeneity in risk is ignored in a proportional hazard model. However, when the results were judged in terms of RMSE, C-Prophet and CHARM performed better than CALM as before. The negative correlation scenario is interesting in that the all-or-nothing regression method increased bias; this also occurred in some other simulations with negative correlation although not all.

Finally, we observe that the resulting correlations between hazard rates and probability of noncompliance from the simulations were smaller compared to the set values. This may be an indication that the pre-specified correlation only controlled underlying hazard rates while observed data exhibited further randomness.

7.7 Discussion

The ITT and simple regression adjustments using binary noncompliance covariates produced large bias. All of the methods performed reasonably well under the null model in terms of bias. However, the simple regression approaches appeared to lead to a bias reduction when noncompliance is treated as all-or-nothing under non-random compliance and the bias was in the opposite direction to the causal treatment effect for time-dependent compliance adjustment. On the basis of the results for the null model alone, the C-Prophet and CHARM methods appeared to be the preferred specialist correction methods in the presence of random and non-random compliances respectively when judged by the RMSE.

In the non-null case of interest, i.e. nonrandom compliance, the simple regression adjustment methods cannot be recommended: they produce large bias which varied to some extent according to the magnitude and direction of correlation between the risk and probability of noncompliance. In this case and the random noncompliance case, the CALM method performed consistently best in terms of bias, i.e. CALM method consistently produced similar and smallest bias for the causal effect in one direction. The better performance of this method, compared to the C-Prophet and CHARM methods would appear to be attributable to the fact that it does not assume constant hazard rates. It is surprising that the all-or-nothing assumption underlying the C-Prophet method did not appear to result in extra bias compared to, say, the CHARM method. The C-Prophet produced the most efficient estimates (smaller RMSE) under non-random compliance albeit with low coverage which may be attributed to its implementation using jackknife procedures that may produce conservative variance estimates Goetghebeur and Loeys (2003). Overall the CALM method produced the best coverage rate for the estimates under both random and non-random compliance compared to other methods.

Considering nonrandom compliance may be more realistic than random compliance given the fact that noncompliance is often associated to other individual patient conditions. In a heterogeneous population, more frail patients are more likely to have higher probability

of treatment noncompliance since more frail patients may fail to see the need to continue treatment in the absence of immediate benefits.

Duration of treatment is an important phenomenon in methods which accounts for partial compliance. The specialist methods considered here may be extended to explore lagged treatment effects (bias), if any, up to some specified period after stoping. Such extension would help identify a method suitable for evaluating effects of treatment known to have long washout effects.

Finally clinical trials often have baseline covariates recorded at the beginning of most studies. Although it may not be straight forward for the specialist methods, incorporating potential prognostic factors while accounting for nonrandom compliance may assist in effective evaluation of treatment efficacy. Specifically, baseline covariates which predict treatment compliance may be used to relax the exclusion restriction assumption (Jo, 2002a) whose tenability is crucial for all the 3 specialist methods considered herein.

Chapter 8

Monte Carlo Study II: Evaluating Performance of Method Adjusting for Noncompliance in Two-Active Treatment Arms

8.1 Introduction

The objective in this chapter is to use simulations to evaluate potential bias induced by noncompliance when estimating efficacy from survival data from a two-active armed treatment trial. Specifically we use statistically designed simulations study to evaluate the performance of the Roy et al. (2008) model in terms of bias and 95% credible intervals as applied to survival data. We begin by outlining the aims of the simulations followed by a description of the simulation study design (set-up). The next section presents the methods of analysis. For both homogeneous and heterogeneous treatment effects cases, we first obtain the ITT estimate by applying the Cox proportional hazard model (ignoring any treatment compliance information) and evaluate resulting bias if viewed as estimating a causal hazard ratio. Under the homogeneous treatment effect assumption, each stratum assumed constant risk of death over time for both treatments A and B. For heterogeneous treatment effect case, the potential treatment effects among non-compliers to treatment A and B were set to be smaller than potential effects among compliers. Finally we present a comparison of the performance with respect to bias and 95% credible intervals of the causal effects in terms of

causal risk ratios obtained as means of posterior median relative risks for each stratum for both homogeneous and heterogeneous scenarios at different values of sensitivity parameter ϕ. The data are generated using Stata Software and analysis done in WinBUGS - see annotated codes in Appendix (Page 273).

8.2 Aims of the simulations

We use statistically designed simulation studies in the framework of a randomized controlled trial to compare two-active treatments in terms of survival to evaluate bias due to noncompliance in two treatment arms. First we evaluate the effect on the intention-to-treat (ITT) hazard ratio due to allocation to treatment B relative to treatment A where there can be noncompliance in either arm. As a check on the simulations, we also evaluate the ITT effects in each stratum. Next we apply Roy et al. (2008) model for survival data analysis (this analysis model was originally proposed for nonrandom compliance for binary outcome). The analysis requires specification a positive sensitivity parameter ϕ which is chosen as a function of arm-specific compliances and the correlation between compliances to treatment, i.e. parameter ϕ is not estimated from data. With two factors separately assumed predictive of compliance, we first construct arm-specific prediction models of compliance using logistic models from which we estimate the probabilities of compliance to treatment in each arm. We then obtain treatment effects for each stratum by in terms of causal risk ratios estimated from means of posterior relative risks parameters (outlined in Chapter 2, Section 2.7, Equation 2.25) as the ratio of probability of experiencing events/death among due to compliance with B relative to A, compliance with either treatment B or A compared to nothing among corresponding compliance types. Specifically, using Bayesian methods we estimate mean posterior median relative risks and corresponding mean 95% credible intervals of death in 3 different strata defined by their respective compliance types (compliance with A and B, A only and B only) while assuming nonrandom compliance under both homogeneous and heterogeneous treatment effects assumptions. We use death as the generic event/endpoint/outcome of interest.

8.3 Simulations design (set-up)

The simulation study mimicked a two-armed randomized trial with active treatments A and B lasting 24 months. There were 2000 replications for each scenario to ensure coverage lies within two standard errors of the nominal 95% coverage probability. Also to mimic Esprit data, each simulation assumed a sample size of 1000 with equal probability of being randomly assigned to either treatment arm.

Each subject had three potential hazard rates: $\lambda_{0i}, \lambda_{Ai}$ and λ_{Bi} corresponding to risk under no treatment and under treatment A and B respectively. The effects of both treatments are assumed better than no treatment at all in all cases. The time-invariant hazard rates $\{\lambda_{0i}\}$ were generated from Gamma distribution with shape and scale parameters 2 and 0.006 respectively so as to have mean 0.012 and variance 7.2×10^{-5}. Each stratum assumed constant risk of death over time for both treatments A and B. The simulation model considered events in each month separately. For a given month the probability of dying if a specific treatment is taken in any stratum were taken as equal to $1-\exp(-\lambda_{Ai})$ and $1-\exp(-\lambda_{Bi})$ for treatment A and B respectively. Random numbers from the uniform distribution were used to decide which subjects actually died from either treatment arm. Time to death was taken as the end of each month, i.e. the minimum time is 1 month for those who died in the first month while the maximum time is taken as 24. Subjects were allocated to treatment arms at random and risks chosen according to arm and potential compliance type. No switching of subjects between the treatment arms was assumed.

We considered all-or-nothing compliance to allocation for both treatments A and B up to 24 months. Each subject belonged to one of four complier groups (principal strata): type 0 ($S=(0,0)$) represent people who would be compliers to neither treatment, type 3 ($S=(1,1)$) represent potential compliers to either treatment, types 1 ($S=(1,0)$) and 2 ($S=(0,1)$) represent compliers to treatment A only and B only respectively. The compliance types were determined independently by a subject's associated risk factors X and her baseline risk of death. The simulation method first set the prevalence of covariates X (Table 8.1). Second

the relationship between probabilities of compliance and a pair of covariates predicting compliance is described by odds ratio (Table 8.2). Nonrandom compliance in each stratum was introduced using a specified positive sensitivity parameter ϕ which was chosen as a function of arm-specific compliances and the correlation between compliances to treatment: we explored different values of $\phi = 0, 0.2, 0.5, 0.8$. Random numbers (multinomial probabilities) were used to determined actual number of compliers for each stratum. Thirdly the compliance type was linked to the baseline hazard: we assigned highest ranked values of baseline risk λ_{0i} to represent subjects who would not comply with either treatment allocation (type 0) while the lowest ranked values of λ_{0i} are assigned to compliers of either treatment (type 3). From the remaining middle set, we assign at random to either compliance to treatment A only (type 1) or treatment B only (type 2) according their respective weighted proportions as set for simulation model (see later).

Compliance with treatment allocation is assumed to be predictable from two binary (0/1) baseline covariates, which we may think of in the context of Esprit data as histories of hysterectomy and cerebrovascular disease (CVD) risks (angina or blood pressure or stroke). The actual prevalence rates of histories of hysterectomy and CVD were set at 25% and 60% respectively. Subjects were assigned and the joint prevalence rates were set as shown in Table 8.1.

Table 8.1: Prevalence of risk factor.

	Hysterectomy	No hysterectomy	
CVD risk factors	0.15	0.45	0.60
No CVD risk factors	0.10	0.30	0.40
	0.25	0.75	

To link the risk factors and compliance, for each treatment arm, we specified three sets of statistics

(a) the probability of compliance to treatment allocation in the absence of both risk factors; this was 0.55 for A and 0.30 for B,

(b) a compliance odds ratio for hysterectomy: 2 for A and 5 for B and

(c) a compliance odds ratio for CVD risks: 4 for A and 3 for B.

The joint effect of both factors on compliance was assumed to be multiplicative on the odds ratio scale. This is the same as using a logistic model with no interaction term to obtain actual compliance probabilities for individual cells (see Table 8.2). These assumptions imply that the probabilities of compliance given a set of covariates X, $\mu_A(\mathrm{x})$ and $\mu_B(\mathrm{x})$, say for groups A and B are such that $\mu_A(\mathrm{x}) > \mu_B(\mathrm{x})$.

Table 8.2: Compliance probabilities for treatments A and B.

	Treatment A			Treatment B	
	Hyst	No hyst		Hyst	No hyst
CVD risks	0.907	0.830	CVD risks	0.865	0.563
No CVD risks	0.710	0.550	No CVD risks	0.682	0.300

We assume that potential compliance to treatment A and to B are positively correlated. The strength of correlation can be expressed as a correlation coefficient or as in Roy et al. (2008), by a positive sensitivity parameter ϕ (Chapter 2, Section 2.7, Equation 2.20) that is a function of arm-specific compliances and the correlation ρ between compliances to treatment. First we specified $\phi = 0.5$, but we also considered other values $\phi = 0$ (conditional independence), $0.2, 0.8$, and 1.0. We then worked out probability that compliance to A

Table 8.3: Compliance proportions by risk factors (for $\phi = 0.5$): μ_A and μ_B are marginal compliance probabilities - $\mu_A = \mu_{11} + \mu_{10}$ and $\mu_B = \mu_{11} + \mu_{01}$.

	Hyst+CVD	Hyst+no CVD	No hyst CVD	No hyst no CVD	Overall (weighted)
Prevalence	0.150	0.100	0.450	0.300	1.000
μ_B	0.865	0.682	0.563	0.300	0.541
μ_A	0.907	0.710	0.830	0.550	0.746
$\mathrm{U}(\mathrm{x}) = \frac{\mu_B}{\mu_A}$	0.954	0.961	0.677	0.545	0.725
μ_{11}	0.825	0.583	0.514	0.233	0.483
μ_{01}	0.040	0.099	0.048	0.068	0.058
μ_{10}	0.082	0.127	0.316	0.318	0.262
μ_{00}	0.053	0.191	0.122	0.383	0.197

is i and compliance to B is j, given x_1 and x_2, i.e. $\mu_{ij}(x_1, x_2)$ for a given value of ϕ (e.g. Table 8.3 for $\phi = 0.5$). We used random numbers and the multinomial probabilities $\mu_{ij}(x_1, x_2)$ to decide the number of subjects for each of the four compliance types in a given simulation.

In the simulation models, we considered both homogeneous and heterogeneous treatment effects. In some scenario, the potential treatment effects among non-compliers to treatment A and B were set to be smaller than potential effects among compliers hence corresponding to the heterogeneous treatment effects assumption. An homogenous treatment effect corresponded to scenario when potential treatment effects were assumed same for all principal strata: $\frac{\lambda_{Ai}}{\lambda_{0i}} = 0.75$ and $\frac{\lambda_{Bi}}{\lambda_{0i}} = 0.50$. For the homogeneous case $\lambda_{Bi} = [\exp(D)]\lambda_{Ai}$, where $\exp(D)$ is the true causal hazard ratio (THR), which was assumed to be 0.667 for each stratum. For the heterogeneous case, the potential treatment effects among non-compliers to treatment A and B were set to be smaller than potential effects among compliers. Specifically we set the causal HR at $0.667, 0.750, 0.778$ and 0.800 for stratum $3, 2, 1$ and 0 respectively, i.e. we set best benefit from treatment B relative to A for patients of type 3 $(1, 1)$, with the hazard ratio the same as in the homogenous case ($\text{THR}_{(1,1)} = 0.667$). The hazard rates for non-compliers among type 2 $(0, 1)$ patients was set to be relatively lower ($\lambda_{Ai} = \frac{2}{3}\lambda_{0i}$) compared to hazard rates for non-compliers among type 3 patients. Conversely the hazard rates for non-compliers among type 1 $(1, 0)$ patients was set to be relatively higher ($\lambda_{Bi} = \frac{7}{12}\lambda_{0i}$) compared to hazard rates for those classified to belong to type 3 (see Table 8.4). We use the ratio $\frac{\lambda_{Bi}}{\lambda_{Ai}}$ to obtain causal effects of treatment B relative to A for the subgroup who would comply with either treatment. (see definition of τ_{ij} later).

For the Roy et al. (2008) model, we obtained the true relative risk (TRR), i.e. causal risk ratio, as the ratio of average risk estimates for each arm in a group. Using moment generating function results of Gamma distribution ($\lambda_i \sim \text{Gamma}(\nu, k)$), we calculated

$$\text{TRR} = \frac{[1 - (1 + 24\bar{\lambda}_B)^{-k}]}{[1 - (1 + 24\bar{\lambda}_A)^{-k}]}, \tag{8.1}$$

i.e. TRR$= 0.729$ ($k = 2$) for the homogeneous case and for the heterogeneous case $\text{TRR}_S = 0.729$, $0.796, 0.824$ and 0.847 for stratum $S = 3, 2, 1$ and 0 respectively (Table 8.4).

Table 8.4: Stratum-specific hazard rates (among patient compliers).

Type	Homogeneous effects				Type	Heterogeneous effects			
(stratum)	$\bar{\lambda}_A$	$\bar{\lambda}_B$	THR	TRR†	(stratum)	$\bar{\lambda}_A$	$\bar{\lambda}_B$	THR	TRR†
3 (1,1)	0.009	0.006	0.667	0.729	3 (1,1)	0.009	0.006	0.667	0.729
2 (0,1)	0.009	0.006	0.667	0.729	2 (0,1)	0.008	0.006	0.750	0.796
1 (1,0)	0.009	0.006	0.667	0.729	1 (1,0)	0.009	0.007	0.778	0.825
0 (0,0)	0.009	0.006	0.667	0.729	0 (0,0)	0.010	0.008	0.800	0.847

†TRR calculated according to Equation (8.1)

8.4 Analysis methods

All ITT results were assumed to provide an estimate $\hat{D}$ of D, the log causal HR in the simulation model; then the mean of the estimators, $\bar{\hat{D}}$, and their corresponding root mean squared errors (RMSE) are calculated where $\text{RMSE}(\bar{\hat{D}}) = \sqrt{[\bar{\hat{D}} - D]^2 + \text{var}(\hat{D})}$. In the table we show mean effect on the HR scale calculated as $\exp(\bar{\hat{D}})$. We use a one-sided t-test with $\alpha = 0.05$ to test for bias with t-statistic $\frac{\bar{\hat{D}} - D}{s/\sqrt{2000}}$, where s is the standard deviation of $\{\hat{D}_i\}$. Assuming that $s = 0.50$ or less, the simulation study was large enough to give 90% power to detect a bias of 0.01 or more on the D scale (i.e. $\bar{\hat{D}} - D$) for any statistical method. A non-significant test was taken as evidence of no important bias.

The analysis begins by providing checks for the parameters set in the simulations. Next we obtain the ITT estimate by applying the Cox proportional hazards (PH) model ignoring treatment compliance in order to evaluate its bias (if any) for estimation of D. Specifically we evaluate the hazard ratio of death due to allocation to treatment B relative to treatment A for both homogeneous and heterogeneous treatment effects cases using the Cox PH model

$$h(t) = h_0(t)\exp[Z]: \qquad Z = \begin{cases} 1; & \text{if treatment B} \\ 0; & \text{if treatment A} \end{cases}, \tag{8.2}$$

where $\text{h}_0(\text{t})$ is the baseline hazard for a subject allocated to treatment A.

To apply the Roy et al. (2008) model, we predicted arm-specific compliance using the logistic models

$$\text{logit}\left[\mu_j(\text{x})\right] = \left(\sum_{i=0}^{2} \gamma_{ji}\text{x}_i\right), j = 0, 1 \tag{8.3}$$

where $\text{x}_0 = 1$ and x_1 and x_2 (in the Esprit context) are histories of hysterectomy and CVD risks. These factors were allowed to be separately predictive of compliance in each arm. From the logistic model, we estimated the probabilities of arm-specific compliance with treatment allocation using:

$$\hat{\mu}_j(\text{x}) = \left[1 + \exp\left(-\sum_{i=0}^{2} \hat{\gamma}_{ji}\text{x}_i\right)\right]^{-1}, j = 0, 1 \tag{8.4}$$

In a Bayesian setting, we used non-informative priors: normal distributions with mean zero and large variance, i.e. $\text{N}(0, 10^6)$, for the two potential predictors of compliance in each arm. We specified uniform $(0, 1)$ priors for the probabilities (risks) of death in each stratum given the arm of allocation and set the sensitivity parameter $\phi = 0, 0.2, 0.5$ and 0.8. Considering both homogeneous and heterogeneous cases, we ran three chains: null starting values for chain one, mean and median values from a trial run for chains two and three respectively. For convergence assessment, we ran simulation for $11,000$ iterations for each of the three chains and excluded the first $1,000$ as burn-in.

For the Roy et al. (2008) model, the results provide a summary of the mean compliance proportions for each treatment arm, performance of the resulting posterior median relative risk for each stratum under both homogeneous and heterogeneous treatment effects assumptions. The stratum-specific relative risks estimates of τ_{ij} are provided as ratio of probabilities of death among potential compliers to treatment B relative to A for each stratum given the arm of allocation. We used the corresponding standard deviation (SD) of the median of the estimators, $\tilde{\hat{\tau}}$, to calculate $\text{RMSE}(\tilde{\hat{\tau}}) = \sqrt{[\tilde{\hat{\tau}} - \tau]^2 + \text{var}(\hat{\tau})}$ and used a one-sided t-test with $\alpha = 0.05$ to test for bias with t statistic $\frac{\tilde{\hat{\tau}} - \tau}{SD/\sqrt{30{,}000}}$, where SD is the standard deviation of $\{\tau_{ij}\}$. Assuming that SD$= 2$ or less, the simulation study was large enough to give 90% power to detect a bias of 0.01 or more on the τ scale for any statistical method. Also a non-significant test was taken as evidence of no important bias.

We obtain causal risk ratio τ as mean of posterior median relative risk (decision based on minimizing linear loss function) for each stratum by estimating the τ_{ij} parameters (Equation 2.25):

(i) τ_{11} compares λ_B with λ_A in stratum $S=(1,1)$ using frailty relation (8.1) given above,

(ii) τ_{01} compares λ_B with λ_0 in stratum $S=(0,1)$, i.e. comparison of λ_B with baseline and

(iii) τ_{10} compares λ_A with λ_0 in stratum $S=(1,0)$, i.e comparison of λ_A with baseline.

8.5 Results

8.5.1 Checking on simulations

We obtained odds ratio estimates by fitting a logistic model to each simulation. For a moderate level of the sensitivity parameter $\phi = 0.5$, the mean compliance odds ratios for hysterectomy were 2.015 and 5.105 for treatment A and B respectively, and the compliance odds ratios for CVD risks were 4.041 and 3.025 respectively for treatment A and B. In general, these results and simulation results for other values of ϕ were in agreement with the odds ratios pre-specified in the simulation design.

Table 8.5: Estimates of mean compliance proportion per stratum for different ϕ values.

Stratum (type)	Homogeneous hazard rates ϕ 0	0.2	0.5	0.8	Heterogeneous hazard rates ϕ 0	0.2	0.5	0.8
$\bar{\hat{\mu}}_B = \bar{\hat{\mu}}_{11} + \bar{\hat{\mu}}_{01}$	0.542	0.543	0.543	0.544	0.541	0.542	0.542	0.541
$\bar{\hat{\mu}}_A = \bar{\hat{\mu}}_{11} + \bar{\hat{\mu}}_{10}$	0.744	0.745	0.747	0.749	0.745	0.746	0.746	0.746
$\bar{\hat{\mu}}_{11}$ (3)	0.426	0.449	0.484	0.518	0.425	0.449	0.483	0.518
$\bar{\hat{\mu}}_{01}$ (2)	0.115	0.092	0.057	0.023	0.116	0.092	0.058	0.023
$\bar{\hat{\mu}}_{10}$ (1)	0.319	0.296	0.261	0.227	0.320	0.297	0.262	0.228
$\bar{\hat{\mu}}_{00}$ (0)	0.139	0.162	0.197	0.231	0.139	0.162	0.197	0.231

Table 8.5 provide the mean (overall) proportion of compliance per stratum at different values of the sensitivity parameter ϕ for both homogeneous and heterogeneous hazard rates. On average, the compliance proportion results were similar to the pre-specified probabilities in the simulation design, for example while in Table 8.3, $\mu_{11} = 0.483$, the mean for the simulations was $\hat{\mu}_{11} = 0.478$. The mean proportions of compliance per stratum are similar for both cases: the mean proportion of compliance to treatment A was higher compared to mean compliance to treatment B (e.g. $\bar{\hat{\mu}}_A = 75\%$ and $\bar{\hat{\mu}}_B = 54\%$ when $\phi = 0.5$). As per our setup, the simulations ensured that potential compliers to either treatment (type 3) were the most frequent type while potential compliers to treatment B only (type 2) would be the least frequent no matter the value of ϕ. Overall, the mean proportion of compliance to either treatment (type 3) and neither treatment (type 0) dominated (increased) as the sensitivity parameter ϕ increased. On the other hand, the mean compliance proportion reduced with increase in ϕ values among those people who would comply with one treatment only (type 1 and 2). We note the small proportion of potential compliers to treatment B only (type 2) which approached total noncompliance as ϕ gets close to 1 (perfect correlation). In general, all the proportions of compliance were comparable to the expected weighted compliances proportions of the preset values (see Table 8.3). Overall, the general patterns/trend of compliance proportion was the same for both homogeneous and heterogeneous hazard rates.

8.5.2 Effect on ITT

Table 8.6: Homogeneous and heterogeneous ITT estimates.

	THR†	$\exp(\hat{D})$	SE$(\hat{D}_i)$	RMSE$(\hat{D}_i)$	p-value
Homogeneous	0.667	0.675	0.161	1.162	< 0.001
Heterogeneous	0.731‡	0.762	0.155	0.161	< 0.001

†True hazard ratio, Table 8.4 (‡weighted using proportions from Table 8.5)

Table 8.6 provide the ITT estimates for both homogeneous and heterogeneous cases. The ITT hazard ratio 0.675 for the homogeneous treatment effects case model suggested

that overall, the risk of death would reduce by 32% for those randomized to treatment B compared to those randomized to treatment A. The resulting small bias (0.008) for the ITT estimate was however statistically significant. For the heterogeneous treatment effects case, the ITT hazard ratio 0.762 indicated an overall reduction of risk of death by 24% for those randomized to treatment B compared to treatment A. The resulting bias (0.031) for the heterogeneous ITT estimate was also statistically significant. We observe a bias-precision tradeoff where as expected the bias due to homogeneous hazard was relatively smaller compared to bias from using heterogeneous hazard rates but the later had relatively smaller SE compared to the former. However, in general we note that a study population is more likely to be heterogeneous than homogeneous.

Table 8.7 provide estimates for each stratum in terms of hazard ratio for both homogeneous and heterogeneous hazard rates at different sensitivity parameter ϕ values. All chosen values of sensitivity parameter essentially produced an unbiased hazard ratio estimate for effects of treatment B relative to A among those who would comply with either treatment ($S=(1,1)$). This may be discerned from ordinary expectation of high compliance rates with treatment for this subgroup which is likely to reveal true effects of both treatments. For a chosen value of sensitivity parameter ϕ, we also note similarity in the standard errors for the corresponding causal hazard ratio estimates for both homogeneous and heterogeneous cases.

In general, we obtain biased hazard ratio estimates for effects of either treatment (A or B) compared to nothing for all chosen values of sensitivity parameter. Although heterogeneous rates produced an unbiased hazard ratio of effects of treatment B compared to nothing at a moderate value of sensitivity parameter ($\phi=0.5$), the tradeoff was a large SE corresponding to the estimate. Specifically, we observe substantial bias in hazard ratio estimates of efficacy due to compliance with treatment B compared to nothing at higher values of sensitivity parameter ($\phi = 0.8$). We also note relatively large standard errors corresponding to these estimates (and the SE dominates the corresponding RMSE) which may be attributed to the almost 'total' noncompliance phenomenon observed above for this stratum at almost perfect correlation.

Table 8.7: ITT treatment effects for each stratum: homogeneous and heterogeneous rates.

	Homogeneous hazard rates				Heterogeneous hazard rates			
ϕ	$\exp(\bar{\hat{D}})$	$\text{SE}(\hat{D}_i)$	$\text{RMSE}(\hat{D}_i)$	p-value	$\exp(\bar{\hat{D}})$	$\text{SE}(\hat{D}_i)$	$\text{RMSE}(\hat{D}_i)$	p-value
	$\text{THR}^{\dagger}_{(1,1)} = 0.667$ $\left(\frac{\bar{\lambda}_B}{\bar{\lambda}_A}\right)$				$\text{THR}_{(1,1)} = 0.667$ $\left(\frac{\bar{\lambda}_B}{\bar{\lambda}_A}\right)$			
0	0.660	0.356	0.356	0.893	0.660	0.356	0.356	0.894
0.2	0.660	0.339	0.339	0.928	0.660	0.339	0.339	0.928
0.5	0.660	0.323	0.323	0.929	0.660	0.323	0.323	0.929
0.8	0.661	0.303	0.303	0.898	0.661	0.303	0.303	0.898
	$\text{THR}_{(0,1)} = 0.500$ $\left(\frac{\bar{\lambda}_B}{\bar{\lambda}_0}\right)$				$\text{THR}_{(0,1)} = 0.500$ $\left(\frac{\bar{\lambda}_B}{\bar{\lambda}_0}\right)$			
0	0.654	0.493	0.561	<0.001	0.740	0.505	0.639	<0.001
0.2	0.625	1.314	1.326	<0.001	0.708	1.323	1.368	<0.001
0.5	0.341	5.336	5.393	<0.001	0.408	5.544	5.547	0.949
0.8	0.018	17.738	18.045	<0.001	0.041	18.441	18.611	<0.001
	$\text{THR}_{(1,0)} = 0.750$ $\left(\frac{\bar{\lambda}_A}{\bar{\lambda}_0}\right)$				$\text{THR}_{(1,0)} = 0.750$ $\left(\frac{\bar{\lambda}_A}{\bar{\lambda}_0}\right)$			
0	0.668	0.274	0.297	<0.001	0.779	0.263	0.266	<0.001
0.2	0.670	0.286	0.409	<0.001	0.782	0.275	0.278	<0.001
0.5	0.667	0.307	0.421	<0.001	0.778	0.296	0.299	<0.001
0.8	0.664	0.329	0.350	<0.001	0.777	0.317	0.319	<0.001

†True hazard ratio (see Table 8.4)

In contrast, using heterogeneous rates with higher sensitivity of parameters ($\phi = 0.5$ and 0.8) produced small bias of 0.027 (though important) for hazard ratio estimate of efficacy due to compliance with treatment A compared to nothing produced. In general, hazard ratio estimates for treatment B compared to nothing among those who would comply with it if offered produced larger biases and corresponding larger standard errors which increased with increase in sensitivity parameter. The large standard errors may be a manifestation of sparseness due to near 'total' noncompliance as ϕ approaches 1 (perfect correlation).

8.5.3 The Roy et al. (2008) method

Table 8.8 provide a comparison of the causal risk ratios (means of posterior median relative risks) for each stratum at different sensitivity parameter ϕ values under both homogeneous and heterogeneous treatment effect assumptions. For the homogeneous case the resulting biases in the relative risk estimates were all statistically significant for all strata at all ϕ values. Compared to other strata, the causal risk ratio estimate of efficacy due to compliance with treatment B relative to A among those who would comply with either treatment (type 3) consistently produced smaller biases for all ϕ values considered. Specifically the resulting bias in the causal risk ratio estimate was smallest (-0.019) for higher values of $\phi = 0.8$. On the other hand, the causal risk ratio estimates of efficacy due to compliance with only one treatment (A or B) produced larger biases for all values of ϕ.

Table 8.8 also show the causal risk ratio under heterogeneous hazard rates assumption. For mild $\phi = 0.2$ value, the causal risk ratio for those who would comply with treatment B only relative to nothing produced small bias (-0.015), although important. Compared to homogeneous case, we observe substantial increase in standard error values corresponding to causal risk ratio estimates under heterogeneous hazard rate assumption. The causal risk ratio estimate for treatment B compared to treatment A among the highly compliant subgroup (potential compliers with either treatment) resulted in small bias (0.035) at moderate $\phi = 0.5$, although the bias was statistically significant. In general, causal risk ratios estimating compliance with treatment B only compared to nothing were less biased (although statistically significant) under heterogeneous hazard rates assumption than for the homogeneous cases.

Table 8.8: Performance of Roy et al. (2008) method: homogeneous and heterogeneous rates.

	Homogeneous hazard rates				Heterogeneous hazard rates			
ϕ	$\tilde{\hat{\tau}}$ (RR)	SE($\hat{\tau}_{ij}$)	RMSE($\hat{\tau}_{ij}$)	p-value	$\tilde{\hat{\tau}}$ (RR)	SE($\hat{\tau}_{ij}$)	RMSE($\hat{\tau}_{ij}$)	p-value
	$\underline{\text{TRR}^{\ddagger}_{(1,1)} = 0.729}$				$\underline{\text{TRR}_{(1,1)} = 0.729}$			
0	0.688	0.153	0.158	¶	0.837	0.211	0.237	¶
0.2	0.676	0.146	0.155	¶	0.804	0.186	0.201	¶
0.5	0.653	0.128	0.149	¶	0.764	0.151	0.155	¶
0.8	0.710	0.114	0.116	¶	0.797	0.176	0.189	¶
	$\underline{\text{TRR}_{(0,1)} = 0.500}$				$\underline{\text{TRR}_{(0,1)} = 0.500}$			
0	0.476	0.291	0.292	¶	0.530	0.302	0.303	¶
0.2	0.579	0.290	0.301	¶	0.485	0.305	0.305	¶
0.5	0.641	0.292	0.324	¶	0.529	0.319	0.320	¶
0.8	0.593	0.311	0.325	¶	0.573	0.803	0.812	¶
	$\underline{\text{TRR}_{(1,0)} = 0.750}$				$\underline{\text{TRR}_{(1,0)} = 0.750}$			
0	0.995	0.091	0.261	¶	0.913	0.256	0.303	¶
0.2	0.989	0.129	0.272	¶	0.841	0.285	0.299	¶
0.5	0.874	0.246	0.275	¶	0.553	2.300	2.308	¶
0.8	0.567	0.428	0.465	¶	0.713	0.720	0.721	¶

‡True relative risk (see Equation 8.1 and Table 8.4); ¶p-value< 0.001

Worth noting is the fact that results presented in Table 8.8 were obtained after using the same value of sensitivity parameter ϕ for both data generation and analysis. An implication here is the likelihood of the simulations to give an optimistic view of the Roy et al. (2008) method, i.e. the method's defining distributional assumption expressed in ϕ may be over represented by first specifying a ϕ value to link the two marginal compliance model for each treatment arm during data generation and using the same value again for compliance models for each stratum in analysis.

Table 8.9 provide causal risk ratios (means of posterior median relative risks) and corresponding mean 95% credible intervals for each stratum under (a) homogeneous and (b) heterogeneous treatment effect scenarios. A relative risk was obtained as ratio of posterior median probability of death to that of survival in a stratum. In general the median probabilities of death were lower among those who would comply with treatment B only compared to those who would comply with treatment A only under both homogeneous and heterogeneous treatment effect assumptions.

Potential compliers to treatment B only who were allocated to it had lowest estimate ($\hat{\pi}_1$) of risk of death at all ϕ values: the risk increased with increase in ϕ. In contrast, risk of death estimates among those who would comply with treatment A only ($\hat{\pi}_4$) decreased with increase in ϕ values. As expected, the probability of death estimates among potential compliers to either treatment ($\hat{\pi}_2$) was intermediate between the two risks of death among those who would comply with only one treatment (A and B). Overall, for potential compliers to either treatment we observe an increase in the risk of death under heterogeneous treatment effect assumptions compared to the homogeneous case.

At low/moderate values of ϕ under homogeneous treatment effect assumptions, the causal risk ratio estimates of efficacy due to compliance with treatment B relative to A among the subgroup who would comply with either treatment to which they were allocated were the least biased (although statistically significant). The resulting biases increased with increase in ϕ values. The corresponding 95% credible intervals for the causal risk ratio estimates became narrower (smaller) as ϕ increased for this subgroup, indicating gain in precision. On the other hand the causal risk estimates had relatively larger biases under heterogeneous treatment effect assumptions for all values of ϕ. Compared to type 3, the mean 95% credible intervals

Table 8.9: Median probability of death, causal risk ratio (mean of posterior median relative risk) of death (mean 95% CI) for each stratum for various values of $\phi(\pi_Z^S = \Pr[Y=1|S,Z] \equiv$ probability of death in a stratum given treatment allocation): (a) Homogeneous and (b) Heterogeneous rates.

	Comply with both A and B		Comply with B only		Comply with A only	
ϕ	$\hat{\pi}_2(\pi_1^{S=(1,1)})$	$\hat{\pi}_6(\pi_0^{S=(1,1)})$	$\hat{\pi}_1(\pi_1^{S=(0,1)})$	$\hat{\pi}_7(\pi_0^{S=(0,1)})$	$\hat{\pi}_4(\pi_1^{S=(1,0)})$	$\hat{\pi}_5(\pi_0^{S=(1,0)})$
	$\text{TRR}_{(1,1)} = 0.729$		$\text{TRR}_{(0,1)} = 0.500$		$\text{TRR}_{(1,0)} = 0.750$	
(a)						
0	0.161	0.238	0.114	0.243	0.238	0.240
	$\tau_{11} = 0.688$	(0.420, 1.007)	$\tau_{01} = 0.476$	(0.028, 1.010)	$\tau_{10} = 0.995$	(0.808, 1.195)
0.2	0.158	0.238	0.139	0.243	0.233	0.236
	$\tau_{11} = 0.676$	(0.423, 0.993)	$\tau_{01} = 0.579$	(0.039, 1.023)	$\tau_{10} = 0.989$	(0.722, 1.258)
0.5	0.154	0.240	0.153	0.242	0.201	0.235
	$\tau_{11} = 0.653$	(0.429, 0.928)	$\tau_{01} = 0.641$	(0.050, 1.055)	$\tau_{10} = 0.874$	(0.260, 1.240)
0.8	0.168	0.240	0.138	0.237	0.123	0.226
	$\tau_{11} = 0.710$	(0.509, 0.956)	$\tau_{01} = 0.593$	(0.038, 1.106)	$\tau_{10} = 0.567$	(0.029, 1.311)
(b)						
0	0.181	0.221	0.125	0.241	0.199	0.221
	$\tau_{11} = 0.837$	(0.531, 1.355)	$\tau_{01} = 0.530$	(0.030, 1.040)	$\tau_{10} = 0.913$	(0.433, 1.464)
0.2	0.178	0.226	0.114	0.240	0.185	0.224
	$\tau_{11} = 0.804$	(0.524, 1.241)	$\tau_{01} = 0.500$	(0.023, 1.035)	$\tau_{10} = 0.841$	(0.266, 1.393)
0.5	0.172	0.230	0.122	0.237	0.117	0.221
	$\tau_{11} = 0.764$	(0.528, 1.119)	$\tau_{01} = 0.529$	(0.028, 1.095)	$\tau_{10} = 0.553$	(0.030, 1.318)
0.8	0.174	0.223	0.119	0.220	0.146	0.215
	$\tau_{11} = 0.797$	(0.567, 1.250)	$\tau_{01} = 0.573$	(0.029, 1.600)	$\tau_{10} = 0.713$	(0.047, 1.695)

for the causal risk ratio estimates of efficacy due to compliance with one treatment only (A or B) were generally wider. These 95% credible intervals for causal risk ratio estimates became wider with increase in ϕ values.

8.6 Conclusion

The sensitivity parameter ϕ had no effect on the ITT results and also the overall and stratum-specific mean proportion of compliance for either treatment. This is expected given that ITT ignores any nonrandom compliance information introduced in the form of ϕ. While the principal effects among the subgroup who would comply with either treatment were smaller than the ITT estimates under homogeneous treatment effects assumption, the effects were larger than ITT for the heterogenous case. Overall, the principal effects for the subgroup who would comply with treatment B only were smaller than ITT for both homogeneous and heterogeneous cases.

Analysis using the Roy et al. (2008) model produced better results (less bias) than those from ITT. In general, causal risk ratios estimating effects of treatment B relative to A among the subgroup who would comply with either treatment produced the least bias (albeit statistically significant) compared to other strata. The corresponding 95% credible intervals for these estimates became narrower as the sensitivity parameter ϕ values increased. Causal risk ratio estimates (posterior median relative risks) for potential compliers to one treatment only produced larger biases and corresponding wider 95% mean credible intervals which became even wider with increase in ϕ values. Potential compliers to treatment B only approached total noncompliance as ϕ approached perfect correlation. Such a phenomenon may be encountered in situations where treatment B produces unpleasant side effects prompting noncompliance among those randomized to it, i.e. resulting in dominance by type 3 compliance at the expense of type 2 (overall B compliance is sum of types 2 and 3).

In general, the causal risk ratio estimates varied a lot depending on the value of the (un-

known) sensitivity parameter. As a result, the Roy et al. (2008) is likely to produce biased results and should only be recommended if there is sufficient knowledge about the compliance behaviours/correlation between the respective treatment arms. Given such knowledge, subgroup (stratum-specific) analyses may be useful in helping understand the nature of ITT bias by utilizing compliance information which would augment ITT results in efficacy estimation. Choosing non-compliers for a known inferior treatment from the tail of hazard rates' distribution may provide a practical and effective evaluation of of principal effects, i.e. it may be considered more meaningful to associate noncompliance with a lower set of ranked baseline hazard rates and corresponding risk factors.

Overall, although the Roy et al. (2008) method performed well with regard to simulations, it failed when applied to Esprit (Chapter 6) where data may have not satisfied the method's underlying assumptions. This apparent failure of Roy et al. (2008) method as applied to Esprit data may be attributed to its implicit (strong) distributional assumption in which the outcome of interest is assumed independent of the set of baseline covariates predictive of compliance to treatment given compliance type/stratum and treatment allocation. For example, the risk factors are likely to be strongly predictive of outcome independent of their relationship with compliance. In general, despite its central role which ensures identification of principal effect estimates, this defining assumption is untestable and the method's application is heavily dependent on availability and selection of suitable predictors of compliance which is rarely a primary objective in trials and may only be feasible by exploiting data from pilot studies which often require more time and resources.

Chapter 9

Discussion and Conclusions

This chapter provides a recap of the whole thesis. It begins with a review of the principal objectives as spelt out at the onset. This is followed by a section providing a summary of the novelty of the present study followed by discussions of the main results from Esprit data analysis and simulations studies. The next section outlines possible extensions and directions for future work while the final section presents a summary and recommendations from the present work.

9.1 Review of the objectives of present work

In this section we review the objectives of the present work as outlined in section 1.12 and examine whether they have been achieved. The introductory chapter began by putting our research question into context followed by a summary of the motivating data. After introducing the concept of causation in medical research, we reviewed research designs with focus on controlled randomized controlled clinical trials highlighting its key design features, types and limitations as contrasted with observational studies. An outline of reasons of association was followed by a a comprehensive review of the counterfactuals framework of causal modelling, key causal modelling assumptions including conditional exchangeability

(no unmeasured confounding) that enables us to make valid causal inference. We briefly discussed the problem of noncompliance to treatment assignment, showed its similarity with nonresponse and how they can be addressed by counterfactual modelling by stratifying on posttreatment variables. After that we reviewed propensity scores as a method of adjusting for confounding extending its tenets to build prediction models for compliance to treatment allocation. A summary of estimation methods of causal effects preceded a review of the use of principal stratification (Section 1.11.5) as a general framework to address noncompliance by conditioning on a bivariate posttreatment variable noncompliance that induces conditional exchangeability to produce well-defined causal estimands.

After introducing key features of survival data (Chapter 2), we reviewed the two common methods of analysing survival data: proportional hazards and accelerated failure time models and outlined the relationship between them. This was followed by three specialists methods of adjusting for noncompliance in one treatment arm. These methods were classified into two types with the structural proportional hazards method C-Prophet adjusting for all-or-nothing noncompliance while the causal accelerated life model (CALM) and causal hazard ratio adjustment regression model (CHARM) adjusting partial compliance by utilizing time till stoppage of treatment/event. We outlined the key assumptions for each of these methods justifying suitability and discussing possible conditions of violations to Esprit data. Chapter 4 provide results for the analysis of the Esprit data using these specialist methods.

Model selection techniques applied to building prediction models for compliance with treatment assignment is a challenge that has received less attention among researchers evaluating treatment compliance/effects. Plausible predictors of compliance can be used to address identification problem of causal estimands (Jo, 2002a; Little et al., 2009). In Chapter 3 we reviewed model selection methods that can be adopted to produce plausible predictors of compliance. We discussed the classical stepwise model selection techniques and pointed its limitations before exploring the merits of penalized regression techniques. An exposition on model validation enabled us evaluate the performance of selected models using optimism, calibration and discrimination indices.

For two-active treatment trials, the ITT provide a biased estimator for the true hazard ratio even under homogeneous treatment effects assumption. Such ITT results may be augmented by efficacy analysis among subgroups likely to comply with their treatment assignment. We applied the principal stratification method of Roy et al. (2008) for survival data analysis to adjust for noncompliance in two arms (Section 2.7). Specifically, we used plausible baseline covariates predictive of compliance in each arm (Chapter 5) to construct arm-specific compliance models from which we apply the Roy et al. model to develop causal models linking the two marginal models using a pre-specified sensitivity parameter ϕ. We used Bayesian methods (Chapter 6), to obtain principal effects for each principal stratum in terms of causal risk ratios obtained from means of posterior median relative risk and their corresponding 95% credible intervals.

We applied statistically designed simulation studies to evaluate the performance of the specialist methods adjusting for noncompliance in one treatment arm in terms of bias, 95% confidence intervals coverage and RMSE (Chapter 7). Using simulations, we also evaluated the performance of the Roy et al. (2008) model in terms of bias and 95% credible intervals (Chapter 8).

9.2 Novelty of present work

Motivated by the Esprit study whose aim was to ascertain whether or not unopposed oestrogen reduced the risk of further cardiac events in postmenopausal women who survive a first myocardial infarction, this thesis:

1. Performed an in-depth analysis of the Esprit data considering two outcomes (all-cause mortality and myocardial reinfarction or cardiac deaths) adjusting for noncompliance in one (active) treatment arm by applying three specialist methods. The structural proportional hazards method C-Prophet assumed all-or-nothing compliance where a woman was considered compliant with medication if she took HRT tablets allocation

up to a day before experiencing event (all-cause mortality or myocardial reinfarction or cardiac deaths) or end of study, whichever came first. On the other hand both CALM and CHARM methods assumed partial compliance by utilizing the time under treatment before stopping or death/study end (Chapter 4).

2. Applied statistically designed simulation studies to compare the performance of six methods (three naive and three specialist above) in terms of bias, RMSE and 95% confidence interval coverage. The performance of both The C-Prophet and CHARM methods performed was similar despite C-Prophet assuming the (restrictive) all-or-nothing compliance. Overall, the results showed that the CALM method performed consistently best to produce smallest bias and largest 95% CI coverage albeit with relatively large RMSE (Chapter 7).

3. Considered the challenge of building compliance prediction models, an issue which seems not to have received much attention before. Specifically we applied model selection techniques to obtain the 'best' predictors of treatment compliance in both separate arms of treatment. Using penalized regression techniques with same predictors of compliance in each arm, the Least Absolute Shrinkage and Selection Operator (Lasso) selection method produced the best calibrated, most discriminative and least optimistic models predictive of compliance to both HRT tablets and placebo: in predicting compliance to both HRT and placebo, the Lasso models were the least optimistic (2%) and almost perfectly calibrated (slope$=0.93$). However, the better performance by the Lasso method in selecting potential predictors of compliance relative to other methods came at the price of severely shrunk log odd estimates/coefficients (Chapter 5).

4. Examined the effects of placebo compliance in compliance-adjusted analysis which may provide a valid method to scientifically investigate possible placebo effects (Kienle and Kiene, 1997). Specifically we used placebo data to assess whether results change when we adjust for placebo compliance by applying Roy et al. (2008) model for survival analysis (originally developed for binary outcome). We applied principal stratification in a Bayesian framework to adjust for noncompliance in two treatment arms. Applied to

the Esprit data, the results showed that for moderate sensitivity parameter ($\phi = 0.5$), compliance with HRT treatment relative to placebo would reduce risk for all-cause mortality by 43% among women who would comply with either treatment. Compared to placebo, HRT tablets would generally reduce the risk of death and reinfarction among the highly compliant women, i.e. subgroup who would comply with either treatment allocation, for all other values of ϕ (Chapter 6).

5. Evaluated the performance the Roy et al. (2008) method which estimates efficacy in a two-active treatment arms' trial in the presence of nonrandom compliance. Specifically we applied statistically designed simulation studies to compared the performance of the Roy et al. (2008) method in terms of bias and 95% credible intervals under both homogeneous and heterogeneous hazard rate assumptions. The results showed more bias under heterogeneous treatment assumption compared to homogenous treatment effects assumption. Generally principal effects among the potentially highly compliant subgroup were less biased (Chapter 8). Overall, the results were sensitive to the unknown sensitivity parameter and hence the Roy et al. (2008) method should not be recommended unless there is sufficient information about compliance behaviour for each treatment arm, i.e. the method's implicit strong distribution assumption not plausible for Esprit data.

9.3 Discussion of results

9.3.1 Esprit data

Similar to the original results published for the Esprit study (Cherry et al., 2002), ITT estimates showed no statistically significant difference in effect between taking HRT tablets and placebo. However, the treatment suggested beneficial effects in reducing the risk of all-cause mortality by about 20%. On the other hand, the analysis revealed that compared to placebo, HRT treatment had no statistically different effect on the risk of myocardial reinfarction or cardiac death (HR$=0.99$). Analysing the Esprit data using the specialist methods produced

similar results for both C-Prophet and CHARM: a beneficial effect of HRT tablets over placebo to reduce the risk for all-cause mortality by about 35% whereas there was no difference in effects between HRT treatment and placebo on the risk for myocardial reinfarction.

As expected, the (simple) CHARM estimate was sensitive to rate of compliance: substantial risk reduction in the presence of higher proportion of compliance and less risk reduction otherwise. CHARM analysis depends on satisfaction of the extended exclusion restriction assumption. While this assumptions may be considered reasonable for the Esprit data, assuming that a subject's past treatment has no effect on her present risk to death may not hold, for example, for an effective treatment with residual effects or a treatment likely to induce unpleasant side effects hence prone to affect compliance rates. A probable solution to address violation of the former is to adopt a suggestion by White et al. (2004) to use time-series model (with additional structural assumptions) to account for residual treatment effects. A general practical limitation for all the specialist methods is their application only to covariate-free models which may deny them producing more efficient efficacy estimates when (baseline) prognostic variables are accommodated.

The overall average rate of compliance with treatment was moderate at 43% and the compliance rate decreased with time as the study progressed till completion. On average, the mean proportion of compliers to placebo (57%) was higher compared to those complying with HRT tablets allocation (46%). Histories of hysterectomy was a predictor of compliance for both HRT tablet and placebo allocation. Smoking status predicted compliance to HRT tablets allocation but not compliance with placebo allocation. Conversely history of alcohol use was a predictor of compliance to placebo allocation only. The proportion of women who would have the same compliance status under either treatment allocation ($S = (1, 1)$ and $S = (0, 0)$) increased with increase in the sensitivity parameter ϕ: the proportion of those who would comply with either treatment allocation ($S = (1, 1)$) were generally higher than those who would not comply with either allocation ($S = (0, 0)$). In general, the relatively low rate of compliance in Esprit data may make it difficult to generalize the results.

At moderate values of the sensitivity parameter ϕ, HRT tablets showed reduction in risk of death among women who would comply with either treatment to which they were allocated: by assuming conditional independence, principal stratification analysis suggested compliance with HRT treatment relative to placebo would reduce risk of mortality by 42% among those who would comply with either treatment. In general HRT tablets indicated beneficial effects compared to placebo for all other values of ϕ among the subgroup who would comply with either treatment allocation. However, causal risk ratios estimating efficacy of HRT treatment only or placebo only (stratum 1 or 2) had relatively wider mean 95% credible intervals compared to estimates for efficacy HRT over placebo among those who would comply with either treatment (stratum 3). The risk for all-cause mortality for those complying with HRT relative to placebo among potential compliers to either treatment increased with increase in the value of sensitivity parameter ϕ. Conversely, the risk of all-cause mortality decreased with increase in ϕ among those who would comply with HRT treatment only. As expected the risk of death was higher regardless of the level of sensitivity ϕ among those who would comply with placebo only.

A comparison of results from the Roy et al. (2008) method which adjusts for placebo compliance and the specialist methods which only consider compliance with the active treatment showed both C-Prophet and CHARM results were comparable to those from Roy et al. method at high values of sensitivity parameter ($\phi = 0.5$ and 0.8) for all-cause mortality outcome. On the other hand, CALM results were comparable to those for Roy et al. method at low value of the sensitivity parameter ($\phi = 0.2$) for the same outcome (all-cause mortality). In general, results from the Roy et al. were heavily dependent on the (unknown) sensitivity parameter, an indication the results were sensitive to the (strong) implicit distributional assumption which would make them less generalizable to situations where the assumption may be breached, for example, where there is no information about the compliance behaviour for either arm. Overall, the CALM method performed best compared to all the methods including the Roy et al. at all the sensitivity parameter values considered. Specifically, CALM results suggested that compliance with HRT treatment would increase survival time 2.8-fold

(equivalently 50% lower risk) compared to placebo for the all-cause mortality outcome. As a result the CALM method should be recommended for analysing the Esprit data.

Variation in the HRT efficacy estimates from the Roy et al. model may be an indication of difference in compliance behaviour between those allocated to placebo and HRT treatment. By adjusting for noncompliance in both arms, the Roy et al. (2008) method perhaps accounts for potential correlation between compliance behaviours in respective arms through the sensitivity parameter which implicitly makes the results depend on ϕ. The fact that the results vary a lot with ϕ and yet we do not know its value suggests benefits of HRT treatment among those who comply when allocated it, i.e. strong monotonicity assumption (strong correlation in compliance behaviour between the two arms).

9.3.2 Monte Carlo: noncompliance in one arm

The ITT analysis produced bias as expected because of ignoring frailty and assuming constant hazard ratios in the proportional hazards models. Simple regression adjustments of noncompliance produced large bias under random compliance. Similarly, simple regression adjustment of noncompliance performed poorest under non-null case in the presence of non-random compliance: produced larger biases whose magnitude and direction depended on correlation between the risk and probability of noncompliance. These sets of results suggest inadequacy of simple regression adjustments for any form of noncompliance (random and nonrandom).

Under ideal conditions mimicking the null model alone, the C-Prophet and CHARM methods produced the most precise results in terms of RMSE under both random and nonrandom compliances scenarios. But the CALM method under random noncompliance performed consistently 'best' to produce similar and smallest bias for the causal effect in one direction, probably because the method effectively accounts for non-constant hazard rates better compared to C-Prophet and CHARM methods. Despite the (limiting) all-or-nothing assumption, the C-Prophet produced the most precise estimates in terms of RMSE for the nonrandom compliance case albeit with low 95% confidence interval coverage.

Overall the CALM method produced the best coverage rate for the estimates under both random and nonrandom compliance compared to other methods. Probably because of its potential robustness in accounting for partial compliance and good coverage probability, the CALM method should be recommended among the specialist methods adjusting for noncompliance in one treatment arm for the Esprit data.

9.3.3 Monte Carlo: noncompliance in two arms

For two active treatments, the ITT analysis produced biased efficacy estimates even under homogeneous treatment effects assumption. As expected, the overall and mean stratum-specific proportion of compliance for either treatment does not depend on the value of the sensitivity parameter ϕ. The principal effects for the subgroup who would comply with either treatment were less than ITT under homogeneous case but larger than the ITT under heterogeneous case. However, the principal effects for the subgroup who would comply with treatment B only (types 2) were smaller than ITT for both homogeneous and heterogeneous cases.

Analysis using the principal stratification framework (Roy et al. 2008 model) produced less biased results compared to ITT. The least biased causal risk ratio estimates were obtained for the subgroup constituted by those who would comply with either treatment. Although these estimates were statistically significant, their corresponding 95% credible intervals became narrower with increase in the sensitivity parameter ϕ values. Conversely, the causal risk ratio estimates of efficacy due to compliance with one treatment only (types 1 and 2) produced larger biases and corresponding wider 95% credible intervals which became even more wider with increase in ϕ values. Potential compliers to treatment B only approached total noncompliance as ϕ approached perfect correlation ($\phi = 1$). Such a phenomenon may be encountered in situations where treatment B produce unpleasant side effects prompting noncompliance among those randomized to it, i.e. resulting in dominance by type 3 compliance at the expense of type 2 (overall compliance with treatment B is the sum of types 2 and 3).

While Roy et al. (2008) originally used one continuous covariate, our simulation used two

dichotomous covariates (mimicking histories of CVD risks and hysterectomy), hence possibly utilizing more information. Also their (Roy et al., 2008) correlation set-up was deterministic and fixed for the model predicting arm-specific compliance and covariate while we specified our compliance membership stochastically using tail distribution. This set-up was motivated by the practicality that choosing non-compliers for a known inferior treatment from the tail of hazard rates' distribution may provide a realistic representation of underlying risk distribution in a heterogeneous population.

Provided there is sufficient compliance information for each arm, subgroup (stratum-specific) analyses may be useful in helping to understand the nature of ITT bias by utilizing compliance information which may augment ITT results in efficacy estimation. Choosing non-compliers for a known inferior treatment from the tail of hazard rates' distribution may provide a practical and effective evaluation of treatment effects.

9.4 Extensions and directions for future work

The causal accelerated life models produce consistent estimates of causal effects (Robins and Tsiatis, 1991). But since the method ignores the compliance selection mechanism, its flexibility may not readily extend to complicated data from studies with multiple switches. Also a practical limitation of the CALM method as implemented using `strbee` Stata command is the key requirement for the user to provide time for recensoring which may become informative (leading to bias) on a different scale (see Section 7.5). However, some trials may not have a fixed time of follow-up and such a situation may make using end of study as the potential recensoring time restrictive (prone to bias). Recensoring itself is meant to account for no assumption in the functional form of the model and an arbitrary choice of a recensoring time would negate the objective of fewer assumptions to facilitate wider credibility. Also failure to effectively account for the resulting potential informative censoring can have noticeable effects on the results (Siannis et al., 2005). A possible extension would be to use frailty models to account for the correlation between failure and recensoring (Huang and Wolfe, 2002).

Another solution may involve application or modification of marginal structural models for so-called partial exposure regimes (Vansteelandt et al., 2009). We however note that although CALM in general is readily able to handle such complex noncompliance scenarios, `strbee` may present a limitation in implementing them.

Most clinical trial studies include records of baseline covariates. Although not straightforward, of interest would be extending the specialist methods considered for the present work (C-Prophet, CHARM and CALM) to utilize such covariates which may help enhance efficiency during efficacy estimation. For the CALM method for example, identification problem may arise from possible informative censoring on the potential event time-scale which is otherwise non-informative on the observed event time scale. Suitable baseline covariates may be used to address identification problem on survival times in the presence of informative censoring (Ding, 2010). In general, recording good baseline predictors of compliance to treatment allocation at the design stage may be useful for making causal conclusions that goes beyond ITT inference (Goetghebeur and Loeys, 2002). For example, conditioning on baseline covariates in a model may help address any residual confounding that may arise from imbalance between randomized arms. Also extending the methods to utilize potential prognostic (baseline) factors may help improve efficiency in efficacy estimation under nonrandom compliance scenario (Angrist and Pischke, 2009; White and Pocock, 1996). Given the central role of exclusion restriction assumption and potential bias if its violated (Hirano et al., 2000), using pretreatment covariates may help relax the exclusion restriction needed by all the 3 specialist methods considered in the present work because such covariates may allow us to examine the assumption's tenability by assuming an additive effect of treatment assignment (Jo, 2002a; Little et al., 2009). Also conditioning on baseline covariates may help preserve (ensure) noninformative censoring that is crucial for the C-Prophet method.

Using Bayesian techniques in the principal stratification framework to adjust for noncompliance in two treatment arms may be more informative by allowing us to incorporate subject matter knowledge through informative prior distribution. The Bayesian setting is also suitable to model missing data (Daniels and Hogan, 2008; Gelman et al., 2004), hence allowing us

to simultaneously address the prevalent twin challenges of noncompliance and missing data. Specifically, posterior simulation using Gibbs sampling framework combining the likelihood with a prior distribution over the model parameters can be used to impute unobserved compliance types and generate sampling parameters from the conditional posterior distribution for efficient estimation. By treating model parameters as random variables, Bayesian inference can be used to minimize ambiguity in inference (Dunson, 2001; Malakoff, 1999; Ntzoufras, 2009). For example, interpretation of 95% credible intervals is devoid of repeated sampling inference implicit in frequentist interpretation of 95% confidence intervals. The application of principal stratification framework could be extended to data with better compliance information to allow sequential prior updating for improved efficiency in the resulting posterior estimates, i.e. as aptly summed by Gelman et al. (2004) that *today's posterior is tomorrow's prior*. Also given that clinical trials are often conducted to provide optimal treatment decisions, adopting a Bayesian decision-theoretic framework would allow integration of the twin tasks of data analysis and decision making, two issues which are classically treated separately (Brophy and Joseph, 1995; Stangl, 1995, 2000), i.e. Bayesian statistical framework provides an intuitive link between data and decision making (Lancaster, 2004; Lilford and Braunholtz, 2000).

Bayesian methods may also be used to address the problem of partial identifiability when estimating efficacy in the presence of more than one active treatment. This is discernible from the fact that proper prior distributions are guaranteed to produce proper posterior distributions even when the parameters are only partially identifiable in the classical sense (Christensen et al., 2010; Gelman and Hill, 2007; Jackman, 2009; Spiegelhalter, 2004). The present work may be extended to more than two active treatment arms and for general outcome (continuous or discrete) following the Bayesian framework introduced recently by Long et al. (2010) to model the principal (arm-specific) compliances directly, instead of joining them using an associational model (and ϕ), by treating the principal compliance status as missing data. Such an approach would avoid the assumption implicit in the Roy et al. (2008) model that the sensitivity parameter is independent of covariates hence allowing estimation under less stringent assumptions and hence permitting broader credibility.

Comparison of two active treatments often involve an additional treatment administered as a placebo to establish assay sensitivity which is defined as the ability of a study to distinguish between active and inactive treatments (Snapinn, 2000; Temple and Ellenberg, 2000), i.e. a study that successfully demonstrates superiority has simultaneously demonstrated assay sensitivity. As a result such a trial ends up involving a minimum of three arms. The principal stratification framework applied in the present work can be extended to multi-arms trials. For a three-armed trials with binary outcomes, Cheng and Small (2006) applied the principal stratification to derived sharp bounds on causal effects within principal strata using two sets of assumptions obtained by decomposing the monotonicity assumption into two: first assuming no access to treatment among those assigned to control and no treatment switches among those assigned to the two active treatments and second assuming similar compliance behaviour among those assigned to the active treatment. These extensions assumed homogeneous treatment effects, correct compliance information and correct drop-out model. An extension of the present work would involve sensitivity analysis to evaluate the robustness from violations of these assumptions given the ubiquitous heterogeneous nature of treatment effects.

Although reporting point estimates of efficacy measures is the most common practice in medical research, point estimation is often achieved at the price of strong untestable assumptions making the validity of the respective methods dependent on the accuracy of such assumptions (Cai et al., 2008, 2007; Chiba, 2009; Chiba et al., 2007). We can broaden credibility under relaxed assumptions by exhausting (partial) information from available data to construct bounds on treatment effects. For example, Cai et al. (2007) used observed covariate information to derive narrower and more informative nonparametric bounds on treatment effects compared to natural (covariate-free) bounds. However, we note that this method used observed covariates and may be extended to use counterfactuals in the principal stratification framework.

Hormone replacement therapy treatment may be used to improve the quality of life in the last third of women's lives to relieve postmenopausal symptoms, for example, va-

somotor and urogenital atrophy (Pecorelli and Fallo, 1998). Apart from prevalent non-compliance to treatment and possible missing data on outcomes, truncation-by-death may further complicate research where quality of life is a primary outcome of interest. The present work can be extended by applying the model proposed by Mattei and Mealli (2007) based on principal stratification framework for jointly handling data with three complications: treatment noncompliance, missing outcomes following treatment noncompliance and truncation-by-death. Stratifying on the posttreatment variable survival outcome would enable us make causal inference (address identification) in the principal stratum, i.e. the subgroup of patients who would have survived under both treatments. The present work can also be extended by adopting the proposal by Imai (2008)) which addresses identification problem by formulating truncation-by-death as a contaminated data problem. Imai (2008) contends that such a formulation may be flexible enough and can be extended to estimating efficacy in three-arm randomized experiments with noncompliance which would enable us extend the HRT study to comparing two competing active treatments and a placebo hence ability to establish assay efficacy.

9.5 Summary and recommendations

The aims and objectives as outlined at the onset of the present work have been achieved (Section 9.1). This work has contributed novelty to modelling survival data with informative noncompliance in one and two treatment arms. We applied specialist methods for adjusting noncompliance in one treatment arm in data from the Esprit study (Chapter 4) and used statistically designed simulation studies to evaluate their performance in terms of bias, root mean squared error and 95% confidence interval coverage (Chapter 7).

Intention-to-treat is the gold standard in evaluating treatment effectiveness by answering the question '*what is the effect of the treatment as randomly allocated?*.' But in trials with significant proportion of non-adherence, the specialist methods could augment the analysis by adjusting for noncompliance to provide a more realistic measure of the treatment's efficacy.

In general, all the three specialist methods (C-Prophet, CALM and CHARM) were applicable to the Esprit study. Key to C-Prophet method is noncompliance classification and exclusion restriction assumption. Given the limited compliance information, noncompliance for the Esprit study may be meaningfully classified as all-or-nothing. However, we note this as a less desirable method although it performed well in simulations. And the exclusion restriction assumption (i.e. no effect of assignment to active arm on survival in the subgroup of non-compliers) may be considered plausible for the C-Prophet method because no switches were allowed from the placebo to HRT arms. The crucial extended exclusion restriction assumption (i.e. randomized allocation has no effect on any survivors to the start of an interval who would stop treatment by the end of that interval) for the CHARM method may be considered plausible for the Esprit study if we assume no carryover effects and probably short monthly evaluations. Finally the exclusion restriction assumption is plausible for CALM method (the baseline prognosis independent of randomized allocation) given the strict inclusion and exclusion criteria for the study which would minimize chances of recruiting patients with prognosis capable of influencing/breaching treatment assignment. Moreover, recensoring using the study duration (24 months) would minimize possible information loss. Although both C-Prophet and CHARM performed equally well, CHARM provided more flexibility in adjusting for possible crossovers from active to placebo arms. Overall, we recommend the CALM method which apart from allowing twin adjustments of crossovers (partial noncompliance) and recensoring also produced 'least' biased treatment effects with 'best' 95% confidence intervals coverage albeit with a relatively large root mean squared error.

Bradley Efron may have correctly identified model selection in regression as one of the most important problems in statistics (Hesterberg et al., 2008) towards the end of last Century. Hitherto this challenge has not been addressed in the context of prediction of compliance to treatment allocation in causal modelling (covered in Chapters 3 and 5). Most authors in the compliance literature say very little about this topic. A record of plausible predictors of compliance can be used to effectively address identification problem of causal estimands by reducing bias and weakening implicit assumptions (Little et al., 2009). From a clinical

perspective, knowledge about predictors of compliance may be a valuable tool to inform treatment decisions. As a result, there is need to adopt existing model selection methods for accurate prediction (of compliance). After model selection, there is further need to use suitable validation measures (e.g. optimism, calibration and discrimination indices) to evaluate performance of selected models. With (many) recorded baseline covariates, using penalized regression techniques is recommended for building compliance prediction models. Advantages of classical model selection may be transferable to the Roy et al. (2008) method which adjusts for noncompliance in two-active treatment arms through use of respective optimal marginal compliance models.

The presence of two active treatments complicates efficacy estimation due to possible noncompliance in both arms. Here the ITT provide a biased estimator for the true hazard ratio even under homogeneous treatment effects assumption. Applying the Roy et al. (2008) model from the present work (Section 2.7 and Chapter 6) may be suitable in utilizing more baseline information to model arm-specific compliance models to develop causal models linking the two marginal models. The resulting principal effects provides efficacy estimates for the different subgroups defined by compliance types. In evaluating performance of the model, simulation studies (Chapter 8) revealed less biased results for the potentially most compliant stratum, i.e. the subgroup that would comply with either treatment. Overall, the method's performance was satisfactory but the results were heavily dependent on the level of sensitivity parameters and hence may not be recommended in the presence of known heterogeneous treatment effects which produced large bias and wider corresponding 95% credible intervals. It may only be recommended in the presence of sufficient information about compliance behaviours in respective arms.

Application of the Roy et al. (2008) method is premised on the plausibility of a defining assumption that potential outcome is independent of the set of covariates predicting compliance for a given stratum. This assumption may not be plausible for the Esprit data especially with regard to history of hysterectomy and cerebrovascular risks which potentially have a higher likelihood to be association with treatment compliance leading to possible efficacy. For

example, while the unpleasant experience of bleeding may affect treatment compliance negatively, those with history of CVD risks may comply with their treatment allocation with a hope to derive any protective benefits. Also the fact that resulting principal effects depended on the choice of covariates predicting compliance may be a reason of failure of the Roy et al. (2008) method as applied to the Esprit data given inadequate compliance information/data.

Results from our simulation studies (Chapter 8) were comparable to those of Roy et al. (2008): less biased treatment effects for the potentially highly compliant subgroup, i.e. women who would comply with either treatment. The difference in simulations set-up for compliance was that while Roy et al. originally set their simulations to have equal compliance probability in each treatment arm, we had a higher proportion of compliance for treatment A compared to treatment B in order to maintain U(x)$= \min\{1, \frac{\hat{\mu}_1(x)}{\hat{\mu}_0(x)}\} < 1$. In addition, we stochastically set the correlation between the two compliance using tail distributions while the original set-up by Roy et al. (2008) was deterministic in nature. As a result our set-up may be more representative of a heterogenous sample and suitable for survival data, i.e. choosing non-compliers for a known inferior treatment from the tail of hazard rates' distribution to provide a practical and effective evaluation of treatment effects.

All the methods considered in the present work were applied to the Esprit study which did not have very high quality compliance data given that compliance was not an original primary objective, i.e. less compliance information recorded. This may limit generalization of the results obtained herein. Since the end of Esprit study, more data has been collected which may be used for more informative analysis that would reflect true/better effects of HRT treatment among different subgroups (principal strata). Also duration of treatment is an important phenomenon in methods which accounts for partial compliance. The specialist methods considered here assumed that the effects of treatment ceased immediately a subject stopped medication. This may not be reflective of the true effects for medications with known residual effects (Nagelkerke et al., 2000). But following suggestion by White et al. (2004), we may incorporate time-series-like models that would extend the methods to account for (latent) lagged treatment effects.

Bibliography

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8:907–925.

Aalen, O. O. (1998). Frailty models. In Everitt, B. S. and Dunn, G., editors, *Statistical Analysis of Medical Data: New Developments*, pages 59–72. Arnold, London.

Aalen, O. O., Borgan, O., and Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Statistics for Biology and Health. Springer, New York.

Aalen, O. O. and Frigessi, A. (2007). What can statistics contribute to a causal Understanding? *Scandinavian Journal of Statistics*, pages 155–168.

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:97–106.

Albert, J. M. and DeMets, D. L. (1994). On a model-based approach to estimating efficacy in clinical trials. *Statistics in Medicine*, 13:2323–2335.

Altman, D. G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall, London.

Altman, D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, 19(4):453–473.

Anderson, W. N. (2005). Statistical techniques for validating logistic regression models. *The Annals of Thoracic Surgery*, 80:1169.

Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology*, 2:23–44.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, New Jersey.

Arjas, E. (2001). Causal analysis and statistics: A social sciences perspective. *European Sociological Review*, 17(1):59–64.

Arjas, E. and Parner, J. (2004). Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics*, 31(2):171–187.

Ashby, D., Smyth, R. L., and Brown, P. J. (1998). Statistical issues in pharmacoepidemiological case-control studies. *Statistics in Medicine*, 17:1839–1850.

Austin, P. C. and Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, 57:1138–1146.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66:411–421.

Baker, S. G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost-effectiveness of a cancer screening program. *Journal of the American Statistical Association*, 93:929–934.

Baker, S. G. and Kramer, B. S. (2005). Simple maximum likelihood estimates of efficacy in randomized trials and before-and-after studies, with implications for meta-analysis. *Statical Methods in Medical Research*, 14:349–367.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

Barnabei, V. M., Aragaki, B. B. C. A. K., Nygaard, I., Williams, R. S., McGovern, P. G., Young, R. L., Wells, E. C., O'Sullivan, M. J., Chen, B., Schenken, R., and Johnson S. R; Women's Health Initiative Investigators (2005). Menopausal symptoms and treatment-related effects of estrogen and progestin in the women's health initiative. *Obstetrics and Gynecology*, 105 (5 Pt 1):1063–1073.

Barnard, J., Frangakis, C. E., Hill, J. L., and Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in new york city. *Journal of the American Statistical Association*, 98(462):299–311.

Barnard, J. and Meng, X.-L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, 8:17–36.

Becque, T. and White, I. R. (2008). Regaining power lost by non-compliance via full probability modelling. *Statistics in Medicine*, 27:5640–5663.

Belin, T. R. and Normand, S.-L. T. (2009). The role of ANCOVA in analyzing experimental data. *Psychiatric Annals*, 39(7):753–759.

Bellamy, S. L., Lin, J. Y., and Ten Have, T. R. (2007). An introduction to causal modelling in clinical trials. *Clinical Trials*, 4:58–73.

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology*, 1:417–433.

Blossfeld, H.-P. and Rohwer, G. (1997). Causal inference, time and observation plans in the social sciences. *Quality and Quantity*, 31:361–384.

Boslaugh, S. and Watters, P. A. (2008). *Statistics in a Nutshell*. O'Reilly Media Inc., Sebastopol, CA.

Bray, B. C., Almirall, D., Zimmerman, R. S., Lynam, D., and Murphy, S. A. (2006). Assessing the total effect of time-varying predictors in prevention research. *Journal of Prevention Science*, 7(1):1–17.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383.

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91(433):14–28.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research*, volume I - The Analysis of Case-Control Studies (IARC Scientific Publications No. 32). International Agency for Reasearch on Cancer, Lyon.

Brittain, E. and Lin, D. (2005). A comparison of intent-to-treat and per-protocol results in antibiotic non-inferiority trials. *Statistics in Medicine*, 24:1–10.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163:1149–1156.

Brophy, J. M. and Joseph, L. (1995). Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *Journal of the American Medical Association*, 273(11):871–875.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25:4279–4292.

Cai, Z., Kuroki, M., Pearl, J., and Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64:695–701.

Cai, Z., Kuroki, M., and Sato, T. (2007). Non-parametric bounds on treatment effects with non-compliance by covariate adjustment. *Statistics in Medicine*, 26:3188–3204.

Campbell, M. J., Machin, D., and Walters, S. J. (2007). *Medical Statistics: A Textbook for the Health Sciences*. Wiley, Chichester, 4th. edition.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Duxbury Press, CA, 2nd. edition.

Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology*, 158:280–287.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419–466.

Checkoway, H., Pearce, N., and Kriebel, D. (2004). *Research Methods in Occupational Epidemiology*. Oxford University Press, Oxford, 2nd. edition.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society*, B68(Part 5):815–836.

Cheng, J., Small, D. S., Tan, Z., and Ten Have, T. R. (2009). Efficient nonparametric estimation of causal effects in randomized trials with noncompliance. *Biometrika*, 96:19–36.

Cherry, N., Gilmour, K., Hannaford, P., Heagerty, A., Khan, M. A., Kitchener, H., McNamee, R., Elstein, M., Kay, C., Seif, M., and Buckley, H. (2002). Oestrogen therapy for prevention of reinfarction in postmenopausal women: A randomised placebo controlled trial. *Lancet*, 360:2001–2008.

Chiba, Y. (2009). Bounds on causal effects in randomized trials with noncompliance under monotonicity assumptions about covariates. *Statistics in Medicine*, 28:3249–3259.

Chiba, Y., Sato, T., and Greenland, S. (2007). Bounds on potential risks and causal risk differennce under assumptions about confounding parameters. *Statistics in Medicine*, 26:5125–5135.

Chilvers, C. E., Knibb, R. C., Armstrong, S. J., Woods, K. L., and Logan, R. F. (2003). Post menopausal hormone replacement therapy and risk of acute myocardial infarction a case control study of women in the East Midlands, UK. *European Heart Journal*, 24:2197–2205.

Cho, L. and Mukherjee, D. (2005). Hormone replacement therapy and secondary cardiovascular prevention: A meta-analysis of randomized trials. *Cardiology*, 104:143–147.

Chow, S.-C. and Liu, J.-P. (2004). *Design and Analysis of Clinical Trials: Concepts and Methodologies*. Wiley, New York, 2nd. edition.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2010). *Bayesian Ideas and Data Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.

Clarke, B., Fokoué, E., and Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer Series in Statistics. Springer, New York.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151.

Cleves, M., Gould, W., Gutierrez, R., and Marchenko, Y. (2008). *An Introduction to Survival Analysis Using Stata*. Stata Press, Texas, 2nd. edition.

Cochran, W. G. (1965). The planning of observational studies of human populattions (with Discussion). *Journal of the Royal Statistical Society, Series A*, 128:134–155.

Cochrane, A. L. (1972). *Effectiveness and Efficacy: Random Reflections on Health Services*. Nuffield Provincial Hospitals Trust, London.

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference: A definition or an assumption? *Epidemiology*, 20(1):3–5.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

Collet, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd. edition.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6:330–351.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series B*, 45:311–354.

Cox, D. R. (1958a). *The Planning of Experiemnts*. Wiley, New York.

Cox, D. R. (1958b). Two further applications of a model for binary regression. *Biometrika*, 45:562–565.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, Series B 34:187–220.

Cox, D. R. and Oakes, D. (1984). *The Analysis of Survival Data*. Chapman & Hall, London.

Cuzick, J., Edwards, R., and Segnan, N. (1997). Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in Medicine*, 16:1017–1029.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17:2265–2281.

D'Agostino Sr, R. B., Massaro, J. M., and Sullivan, L. M. (2003). Non-inferiority trials: design concepts and issues the encounters of academic consultants in statistics. *Statistics in Medicine*, 22:169–186.

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall/CRC, Boca Raton, FL.

Datta, M. (1993). You cannot exclude the explanation you have not considered. *Lancet*, 342(8867):345–347.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, UK.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B (Statisical Methodology)*, 41(1):1–31.

Dawid, A. P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association*, 95:407–448.

Dawid, A. P. (2004). Probability,causality and the empirical world: A Bayes–de Finetti-Popper-Borel synthesis. *Statistical Science*, 19(1):44–57.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Desousa, C. and Murrells, T. (2005). The use of Splus software to analyse event history data: an application to the early career promotion of nurses in UK. *Quality & Quantity*, 39:453–465.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, 2nd. edition.

Ding, A. A. (2010). Identifiability conditions for covariate effects model on survival times under informative censoring. *Statistics & Probability Letters*, 80(11-12):911–915.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms*. Oxford University Press, Oxford.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley, New York, 3rd. edition.

Driscoll, M. (1977). The ten commandments of statistical inference. *American Mathematical Monthly*, 84:628.

Dunn, G. and Everitt, B. S. (1995). *Clinical Biostatistics: An Introduction to Evidence-Based Medicine*. Edward Arnold, London.

Dunn, G. and Goetghebeur, E. T. (2005). Editorial: Analysing compliance in clinical trials. *Statistical Methods in Medical Research*, 14:325–326.

Dunn, G., Maracy, M., Dowrick, C., Ayuso-Mateos, J. L., Dalgard, O. S., Page, H., Lehtinen, V., Casey, P., Wilkinson, C., Vázquez-Barquero, J. L., and Wilkinson, G. on behalf of the ODIN group (2003). Estimating psychological treatment effects from a randomised controlled trial with both non-compliance and loss to follow-up. *British Journal of Psychiatry*, 18(3):323–331.

Dunson, D. B. (2001). Practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, 153(12):1222–1226.

Eells, E. (1991). *Probabilistic Causality*. Cambridge University Press, Cambridge.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Efron, B. (1983). Estimating the error rate of prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331.

Efron, B. (1986). How biased is the apparent error rate of prediction rule? *Journal of the American Statistical Association*, 81:461–470.

Efron, B. (1998). Forward in "Special issue on analyzing non-compliance in clinical trials". *Statistics in Medicine*, 17:249–250.

Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials (with discussion). *Journal of the American Statistical Association*, 86:9–26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Egleston, B. L., Cropsey, K. L., Lazev, A. B., and Heckman, C. J. (2010). A tutorial on principal stratification-based sensitivity analysis: application to smoking cessation studies. *Clinical Trials*, 7:286–298.

Enders, C. K. (2010). *Applied Missing Data Analysis*. Methodology in the Social Sciences. The Guilford Press, New York.

Everitt, B. S. and Pickles, A. (1999). *Statistical Aspects of the Design and Analysis of Clinical Trials*. Imperial College Press, London.

Feudjo-Tepie, M., Frost, C., Wang, D., and Bakhai, A. (2006). Missing data. In Wang, D. and Bakhai, A., editors, *Clinical Trials: A Practical Guide to Design, Analysis, and Reporting*, pages 339–351. Remedica, London, UK.

Fewell, Z., Hernán, M. A., Wolfe, F., Tilling, K., Choi, H., and Sterne, J. A. C. (2004). Controlling for time-dependent confounding using marginal structural models. *The Stata Journal*, 4(4):402–420.

Fischer, K., Goetghebeur, E., Vrijens, B., and White, I. R. (2011). A structural mean model to allow for noncompliance in a randomized trial comparing 2 active treatments. *Biostatistics*, 12(2):247–257.

Fischer-Lapp, K. and Goetghebeur, E. (1999). Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials*, 20:531–546.

Fischer-Lapp, K. and Goetghebeur, E. (2004). Structural mean effects of noncompliance: estimating interaction with baseline prognosis and selection effects. *Journal of the American Statistical Association*, 99(468):918–928.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of Ministry of Agriculture*, 33:503–513.

Follmann, D. A. (2000). On the effect of treatment among would-be treatment compliers: an analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*, 95(452):1101–1109.

Frangakis, C. E. (2004). Principal stratification. In Gelman, A. and Meng, X.-L., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 97–108. Wiley, New York.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D., and Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association*, 99(465):239–249.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86:365–379.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58:21–29.

Frangakis, C. E., Rubin, D. B., An, M.-W., and MacKenzie, E. (2007). Principal stratification designs to estimate input data missing due to death. *Biometrics*, 63:641–662.

Frangakis, C. E., Rubin, D. B., and Zhou, X.-H. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics*, 3(2):147–164.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135.

Freedman, D. A. (2006). Statistical models for causation: What inferential leverage do they provide? *Evaluation Review*, 30(691):692–713.

Freedman, D. A., Collier, D., Sekhon, J. S., and Stark, P. B. (2010). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge University Press, Cambridge, NY.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (1998). *Fundamentals of Clinical Trials*. Springer, New York, 3rd. edition.

Frölich, M. (2003). *Programme Evaluation and Treatment Choice*. Lecture Notes in Economics and Mathematical Systems (524). Springer-Verlag, Berlin.

Furnival, G. M. and Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, 16:499–511.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd. edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, Cambridge.

Geng, Z. (2003). Causal inference. In Lu, Y. and Fang, J.-Q., editors, *Advanced Medical Statistics*, chapter 16, pages 777–812. World Scientific Publishing, River Edge, NJ.

Goetghebeur, E. J. T. and Fischer-Lapp, K. (1997). The effects of treatment compliance in a placebo-controlled trial: regression with unpaired data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46:351–364.

Goetghebeur, E. J. T. and Loeys, T. (2002). Beyond intention to treat. *Epidemiologic Reviews*, 24:85–90.

Goetghebeur, E. J. T. and Loeys, T. (2003). A sensitivity analysis for causal parameters in structural proportional hazards models. *Sort*, 27(1):31–40.

Goetghebeur, E. J. T. and Shapiro, S. H. (1996). Analysing non-compliance in clinical trials: Ethical imperative or mission impossible? *Statistics in Medicine*, 15:2813–2826.

Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies*. Academic Press, New York.

Grady, D., Herrington, D., Bittner, V., Blumenthal, R., Davidson, M., Hlatky, M., Hsia, J., Hulley, S., Herd, A., Khan, S., Newby, L. K., Waters, D., Vittinghoff, E., Wenger, N., and HERS Research Group (2002). Cardiovascular disease outcomes during 6.8 years of hormone therapy: Heart and estrogen/progestin replacement study follow-up (HERS II). *Journal of the American Medical Asssociation*, 288(1):49–57.

Grady, D., Rubin, S. M., Petitti, D. B., Fox, C. S., Black, D., Ettinger, B., Ernster, V. L., and Cummings, S. R. (1992). Hormone therapy to prevent disease and prolong life in postmenopausal women. *Annals of Internal Medicine*, 117(12):1016–1037.

Greenland, S. (2004). An overview of methods for causal inference from observational studies. In Gelman, A. and Meng, X.-L., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 3–13. Wiley, New York.

Greenland, S., Lanes, S., and Jara, M. (2008). Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. *Clinical Trials*, 5:5–13.

Greenland, S. and Morgenstern, H. (2001). Confounding in health research. *Annual Reviews of Public Health*, 22:189–212.

Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiologic confounding. *International Journal of Epidemiology*, 15:413–419.

Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

Grodstein, F. and Stampfer, M. (1995). The epidemiology of coronary heart disease and estrogen replacement in postmenopausal women. *Progress in Cardiovascular Diseases*, 38(3):199–210.

Hackshaw, A. (2009). *A Concise Guide to Clinical Trials*. BMJ Books. Wiley-Blackwell, Chichester, UK.

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology*, 30:507–544.

Hammal, D. M. and Bell, C. L. (2002). Confounding and bias in epidemiological investigations. *Paediatric Haematology and Oncology*, 19:375–381.

Harrell, F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer Verlag, New York.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag, New York, 2nd. edition.

Heckman, J. J. (1996). Comment on "Identification of causal effects using instrumental variables" by J. D. Angrist, G. W. Imbens, and D. B. Rubin. *Journal of the American Statistical Association*, 91:459–462.

Heckman, J. J. (2000). Causal parameters and policy analysis in economics: A twentieth Century retrospective. *Quarterly Journal of Economics*, 115:45–97.

Heckman, J. J. and Hotz, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training (with Discussion). *Journal of the American Statistical Association*, 84:862–880.

Heckman, J. J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In Heckman, J. J. and Singer, B., editors, *Longitudinal Analysis of Labor Market Data*, pages 156–245. Cambridge University Press, Cambridge, UK.

Heckman, J. J. and Robb, R. (1988). The value of longitudinal data for solving the problem of selection bias in evaluation the impact of treatment on outcomes. In Duncan, G. and Kalton, G., editors, *Panel Surveys*, pages 512–538. Wiley, New York.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432.

Heitjan, D. F. (1999). Causal inference in a clinical trial: A comparative example. *Controlled Clinical Trials*, 20:309–318.

Hennekens, C. H. and Buring, J. E. (1987). Evaluating the role of confounding. In Mayrent, S. L., editor, *Epidemiology in Medicine*. Little, Brown and Company, Boston, MA.

Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health*, 58:265–271.

Hernán, M. A., Cole, S. R., Margolick, J., Cohen, M., and Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*, 14:477–491.

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60:578–586.

Hertogh, E. M., Schuit, A. J., Peeters, P. H. M., and Monninkhof, E. M. (2010). Noncompliance in lifestyle intervention studies: the instrumental variable method provides insight into the bias. *Journal of Clinical Epidemiology*, 63:900–906.

Hesterberg, T., Choi, N. H., Meier, L., and Fraley, C. (2008). Least angle and ℓ_1 penalized regression: A review. *Statistics Surveys*, 2:61–93.

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300.

Hill, K. (1996). The demography of menopause. *Maturitas*, 23:113–27.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, 2:259–278.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistcs*, 1(1):69–88.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Höfler, M. (2005). Causal inference based on counterfactuals. *BMC Medical Research Methodology*, 5(28).

Höfler, M., Pfister, H., Lieb, R., and Wittchen, H.-U. (2005). The use of weights to account for non-response and drop-out. *Society of Psychiatry and Psychiatric Epidemiology*, 40:291–299.

Holland, P. W. (1986). Statistics and causal inference (with Discussion). *Journal of the American Statistical Association*, 81(396):945–970.

Horiuchi, Y., Imai, K., and Taniguchi, N. (2007). Designing and analyzing randomized experiments: Application to a Japanese election survey experiment. *American Journal of Political Science*, 51(3):669–687.

Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics*, 58:510–520.

Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., and Hearst, N. (2001). *Designing Clinical Research: An Epidemiologic Approach*. Lippincott Williams & Wilkins, Philadelphia, 2nd. edition.

Hume, D. (1739). *A Treatise of Human Nature*. John Noon, London.

Hume, D. (1748). *An Inquiry Concerning Human Understanding*. Open Court Press, Lasalle.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Imai, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statistics & Probability Letters*, 78:144–149.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrics*, 62(2):467–475.

Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with non-compliance. *The Annals of Statistics*, 25:305–327.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer, New York.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley, Chichester, UK.

Jadad, A. R. and Enkin, M. W. (2007). *Randomised Controlled Trials: Questions, Answers and Musings*. BMJ Books. Blackwell Publishing, London, 2nd. edition.

Jepsen, P., Johnsen, S. P., Gillman, M. W., and Sørensen, H. T. (2004). Interpretation of observational studies. *Heart*, 90:956–960.

Jewell, N. P. (2003). *Statistics for Epidemiology*. Chapman & Hall/CRC, Boca Raton.

Jin, H., Barnard, J., and Rubin, D. B. (2010). A modified general location model for non-compliance with missing data: Revisiting the New York City School Choice Scholarship Program using principal stratification. *Journal of Educational and Behavioral Statistics*, 35(2):154–173.

Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481):101–111.

Jo, B. (2002a). Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. *Statistics in Medicine*, 21:3161–3181.

Jo, B. (2002b). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods*, 7:1124–1129.

Jo, B., Ginexi, E. M., and Ialongo, N. S. (2010). Handling missing data in randomized experiments with noncompliance. *Prevention Science*, 1:1–13.

Jo, B. and Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, 28:2857–2875.

Joffe, M. M. and Brensinger, C. (2003). Weighting in instrumental variables and G-estimation. *Statistics in Medicine*, 22(8):1285–1303.

Julious, S. A. and Mullee, M. A. (1994). Confounding and Simpson's paradox. *British Medical Journal*, 309:1480–1481.

Kapadia, A. S., Chan, W., and Moyé, L. A. (2005). *Mathematical Statistics with Applications*, volume 176 of *Statistics, Textbooks and Monographs*. Chapman & Hall/CRC Press, Boca Raton, FL.

Keiding, N., Andersen, P. K., and Klein, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine*, 16:215–224.

Kienle, G. S. and Kiene, H. (1997). The powerful placebo effect: Fact or fiction? *Journal of Clinical Epidemiology*, 50(12):1311–1318.

Kim, L. G. and White, I. R. (2004). Compliance-adjusted intervention effects in survival data. *The Stata Journal*, 4(3):257–264.

Kim, M. Y. (2010). Using the instrumental variables estimator to analyze noninferiority trials with noncompliance. *Journal of Biopharmaceutical Statistics*, 20(4):745–758.

Kleijnen, J. P. C. (1974). A note on sampling two correlated variables. *Simulation*, pages 45–46.

Klungel, O. H., Martens, E. P., Psaty, B. M., Grobbee, D. E., Sullivan, S. D., Stricker, B. H. C., Leufkens, H. G. M., and de Boer, A. (2004). Methods to assess intended effects of drug treatment in observational studies are reviewed. *Journal of Clinical Epidemiology*, 57:1223–1231.

Kluve, J. (2004). On the role of counterfactuals in inferering causal effects. *Foundations of Science*, 9:65–101.

Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378.

Korhonen, P. A., Laird, N. M., and Palmgren, J. (1999). Correcting for non-compliance in randomized trials: an application to the ATB study. *Statistics in Medicine*, 18:2879–2897.

Kravitz, R. L., Duan, N., and Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4):661–687.

Kuper, H. and Gilbert, C. (2005). Epidemiology for ophthalmologists: an introduction to concepts, study designs, and interpreting findings. *British Journal of Ophthalmology*, 89:378–384.

Kurland, B. F., Johnson, L. L., Egleston, B. L., and Diehr, P. H. (2009). Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. *Statistical Science*, 24(2):211–222.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, 5(2):369–412.

Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*, 21:167–189.

Laine, C., De Angelis, C., Delamothe, T., Drazen, J. M., Frizelle, F. A., Haug, C., Hébert, P. C., Horton, R., Kotzin, S., Marusic, A., Sahni, P., Schroeder, T. V., Sox, H. C., van der Weyden, M. B., and Verheugt, F. W. (2007). Clinical trial registration: looking back and moving ahead. *Annals of Internal Medicine*, 147(4):275–277.

Lancaster, T. (2004). *Introduction to Modern Bayesian Econometrics*. Wiley-Blackwell, New York.

Last, J. M. (2001). *A Dictionary of Epidemiology*. Oxford University Press, New York, 4th. edition.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lauritzen, S. L. (2001). Causal inference from graphical models. In Barndorff-Nielsen, O. E., Cox, D. R., and Klüppelberg, C., editors, *Complex Stochastic Systems*, pages 63–107. Chapman and Hall/CRC Press, Baton Raton.

Lee, E. T. and Go, O. T. (1997). Survival analysis in public health research. *Annual Review of Public Health*, 18:105–134.

Lee, Y. J., Ellenberg, J. H., Hirtz, D. G., and Ne, K. B. (1991). Analysis of clinical trials by treatment actually received: Is it really an option? *Statistics in Medicine*, 10:1595–1605.

Lewis, D. (1973). Causation. *Journal of Philosophy*, 70:556–567.

Lilford, R. J. and Braunholtz, D. (2000). Who's afraid of Thomas Bayes? *Journal of Epidemiology and Community Health*, 54:731–739.

Lin, D. Y. and Ying, Z. (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. *Journal of Statistical Planning and Inference*, 44(1):47–63.

Linden, A. and Adams, J. L. (2006). Evaluating disease management programme effectiveness: and introduction to instrumental variables. *Journal of Evaluation in Clinical Practice*, 12(2):148–154.

Lindley, D. V. and Singpurwalla, N. D. (2002). On exchangeable, causal and cascading failures. *Statistical Science*, 17(2):209–219.

Lindsey, J. K. (2004). *Statistical Analysis of Stochastic Processes in Time*. Cambridge University Press, Cambridge.

Little, R. J. A., Long, Q., and Lin, X. (2009). A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics*, 65:640–649.

Little, R. J. A. and Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21:121–145.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. Wiley, New York, 2nd. edition.

Loeys, T. and Goetghebeur, E. J. T. (2003). A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics*, 59:100–105.

Loeys, T., Goetghebeur, E. J. T., and Vandebosch, A. (2005). Causal proportional hazards models and time-constant exposure in randomized clinical trials. *Lifetime Data Analysis*, 11:435–449.

Lok, J. J. (2008). Statistical modeling of causal effects in continuous time. *The Annals of Statistics*, 36(3):1464–1507.

Lok, J. J., Gill, R., van der Vaart, A., and Robins, J. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Statistica Neerlandica*, 58(3):271–295.

Long, Q., Little, R. J. A., and Lin, X. (2010). Estimating causal effects in trials involving multitreatment arms subject to non-compliance: a Bayesian framework. *Journal of the Royal Statistical Society, C*, 59(3):513–531.

Longford, N. (2008). *Studying Human Populations*. Springer, New York.

Luellen, J. K., Shadish, W. R., and Clark, M. H. (2005). Propensity scores: an introduction and experimental test. *Evaluation Review*, 29(6):530–558.

Lui, K.-J. (2011). *Binary Data Analysis of Randomized Clinical Trials with Noncompliance*. Statistics in Practice. Wiley, New York.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley, New York.

Malakoff, D. (1999). Bayes offers a 'new' way to make sense of numbers. *Science*, 286(5444):1460–1464.

Maldonado, G. and Greenland, S. (2002). Estimating causal effects. *International Journal of Epidemiology*, 31:422–429.

Manski, C. F. (1990). Non-parametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319–323.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer, New York.

Manski, C. F. (2007). *Identification for Prediction and Decision*. Harvard University Press, Cambridge, MA.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, 12:621–625.

Marcus, S. M. and Gibbons, R. D. (2001). Estimating the efficacy of receiving treatment in randomized clinical trials with noncompliance. *Health Services & Outcomes Research Methodology*, 2:247–258.

Mark, S. D. and Robins, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statistics in Medicine*, 12:1605–1628.

Marubini, E. and Valsecchi, M. G. (1995). *Analysising Survival Data from Clinical Trials and Observational Studies*. Wiley, Chichester.

Mathews, J. N. S. (2006). *Introduction to Randomized Controlled Clinical Trials*. Chapman & Hall/CRC, Boca Raton, FL, 2nd. edition.

Mattei, A. and Mealli, F. (2007). Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics*, 63:437–446.

McNamee, R. (2003). Confounding and confounders. *Journal of Occupational and Environmental Medicine*, 60:227–234.

McNamee, R. (2009). Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. *Statistics in Medicine*, 28:2639–2652.

Mealli, F., Imbens, G. W., Ferro, S., and Biggeri, A. (2004). Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics*, 5(2):207–222.

Mealli, F. and Rubin, D. B. (2002). Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services & Outcomes Research Methodology*, 3:225–232.

Mevik, B.-H. and Wehrens, R. (2007). The `pls` package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(2):1–23.

Miller, A. J. (2002). *Subset Selection in Regression*. Chapman & Hall, Boca Raton, Florida, 2nd. edition.

Miller, M. E., Langefeld, C. D., Tierney, W. M., Hui, S. L., and McDonald, C. L. (1993). Validation of probabilistic predictions. *Medical Decision Making*, 13:49–58.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Trials*. Statistics in Practice. Wiley, Chichester, UK.

Moons, K. G. M., Donders, A. R. T., Steyerberg, E. W., and Harrell, Jr., F. E. (2004). Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: A clinical example. *Journal of Clinical Epidemiology*, 57:1262–1270.

Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. *Sociology of Education*, 74:341–374.

Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York.

Nagelkerke, N. J. D., Fidler, V., Bernsen, R., and Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, 19:1849–1864.

Nesslroader, J. R. and Baltes, P. B., editors (1979). *Longitudinal Research in the Study of Behavior and Development*. Academic Press, New York.

Newell, D. (1992). Intention-to-treat analysis: Implications for quantitative and qualitative research. *International Journal of Epidemiology*, 21(5):837–841.

Newhouse, J. P. and McClellan, M. (1998). Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health*, 19:17–34.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experments. Essay on Principles. Section 9. English translation of excerpts by D. Dabrowska and T. Speed. *Statistical Science*, 5:465–472.

Nie, H., Cheng, J., and Small, D. S. (2011). Inference for the effect of treatment on survival probability in randomized trials with noncompliance and administrative censoring. *Biometrics*, 0:1–9. `Published online 8 MAR 2011 (doi: 10.1111/j.1541-0420.2011.01575.x).`

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Wiley, New York.

Öjelund, H., Madsen, H., and Thyregod, P. (2001). Calibration with absolute shrinkage. *Journal of Chemometrics*, 15:497–509.

O'Malley, A. J. and Normand, S.-L. T. (2005). Likelihood methods for treatment concompliance and subsequent nonresponse in randomized trials. *Biometrics*, 61:325–334.

O'Quigley, J. (2008). *Proportional Hazards Regression*. Statistics for Biology and Health. Springer, New York.

Palmgren, J. and Goetghebeur, E. J. T. (2004). Methods incorporating compliance in treatment evaluation. In Geller, N. L., editor, *Advances in Clinical Trial Biostatistics*, chapter 10, pages 189–212. Marcel Dekker, New York.

Pan, W. (2001). Using frailties in the accelerated failure time model. *Lifetime Data Analysis*, 7:55–64.

Panik, M. J. (2005). *Advanced Statistics from an Elementary Point of View*. Elsevier Academic Press, London, UK.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman Publishers, San Mateo, CA.

Pearl, J. (1995). Causal diagrams for empirical research (with Discussions). *Biometrika*, 82(4):669–710.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research*, 27:226–284.

Pearl, J. (2001). Causal inference in the health sciences: A conceptual introduction. *Health Services & Outcomes Research Methodology*, 2:189–220.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

Pearl, J. (2009b). *Causality: Model Reasoning and Inference*. Cambridge University Press, New York, 2nd. edition.

Pecorelli, S. and Fallo, L. (1998). Hormone replacement therapy in gynecological cancer survivors. *Critical Reviews in Oncology/Hematology*, 27:1–10.

Peng, Y., Little, R. J. A., and Raghunathan, T. E. (2004). An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics*, 60:598–607.

Phillips, C. V. and Goodman, K. J. (2006). Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. *Emerging Themes in Epidemiology*, 3(5).

Piantadosi, S. (2005). *Clinical Trials: A Methodologic Perspective*. John Wiley, New York, 2nd. edition.

Platt, R. W., Schisterman, E. F., and Cole, S. R. (2009). Time-modified confounding. *Americal Joournal of Epidemiology*, 170:687–694.

Pocock, S. J. (1983). *Clinical Trials: A Practical Approach*. John Wiley & Sons, Chichester.

Pocock, S. J. and Abdalla, M. (1998). The hope and the hazards of using compliance data in randomized controlled trials. *Statistics in Medicine*, 17:303–317.

Quan, H., Sun, Q., Zhang, J., and Shih, W. J. (2008). Comparisons between ITT and treatment emergent adverse event analyses. *Statistics in Medicine*, 27:5356–5376.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rees, M. (2005). Unravelling the confusion about HRT in women. *The Journal of Men's Health & Gender*, 2(3):287–291.

Robins, J. M. (1989a). The analysis of randomised and non-randomised AIDS treatment trials using a new approach to causal inference in longitudinal studies. In Sechrest, L., Freeman, H., and Mulley, A., editors, *Health Service Research Methodology: A Focus on AIDS*, pages 113–159. US Public Health Service, National Center for Health Services Research, Washington, DC.

Robins, J. M. (1989b). The control of confounding by intermediate variables. *Statistics in Medicine*, 8:679–701.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23:2379–2412.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In Berkane, M., editor, *Latent Variable Modeling and Applications to Causality*, number 120 in Lecture Notes in Statistics, pages 69–117. Springer Verlag, New York.

Robins, J. M. (1998a). Correction for non-compliance in equivalent trials. *Statistics in Medicine*, 17:269–302.

Robins, J. M. (1998b). Marginal structural models. In 1997 *Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pages 1–10.

Robins, J. M. (1998c). Structural nested failure time models. In Armitage, P. and Colton, T., editors, *The Encyclopedia of Biostatistics*, pages 4372–4389. John Wiley & Sons, Chichester, UK.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, 121:151–79.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In 1999 *Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pages 6–10.

Robins, J. M. (2001). Data, design and background knowledge in etiologic inference. *Epidemiology*, 11:313–320.

Robins, J. M. (2008). Causal models for estimating the effects of weight gain on mortality. *International Journal of Obesity*, 32:S15–S41.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for *pneumocystis carinii pneumonia* on the survived of AIDS patients. *Epidemiology*, 3:319–336.

Robins, J. M. and Greenland, S. (1986). The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, 123(3):392–402.

Robins, J. M. and Hernán, M. A. (2009). Estimation of the causal effects of time-varying exposures. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, Handbooks of Modern Statistical Methods, chapter 23, pages 553–599. Chapman & Hall/CRC, Boca Raton, FL.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:561–570.

Robins, J. M. and Rotznitzky, A. (2004). Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, 91:763–783.

Robins, J. M. and Tsiatis, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics - Theory and Methods*, A 20(8):2609–2631.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer Verlag, New York, 2nd. edition.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer Series in Statistics. Springer, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rossi, R. J. (2010). *Applied Biostatistics for the Health Sciences*. Wiley, New York.

Rothman, K. J. (2002). *Epidemiology: An Introduction*. Oxford University Press, New York.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins, Philadelphia, 3rd. edition.

Roy, J., Hogan, J. W., and Marcus, B. H. (2008). Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics*, 9(2):277–289.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701.

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the Social Statistics Section of the American Statistical Association*, pages 233–239.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 23(13):1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.

Rubin, D. B. (1980). Discussion of 'Randomization analysis of experimental data in the Fisher randomization test' by D. Basu. *Journal of the American Statistical Association*, 75:591–593.

Rubin, D. B. (1986). Comment: which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

Rubin, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, 29(3):343–367.

Rubin, D. B. (2006a). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statistical Science*, 21(3):299–309.

Rubin, D. B. (2006b). Conceptual, computational and inferential benefits of the missing data perspective in applied and theoretical statistical problems. *Allgemeines Statistisches Archiv*, 90:501–513.

Rubin, D. B. (2006c). *Matched Sampling for Causal Effects*. Cambridge Univerisity Press, New York.

Rubin, D. B. (2007). The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26:20–36.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West & Thoemmes (2010). *Psychological Methods*, 15(1):38–46.

Rubin, D. B. and Zell, E. R. (2010). Dealing with noncompliance and missing outcomes in a randomized trial using bayesian technology: Prevention of perinatal sepsis clinical trial, Soweto, South Africa. *Statistical Methodology*, 7:338–350.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12:487–508.

Senn, S. J. (2002). *Crossover Trials in Clinical Research*. Statistics in Practice. Wiley, Chichester, UK, 2nd. edition.

Senn, S. J. (2007). *Statistical Issues in Drug Development*. Statistics in Practice. Wiley, Chichester, UK, 2nd. edition.

Sheiner, L. B. and Rubin, D. B. (1995). Intention to treat analysis and the goals of clinical trials. *Clinical Pharmacology Therappy*, 57:6–15.

Sheng, D. and Kim, M. Y. (2006). The effects of non-compliance on intent-to-treat analysis of equivalence trials. *Statistics in Medicine*, 25:1183–1199.

Shervish, M. (1995). *Theory of Statistics*. Springer Texts in Statistics. Springer-Verlag, New York.

Siannis, F., Copas, J., and Lu, G. (2005). Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*, 6(1):77–91.

Snapinn, S. M. (2000). Noninferiority trials. *Current Controlled Trials in Cardiovascular Medicine*, 1(1):19–21.

Snow, J. (1936). On the mode of communication of cholera. In *Snow On cholera*. The Commonwealth Fund, New York, 2nd. edition. Reprint.

Sobel, M. E. (2000). Causal inference in the social sciences. *Journal of the American Statistical Association*, 95(450):647–651.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27:799–811.

Sommer, A. and Zeger, S. L. (1991). On estimating efficacy from clinical trials. *Statistics in Medicine*, 10:45–52.

Sonis, J. (1998). A closer look at confounding. *Family Medicine*, 30(8):584–588.

Spiegelhalter, D. J. (2004). Incorporating Bayesian ideas into health-care evaluation. *Statistical Science*, 19(1):156–174.

Spiegelhalter, D. J., Thomas, A., Best, N., and Lunn, D. (2004). *WinBUGS version* 1.4.1. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Spirtes, P. (2005). Graphical models, causal inference, and econometric models. *Journal of Economic Methodology*, 12(1):3–34.

Stangl, D. (1995). Prediction and decision making using Bayesian hierarchichal models. *Statistics in Medicine*, 14:2173–2190.

Stangl, D. (2000). The use of reference priors and Bayes factors in the analsysis of clinical trials. In Halloran, M. E. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment and Clinical Trials, (Minneapolis, MN,* 1997*)*, pages 237–249. Springer, New York.

StataCorp (2008). *Stata Statistical Software: Release 10.* College Station, TX, StataCorp LP.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151.

Stephenson, J. M. and Babiker, A. (2000). Overview of study design in clinical epidemiology. *Sexually Transmitted Infections*, 76:244–247.

Sterne, J. A. C. and Tilling, K. (2002). G-estimation of causal effects, allowing for time varying confounding. *The Stata Journal*, 2:164–182.

Steyer, R., Gabler, S., von Davier, A., and Nachtigall, C. (2000a). Causal regression models II: unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online*, 5(3):55–86.

Steyer, R., Gabler, S., von Davier, A., Nachtigall, C., and Buhl, T. (2000b). Causal regression models I: individual and average causal effects. *Methods of Psychological Research Online*, 5(2):39–71.

Steyer, R., Nachtigall, C., Wüthrich-Martone, O., and Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research Online*, 7(1):41–68.

Steyerberg, E. W. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Statistics for Biology and Health. Springer, New York.

Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C., and Habbema, J. D. F. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine*, 23:2567–2586.

Steyerberg, E. W., Eijkemans, M. J. C., and Habbema, J. D. F. (1999). Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, 52:935–942.

Steyerberg, E. W., Eijkemans, M. J. C., Harrell, Jr., F. E., and Habbema, J. D. F. (2000). Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, 19:1059–1079.

Stolberg, H. O., Norman, G., and Trop, I. (2004). Randomized controlled trials. *American Journal of Roentgenology*, 183:1539–1544.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21.

Suppes, P. (1970). A probabilistic theory of causality. *Acta Philosophica Fennica*, 24:5–130.

Symons, M. J. and Moore, D. T. (2002). Hazard rate ratio and prospective epidemiological studies. *Journal of Clinical Epidemiology*, 55:893–899.

Tai, S. S. and Iliffe, S. (2000). Considerations for the design and analysis of experimental studies in physical activity and exercise promotion: advantages of the randomised controlled trial. *British Journal of Sports Medicine*, 34:220–224.

Tang, L., Song, J., Belin, T. R., and Unützer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 24:2111–2128.

Temple, R. and Ellenberg, S. S. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. *Annals of Internal Medicine*, 133(6):455–463.

Thompson, J., Palmer, T., and Moreno, S. (2006). Bayesian analysis in Stata with WinBUGS. *The Stata Journal*, 6(4):530–549.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Tilling, K., Sterne, J. A. C., and Szklo, M. (2002). Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using G-Estimation - The atherosclerosis risk in communities study. *American Journal of Epidemiology*, 155(8):710–718.

Toh, S. and Hernán, M. A. (2008). Causal inference from longitudinal studies with baseline randomization. *The International Journal of Biostatistics*, 4(1). Article 22.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Spinger, New York.

van Belle, G., Fisher, L. D., Heargerty, P. J., and Lumbley, T. (2004). *Biostatistics: A Methodology for the Health Sciencs*. Wiley, New York, 2nd. edition.

van Houwelingen, J. C. (2001). Shrinkage and penalized likelihood methods to improve diagnostic accuracy. *Statistic Nederlica*, 55:17–34.

van Houwelingen, J. C. and le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine*, 8:1303–1325.

Vandenbroeck, P., Wouters, L., Molenberghs, G., Van Gestel, J., and Bijnens, L. (2006). Teaching statistical thinking to life scientists. a case-based approach. *Journal of Biopharmaceutical Statistics*, 16:61–75.

Vansteelandt, S. and Goetghebeur, E. (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society (section B)*, 65:817–835.

Vansteelandt, S., Mertens, K., Suetens, C., and Goetghebeur, E. (2009). Marginal structural models for partial exposure regimes. *Biostatistics*, 10(1):46–59.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454.

Verweij, P. J. M. and van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, 13:2427–2436.

Vigneau, E., Deneau, M. F., Qannari, E. M., and Robert, P. (1997). Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *Journal of Chemometrics*, 11:239–249.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., and McCulloch, C. E. (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Springer-Verlag, Inc., New York.

Walker, A. S., White, I. R., and Babiker, A. G. (2004). Parametric randomization-based methods for correcting for treatment changes in the assessment of the causal effect of treatment. *Statistics in Medicine*, 23:571–590.

Wang, D., Zhang, W., and Bakhai, A. (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. *Statistics in Medicine*, 23:3451–3467.

Wang, J.-D. (2002). *Basic Principles and Practical Applications in Epidemiological Research*, volume 1 of *Quantitative Sciences in Biology and Medicine*. World Scientific, Singapore.

Wasserman, L. (1999). Comment on "Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome" by J. M. Robins, S. Greenland and F.-C. Hu. *Journal of the American Statistical Association*, 49(447):704–706.

Wasserman, L. (2000). Comment on "Causal inference without counterfactuals" by A. P. Dawid. *Journal of the American Statistical Association*, 95:442–443.

Weisberg, H. I. (2010). *Bias and Causation: Models and Judgment for Valid Comparisons*. Wiley, New York.

White, I. R. (2002). `adjhr`: Hazard ratio adjustment of treatment effect for treatment crossovers. Unpublished Stata Software version 2.7, Cambridge (`www.mrc-bsu.cam.ac.uk`).

White, I. R. (2005). Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research*, 14:327–347.

White, I. R. and Goetghebeur, E. J. T. (1998). Clinical trials comparing two treatment policies: which aspects of the treatment policies make a difference? *Statistics in Medicine*, 17(3):319–339.

White, I. R. and Pocock, S. J. (1996). Statistical reporting of clinical trials with individual changes from allocated treatment. *Statistics in Medicine*, 15:249–262.

White, I. R., Walker, A. S., and Babiker, A. G. (2002). `strbee`: Randomization-based efficacy estimator. *The Stata Journal*, 2(2):140–150.

White, I. R., Walker, A. S., and Babiker, A. G. (2004). An approximate randomisation-respecting adjustment to the hazard ratio for time-dependent treatment switches in clinical trials. Unpublished Manuscript: available at `www.mrc-bsu.cam.ac.uk/BSUsite/Publications/Preprints/Time-dependent.pdf`.

Wienke, A. (2010). *Correlated Frailty Models in Survival Analysis*. Biostatistics Series. Chapman & Hall/CRC, Boca Raton, FL.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Massachusetts, 2nd. edition.

Wright, D. B. and London, K. (2009). *Modern Regression Techniques Using R: A Practical Guide for Students and Researchers*. Sage Publications, London.

Wright, P. G. (1928). *Appendix: The Tariff on Animal and Vegetable Oils*. Macmillan, New York.

Xie, H. and Heitjtan, D. F. (2004). Sensitivity analysis of causal inference in a clinical trial subject to crossover. *Clinical Trials*, 1:21–30.

Xing, C., Schumacher, F. R., Conti, D. V., and Witte, J. S. (2003). Comparison of missing data approaches in linkage analysis. *BMC Genetics*, 4(Suppl.1):S44.

Young, J. G., Hernán, M. A., Picciotto, S., and Robins, J. M. (2010). Relation between three classes of structural models for the effect of a time-varying exposure on survival. *Lifetime Data Analysis*, 16:71–84.

Zaccai, J. H. (2004). How to assess epidemiological studies. *Postgraduate Medical Journal*, 80:140–147.

Zhang, J. L. (2004). Causal inference with instrumental variables. In Gelman, A. and Meng, X.-L., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 85–96. Wiley, New York.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratifcation when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104(485):166–176.

Appendices: Annotated Stata, R and WinBUGS codes

Appendix I: Annotated Stata codes for Esprit data analysis (Chapter 4)

```
use espritdata.dta, clear
* ITT analysis for all-cause mortality and myocardial reinfarction
stset monthsacm, failure(dead)
stcox treat hysterec age bmi smk bpress diab
stcox treat
streg treat,dist(weib) nolog
streg treat hysterec age bmi smk bpress diab,dist(weib) nolog
*weibull _t treat,dead(_d) t0(_t0) hr
stset monthsmrcd, failure(mrcd)
stcox treat hysterec age bmi smk bpress diab, nolog
stcox treat
streg treat,dist(weib) nolog
streg treat hysterec age bmi smk bpress diab,dist(weib) nolog

* Testing PH assumption
qui stset monthsacm, failure(dead)
qui stcox treat hysterec age_admi bmi smk1 bldpress diab, /*
    */schoenfeld(sch1*) scaledsch(sca1*)
stphtest, detail
qui stset monthsmrcd, failure(mrcd)
qui stcox treat hysterec age_admi bmi smk1 bldpress diab, /*
    */schoenfeld(sch2*) scaledsch(sca2*)
stphtest, detail

* Per-protocol and as-treated analysis
stset monthsacm,failure(dead)
stcox treatpp
stset monthsmrcd,failure(mrcd)
stcox treatpp
stset monthsacm,failure(dead)
stcox treatat,tvc(treatat)
stset monthsmrcd,failure(mrcd)
stcox treatat,tvc(treatat)
```

```
* Simple regression adjustment of noncompliance
stset monthsacm,failure(dead)
stcox treat noncomply
stcox treat,tvc(noncomply)
stset monthsmrcd,failure(mrcd)
stcox treat noncomply
stcox treat,tvc(noncomply)

* C-PROPHET analysis
replace  complyacm=. if treat==0
stset monthsacm, failure(dead) id(idno)
stcomply treat complyacm, grfit convcrit(0.01)
replace  complymrcd=. if treat==0
stset monthsmrcd, failure(mrcd) id(idno)
stcomply treat complymrcd, grfit convcrit(0.01)

* CHARM analysis
stset monthsacm,failure(dead)
adjhr treat,const
adjhr treat,xo1(monthstab1 comply1) const

stset monthsmrcd,failure(mrcd)
adjhr treat,const
adjhr treat,xo1(monthstab2 comply2) const

* CALM analysis (ITT, adjusting for crossovers and recensoring)
stset monthsacm, failure(dead)
strbee treat, psimin(-3) psimax(2) graph title(ACM: ITT under CALM)
*adjusting for crossovers only
strbee treat, xo1(monthstab1 comply1) psimin(-3) psimax(2)
*adjusting for crossovers and recensoring
strbee treat, xo1(monthstab1 comply1) endstudy(endsty) test(logrank) /*
*/psimin(-3) psimax(2) psistep(0.6) graph title(ACM: CALM with recensoring)

stset monthsmrcd, failure(mrcd)
strbee treat,psimin(-2.5) psimax(2) graph title(MRCD: ITT under CALM)
*adjusting for crossovers only
strbee treat, xo1(monthstab2 comply2) psimin(-2.5) psimax(2)
*adjusting for crossovers and recensoring
strbee treat, xo1(monthstab2 comply2) endstudy(endsty) test(logrank) /*
*/psimin(-2.5) psimax(2) psistep(0.4) graph title(MRCD: CALM with recensoring)
```

Appendix II: Annotated R codes for compliance prediction models (Chapter 5)

```
set.seed(876543)
library(foreign)
library(Hmisc,T)
library(Design,T)
library(glmpath)
options(digits=3)
esp<-read.dta('p:esprit.dta')
attach(esp)

# full model
hyst<-as.factor(hyst)
sclass<-as.factor(sclass)
smoker<-as.factor(smoker)
diabet<-as.factor(diabet)
risks<-as.factor(risks)
alcoh<-as.factor(alcoh)
hrt<-as.factor(hrt)
fract<-as.factor(fract)
trt<-as.factor(trt)
esp<-data.frame(comply,hyst,smoker,sclass,age,risks,diabet,fract,alcoh,hrt,trt)
ddist<-datadist(esp)
options(datadist='ddist')
fit.full<-lrm(comply~hyst+scored(sclass)+risks+diabet+age+smoker+fract+alcoh+hrt,
             x=T,y=T)
validate(fit.full,method="boot",B=200,bw=T,rule='aic',sls=.1,type='individual')
# full reduced model
esp.fred<-data.frame(comply,hyst,risks,alcoh)
ddist.fred<-datadist(esp.fred)
options(datadist='ddist.fred')
fit.fred<-lrm(comply~hyst+risks+alcoh,x=T,y=T)
fit.fred
validate(fit.fred,method="boot",B=200)

# placebo arm
fit0<-update(fit.full, subset=trt==0)
validate(fit0,method="boot",B=200)
# placebo arm reduced model
hyst0<-as.factor(hyst[trt==0])
smk0<-as.factor(smoker[trt==0])
alc0<-as.factor(alcoh[trt==0])
```

```
comply0<-as.factor(comply[trt==0])
esp.red.0<-data.frame(comply0,hyst0,smk0,alc0)
ddist.red.0<-datadist(esp.red.0)
options(datadist='ddist.red.0')
fit.red.0<-lrm(comply0~hyst0+smk0+alc0,x=T,y=T)
fit.red.0
validate(fit.red.0,method="boot",B=200)

# active arm
fit1<-update(fit.full, subset=trt==1)
validate(fit1,method="boot",B=200)
# active arm reduced model
hyst1<-as.factor(hyst[trt==1])
risks1<-as.factor(risks[trt==1])
smk1<-as.factor(smoker[trt==1])
comply1<-as.factor(comply[trt==1])
esp.red.1<-data.frame(comply1,hyst1,risks1,smk1)
ddist.red.1<-datadist(esp.red.1)
options(datadist='ddist.red.1')
fit.red.1<-lrm(comply1~hyst1+risks1+smk1,x=T,y=T)
fit.red.1
validate(fit.red.1,method="boot",B=200)

# LASSO regression and bootstrapping
xpred<-matrix(c(hyst,smoker,sclass,age,risks,diabet,fract,alcoh,hrt),ncol=9)
yresp<-matrix(c(comply),ncol=1)
esppath<-glmpath(xpred,yresp,family=binomial)
par(mfrow=c(2,1))
plot.glmpath(esppath,type='coefficients')
plot.glmpath(esppath,type='aic')
esppath$b.predictor[esppath$aic==min(esppath$aic),]
boot.lasso<-bootstrap.path(xpred,yresp,B=200,method='aic',trace=F)
plot(boot.lasso)
plot(boot.lasso,type='pairplot')
# validating final selected Lasso model
fit.lasso<-lrm(comply~hyst+risks+alcoh,x=T,y=T)
validate(fit.lasso,method="boot",B=200)
```

Appendix III: Annotated WinBUGS+Stata codes-noncompliance in 2 arms (Chapter 6)

```
model Roy08modelEsprit{
# WinBUGS: predicting compliance in placebo and active arms
for (i in 1:1017){
comply0[i]<-comply[i]*equals(trt[i],0)
comply0[i]~dbern(p0[i])
logit(p0[i])<-lambda00+lambda0[1]*hyst[i]+lambda0[2]*smoker[i]+
                    lambda0[3]*(age[i]-mean(age[]))+lambda0[4]*risks[i]+
                    lambda0[5]*fract[i]+lambda0[6]*alcoh[i]
comply1[i]<-comply[i]*equals(trt[i],1)
comply1[i]~dbern(p1[i])
logit(p1[i])<-lambda10+lambda1[1]*hyst[i]+lambda1[2]*smoker[i]+
                    lambda1[3]*(age[i]-mean(age[]))+lambda1[4]*risks[i]+
                    lambda1[5]*fract[i]+lambda1[6]*alcoh[i]
# compliance probability for each treatment arm
mu0[i]<-pow((1+exp(-((lambda00+lambda0[1]*hyst[i]+lambda0[2]*smoker[i]+
        lambda0[3]*(age[i]-mean(age[]))+lambda0[4]*risks[i]+lambda0[5]*fract[i]+
        lambda0[6]*alcoh[i])*equals(trt[i],0)))),-1)
mu1[i]<-pow((1+exp(-((lambda10 + lambda1[1]*hyst[i]+lambda1[2]*smoker[i]+
        lambda1[3]*(age[i]-mean(age[]))+lambda1[4]*risks[i]+lambda1[5]*fract[i]+
        lambda1[6]*alcoh[i])*equals(trt[i],1)))),-1)

# likelihoods- risks of death for individual stratum (same for MRCD outcome)
  # L(pi,beta|Y=1,A=1,Z=1,X) - u(0,1) prior for pis
dead3[i]<-dead[i]*equals(strata[i],3)
dead3[i] ~ dbern(pi11_1[i])
       pi11_1[i]<-pi1[1]*mu01[i]+pi1[2]*mu11[i]
# L(pi,beta|Y=1,A=0,Z=1,X) - u(0,1) prior for pis
dead2[i]<-dead[i]*equals(strata[i],2)
dead2[i] ~ dbern(pi10_1[i])
pi10_1[i]<-pi1[3]*mu00[i]+pi1[4]*mu10[i]
# L(pi,beta|Y=1,A=1,Z=0,X) - u(0,1) prior for pis
dead1[i]<-dead[i]*equals(strata[i],1)
dead1[i] ~ dbern(pi11_0[i])
      pi11_0[i]<-pi1[5]*mu10[i]+pi1[6]*mu11[i]
# L(pi,beta|Y=1,A=0,Z=0,X) - u(0,1) prior for pis
dead0[i]<-dead[i]*equals(strata[i],0)
dead0[i] ~ dbern(pi10_0[i])
       pi10_0[i]<-pi1[3]*mu00[i]+pi1[7]*mu01[i]
#compliance probability ratio
u_x[i]<-min(1,mu1[i]/mu0[i])
```

```
# compliance probabilities for each stratum - phi=0,0.2,0.5 and 0.8
mu11[i]<-mu0[i]*mu1[i]+0.5*mu0[i]*(u_x[i]-mu1[i])
mu01[i]<-mu1[i]-mu0[i]*mu1[i]-0.5*mu0[i]*(u_x[i]-mu1[i])
mu10[i]<-mu0[i]-mu0[i]*mu1[i]-0.5*mu0[i]*(u_x[i]-mu1[i])
mu00[i]<-1-mu0[i]-mu1[i]+mu0[i]*mu1[i]+0.5*mu0[i]*(u_x[i]-mu1[i])
}
# non-informative priors for parameters
   for (k in 1:6) {
lambda0[k] ~ dnorm(0.0,1.0E-6)
lambda1[k] ~ dnorm(0.0,1.0E-6)
}
  lambda00 ~ dnorm(0.0,1.0E-6)
  lambda10 ~ dnorm(0.0,1.0E-6)
# uniform priors for the pis
 for (w in 1:7) {
pi1[w] ~ dunif(0,1.0)
}
# mean posterior estimates
m_mu0<-mean(mu0[]); m_mu1<-mean(mu1[]); m_ux<-mean(u_x[]); m_mu11<-mean(mu11[]);
m_mu01<-mean(mu01[]); m_mu10<-mean(mu10[]); m_mu00<-mean(mu00[]);
tau11<-pi1[2]/pi1[6]; tau01<-pi1[1]/pi1[7]; tau10<-pi1[4]/pi1[5]
}
# WinBUGS calling script
display('log')
check('C:/model.txt')
data('C:/data.txt')
compile(3)
inits(1,'C:/inits1.txt')
inits(2,'C:/inits2.txt')
inits(3,'C:/inits3.txt')
gen.inits()
update(1000)
set('lambda00')
set('lambda0')
set('lambda10')
set('lambda1')
set('m_mu0')
set('m_mu1')
set('m_ux')
set('m_mu11')
set('m_mu01')
set('m_mu10')
```

```
set('m_mu00')
set('pi1')
set('tau11')
set('tau01')
set('tau10')
dic.set()
update(10000)
dic.stats()
coda(*,'C:/coda')
save('C:/log.txt')
quit()

* Calling WinBUGS from Stata
use "C:\espritdata.dta", clear
* transforming data to stata format
wbarray comply trt hyst smoker age risks fract alcoh dead strata, /*
*/ format(%3.0f %3.0f %3.0f %3.0f %8.2f %3.0f %3.0f %3.0f %3.0f %3.0f) /*
*/ noprint saving(C:\data.txt,replace)

* Model fitting
wbscript, sav(C:\script.txt,replace) model(C:\model.txt) data(C:\data.txt) /*
*/ inits(C:\inits.txt) coda(C:\coda) burn(1000) update(10000) chain(3) /*
*/ set(lambda00 lambda0 lambda10 lambda1 m_mu0 m_mu1 m_ux m_mu11 m_mu01 /*
*/ m_mu10 m_mu00 pi1 tau11 tau01 tau10) dic log(C:\log.txt) quit

* running the bugs program
wbrun, script(C:\script.txt) winbugs(c:\Program Files\WinBUGS14\winbugs14.exe)
* Reading the MCMC results
wbcoda,root(C:\coda) clear multichain chain(3) id(chain)
* summarising the results
wbstats lambda00 lambda0* lambda10 lambda1* m_mu0 m_mu1 m_ux /*
*/ m_mu11 m_mu01 m_mu10 m_mu00 pi1* tau11 tau01 tau10
```

Appendix IV: Annotated Stata codes for simulations comparing methods (Chapter 7)

```
capture program drop thesis2010C7
set seed 819726345
set more off
program thesis2010C7, rclass
set obs 1000
gen endsty=24
gen id=_n
gen byte trt=uniform()<=0.5
gen death=0
gen comply=1
*initializing for 95% CI calculation
gen cvitt=0
gen cvcox1=0
gen cvcox2=0
gen cvcproph=0
gen cvcalm=0
gen cvcharm=0
*generating correlated hazard rate and probability of nocompliance-Kleijnen (1974)
*baseline hazard rate=haz0: mu1=0.012, sd1=0.00845, alpha: mu2=0.05, sd2=0.0354
gen rho=0.5
gen z=invnormal(uniform())
gen haz0=rgamma(2,0.006)
quietly sum haz
local mu1haz=r(mean)
local sd1haz=r(sd)
gen alpha=exp(-3.199+0.6374*invnorm(uniform()))
quietly sum alpha
local mu2alpha=r(mean)
local sd2alpha=r(sd)
gen probnoncomply=abs(rho*(‘sd2alpha’/‘sd1haz’)*haz+‘mu2alpha’- /*
*/rho*(‘sd2alpha’/‘sd1haz’)*‘mu1haz’+sqrt(1-rho*rho)*‘sd2alpha’*z)
gen haz1=haz0/2
sort probnoncomply
gen noncomply1=probnoncomply
gen noncomply2=0.4*noncomply1
gen noncomply3=0.2*noncomply1
.....
gen rno1=uniform()
gen event1=0
replace event1 =1 if trt==0 & rno1<(1-exp(-haz0))
```

```
replace event1 =1 if trt==1 & rno1<(1-exp(-haz1))
recode death 0=1 if event1==1
gen time=0.5 if event1==1
recode comply 1=0 if trt==1 & rno1<=noncomply1 & death==0
gen noncomply_t = 0.51 if trt==1 & rno1<=noncomply1 & death==0
.....
gen rno7=uniform()
gen event7=0 if death~=1
recode event7 0=1 if death~=1 & trt==0 & rno7<(1-exp(-haz0))
recode event7 0=1 if death~=1 & trt==1 & comply==1 & rno7<(1-exp(-haz1))
recode event7 0=1 if death~=1 & trt==1 & comply==0 & rno7<(1-exp(-haz0))
recode death 0=1 if event7==1
recode time .=6.5 if event7==1
recode comply 1=0 if trt==1 & rno7<=noncomply2 & death==0
recode noncomply_t .= 6.51 if trt==1 & rno7<=noncomply2 & death==0
.....
gen rno13=uniform()
gen event13=0 if death~=1
recode event13 0=1 if death~=1 & trt==0 & rno13<(1-exp(-haz0))
recode event13 0=1 if death~=1 & trt==1 & comply==1 & rno13<(1-exp(-haz1))
recode event13 0=1 if death~=1 & trt==1 & comply==0 & rno13<(1-exp(-haz0))
recode death 0=1 if event13==1
recode time .=12.5 if event13==1
recode comply 1=0 if trt==1 & rno13<=noncomply3 & death==0
recode noncomply_t .= 12.51 if trt==1 & rno13<=noncomply3 & death==0
.....
gen rno19=uniform()
gen event19=0 if death~=1
recode event19 0=1 if death~=1 & trt==0 & rno19<(1-exp(-haz0))
recode event19 0=1 if death~=1 & trt==1 & comply==1 & rno19<(1-exp(-haz1))
recode event19 0=1 if death~=1 & trt==1 & comply==0 & rno19<(1-exp(-haz0))
recode death 0=1 if event19==1
recode time .=18.5 if event19==1
.....
gen rno24=uniform()
gen event24=0 if death~=1
recode event24 0=1 if death~=1 & trt==0 & rno24<(1-exp(-haz0))
recode event24 0=1 if death~=1 & trt==1 & comply==1 & rno24<(1-exp(-haz1))
recode event24 0=1 if death~=1 & trt==1 & comply==0 & rno24<(1-exp(-haz0))
recode death 0=1 if event24==1
recode time .=23.5 if event24==1
gen cens=0
recode cens 0=1 if death==1
```

```
recode time .=24 if death==0
gen noncomply=1 if comply==0
replace noncomply=0 if comply==1
.....

quietly count if death==1 /*totdeath*/
quietly count if noncomply==1 /*noncomply10*/
quietly count if noncomply==1 & trt==0 /*noncomply10*/
quietly count if noncomply==1 & trt==1 /*noncomply10*/

* ITT analysis
stset time death, id(id)
stcox trt,nohr
local ittest=_b[trt]
local sditt=_se[trt]
replace cvitt=1 if `ittest'-1.96*`sditt'<=ln(0.5)&`ittest'+1.96*`sditt'>=ln(0.5)
quietly count if cvitt==1
gen totcvitt=r(N)
sum totcvitt
local covitt=r(mean)/1000
di "`covitt'" "`ittest'"
return scalar covitt=`covitt'
return scalar ittest=`ittest'
* coxreg1- simple regression adjustment with 0/1 time-invariant noncompliance
stset time death, id(id)
stcox trt noncomply,nohr
local cox1=_b[trt]
local sdcox1=_se[trt]
local noncox1=_b[noncomply]
replace cvcox1=1 if `cox1'-1.96*`sdcox1'<=ln(0.5)&`cox1'+1.96*`sdcox1'>=ln(0.5)
quietly count if cvcox1==1
gen totcvcox1=r(N)
sum totcvcox1
local covcox1=r(mean)/1000
di "`cox1'" "`noncox1'" "`covcox1'"
return scalar covcox1=`covcox1'
return scalar cox1=`cox1'
return scalar noncox1=`noncox1'
* coxreg2 - simple regression adjustment with 0/1 time-varying noncompliance
stset time death, id(id)
stcox trt, tvc(noncomply) nohr
matrix bcox2=e(b)
matrix vcox2=e(V)
```

```
local cox2=bcox2[1,1]
local noncox2=bcox2[1,2]
local sdcox2=sqrt(vcox2[1,1])
replace cvcox2=1 if `cox2'-1.96*`sdcox2'<=ln(0.5)&`cox2'+1.96*`sdcox2'>=ln(0.5)
quietly count if cvcox2==1
gen totcvcox2=r(N)
sum totcvcox2
local covcox2=r(mean)/1000
di "`cox2'" "`noncox2'" "`covcox2'"
return scalar covcox2=`covcox2'
return scalar cox2=`cox2'
return scalar noncox2=`noncox2'
*C-PROPHET analysis: HR estimate got as cprophet in stcomply2 ado file
replace  comply=. if trt==0
stset time,fail(death)
stcomply2 trt comply
local cprophet=log(r(cprophet))
local sdcproph=_se[trt]
replace cvcproph=1 if `cprophet'-1.96*`sdcproph'<=ln(0.5)& /*
*/`cprophet'+1.96*`sdcproph'>=ln(0.5)
quietly count if cvcproph==1
gen totcproph=r(N)
sum totcproph
local covcproph=r(mean)/1000
di "`cprophet'" "`covcproph'"
return scalar cprophet=`cprophet'
return scalar covcproph=`covcproph'
*CALM analysis
stset time death, id(id)
strbee trt, xo1(noncomply_t noncomply) endstudy(endsty) /*
*/test(logrank) psimin(-2) psimax(2) psistep(0.01)
local calm=r(psi)
local sdcalm=_se[trt]
replace cvcalm=1 if `calm'-1.96*`sdcalm'<=ln(0.5)& `calm'+1.96*`sdcalm'>=ln(0.5)
quietly count if cvcalm==1
gen totcvcalm=r(N)
sum totcvcalm
local covcalm=r(mean)/1000
di "`calm'" "`covcalm'"
return scalar calm=`calm'
return scalar covcalm=`covcalm'
*true correlation from the simulations
corr haz0 probnoncomply
```

```
local corrhaznonc=r(rho) /*corrhaznonc*/

*CHARM analysis
stset time death, id(id)
adjhr trt, xo1(noncomply_t noncomply) const
matrix bcharm=e(b)
matrix vcharm=e(V)
local charm=bcharm[1,1]
local sdcharm=sqrt(vcharm[1,1])
replace cvcharm=1 if `charm'-1.96*`sdcharm'<=ln(0.5)&`charm'+1.96*`sdcharm'>=ln(0.5)
quietly count if cvcharm==1
gen totcvcharm=r(N)
sum totcvcharm
local covcharm=r(mean)/1000
di "`charm'" "`covcharm'"
return scalar covcharm=`covcharm'
return scalar charm=`charm'
*simple CHARM approximation (White et al., 2004): adjusting ITT with theta
* calculating theta=pr(death) among non-compliers in the treated group
quietly count if group==1 & death==1
gen totdeadtrt=r(N)
sum totdeadtrt
local deadtrt=r(mean) /*deadtrt*/
quietly count if noncomply==1 & trt==1 & death==1
gen tottheta=r(N)
sum tottheta
local theta0=r(mean)
local theta=`theta0'/`deadtrt' /*theta*/
local charms=(exp(`ittest')*(1-`theta'))/(1-`theta'*exp(`ittest')) /*charms*/
.....
simulate totdeath=r(totdeath) noncomply10=r(noncomply10) /*
*/ noncomply0=r(noncomply0)noncomply1=r(noncomply1) theta=r(theta) /*
*/ corrhaznonc=r(corrhaznonc) ittest=r(ittest) cox1=r(cox1) /*
*/ noncox1=r(noncox1) cox2=r(cox2) charms=r(charms) noncox2=r(noncox2) /*
*/ cprophet=r(cprophet) calm=r(calm) charm=r(charm) covitt=r(covitt) /*
*/ covcox1=r(covcox1) covcox2=r(covcox2) covcalm=r(covcalm) /*
*/ covcproph=r(covcproph) covcharm=r(covcharm), reps(2000):thesis2010C7
```

Appendix V: Annotated Stata+WinBUGS codes for simulations implementing Roy et al. (2008) model (Chapter 8)

```
/*Stata codes generating simulation data*/
capture program drop Roy08simthesis2010C8
set seed 918273645
set more off
program define Roy08simthesis2010C8, rclass
drop _all
set obs 1000
gen id=_n
/* two risk factor distribution*/
gen hyst=_n<251
gen risk=_n<151
replace risk=1 if _n>250 & _n<701
gen byte trt=uniform()<=0.5
/* setting marginal compliance probs in H=R=0 for each stratum - state 1*/
gen compA_00=0.55
gen compB_00=0.30
/* odds ratios-effects of hyst*/
gen orhystA=2
gen orriskA=4
gen orhyriA=8
/* odds ratios - effects of risk factors*/
gen orhystB=5
gen orriskB=3
gen orhyriB=15
/* filling marginal compliance probs in remaining risk factor subgroups*/
gen compA_11=orhyriA*compA_00/(1-compA_00+orhyriA*compA_00)
gen compA_10=orhystA*compA_00/(1-compA_00+orhystA*compA_00)
gen compA_01=orriskA*compA_00/(1-compA_00+orriskA*compA_00)
gen compB_11=orhyriB*compB_00/(1-compB_00+orhyriB*compB_00)
gen compB_10=orhystB*compB_00/(1-compB_00+orhystB*compB_00)
gen compB_01=orriskB*compB_00/(1-compB_00+orriskB*compB_00)
gen u11=compB_11/compA_11
gen u10=compB_10/compA_10
gen u01=compB_01/compA_01
gen u00=compB_00/compA_00
replace u11=1 if u11>1
replace u10=1 if u10>1
replace u01=1 if u01>1
replace u00=1 if u00>1
/*joint compliance prob following Roy etal 2008 for different phi values - state 2*/
```

```
gen phi=0.5
gen mu11_11=compA_11*compB_11+phi*compA_11*(u11-compB_11)
gen mu01_11=compB_11-compA_11*compB_11-phi*compA_11*(u11-compB_11)
gen mu10_11=compA_11-compA_11*compB_11-phi*compA_11*(u11-compB_11)
gen mu00_11=1-compA_11-compB_11+compA_11*compB_11+phi*compA_11*(u11-compB_11)
gen mu11_01=compA_01*compB_01+phi*compA_01*(u01-compB_01)
gen mu01_01=compB_01-compA_01*compB_01-phi*compA_01*(u01-compB_01)
gen mu10_01=compA_01-compA_01*compB_01-phi*compA_01*(u01-compB_01)
gen mu00_01=1-compA_01-compB_01+compA_01*compB_01+phi*compA_01*(u01-compB_01)
gen mu11_10=compA_10*compB_10+phi*compA_10*(u10-compB_10)
gen mu01_10=compB_10-compA_10*compB_10-phi*compA_10*(u10-compB_10)
gen mu10_10=compA_10-compA_10*compB_10-phi*compA_10*(u10-compB_10)
gen mu00_10=1-compA_10-compB_10+compA_10*compB_10+phi*compA_10*(u10-compB_10)
gen mu11_00=compA_00*compB_00+phi*compA_00*(u00-compB_00)
gen mu01_00=compB_00-compA_00*compB_00-phi*compA_00*(u00-compB_00)
gen mu10_00=compA_00-compA_00*compB_00-phi*compA_00*(u00-compB_00)
gen mu00_00=1-compA_00-compB_00+compA_00*compB_00+phi*compA_00*(u00-compB_00)
.....
/*compliance types: 0=00, 1=10, 2=01,3=11 */
gen compno=runiform()
gen type=cond(compno<mu00_00, 0,cond(compno<mu00_00+psi10_00, 1,/*
*/cond(compno<mu00_00+psi10_00+mu01_00, 2,3))) if hyst==0 & risk==0
return scalar n00_00=‘n00_00’ /*mean count if type==0 & hyst==0 & risk==0*/
return scalar n10_00=‘n10_00’ /*mean count if type==1 & hyst==0 & risk==0*/
return scalar n01_00=‘n01_00’ /*mean count if type==2 & hyst==0 & risk==0*/
return scalar n11_00=‘n11_00’ /*mean count if type==3 & hyst==0 & risk==0*/
replace type=cond(compno<mu00_10, 0,cond(compno<mu00_10+psi10_10, 1,/*
*/cond(compno<mu00_10+mu10_10+mu01_10, 2,3))) if hyst==1 & risk==0
return scalar n00_10=‘n00_10’;n10_10=‘n10_10’;n01_10=‘n01_10’;n11_10=‘n11_10’
replace type=cond(compno<mu00_01, 0,cond(compno<mu00_01+psi10_01, 1,/*
*/cond(compno<mu00_01+mu10_01+mu01_01, 2,3))) if hyst==0 & risk==1
return scalar n00_01=‘n00_01’;n10_01=‘n10_01’;n01_01=‘n01_01’;n11_01=‘n11_01’
replace type=cond(compno<mu00_11, 0,cond(compno<mu00_11+psi10_11, 1,/*
*/cond(compno<mu00_11+mu10_11+mu01_11, 2,3))) if hyst==1 & risk==1
return scalar n00_11=‘n00_11’;n10_11=‘n10_11’;n01_11=‘n01_11’;n11_11=‘n11_11’
.....
return scalar enotype00=‘enotype00’=round(‘n00_00’+‘n00_10’+‘n00_01’+‘n00_11’,1)
return scalar enotype10=‘enotype10’=round(‘n10_00’+‘n10_10’+‘n10_01’+‘n10_11’,1)
return scalar enotype01=‘enotype01’=round(‘n01_00’+‘n01_10’+‘n01_01’+‘n01_11’,1)
return scalar enotype11=‘enotype11’=round(‘n11_00’+‘n11_10’+‘n11_01’+‘n11_11’,1)
.....

drop type   /* allow (start) independent generation of type - state 3*/
gen haz0=rgamma(2,0.006)
sort haz0 hyst risk /*ranking wrt to baseline hazard to help choose the 4 types */
gen partno=runiform()
```

```
/*choosing number compliant for H=R=0: n from 1 to 300 */
gen comp23=n1000/(n1000+n0100)
gen type=3 if _n<=n1100
replace type=0 if _n>300-n0000 & _n<=300
replace type=1 if _n>n1100 & _n<=300-n0000 & partno<=comp23
replace type=2 if _n>n1100 & _n<=300-n0000 & partno>comp23
/*choosing number compliant for H=0, R=1: n from 301 to 750 */
replace comp23=n1001/(n1001+n0101)
replace type=3 if _n>300 & _n<=300+n1101
replace type=0 if _n>750-n0001 & _n<=750
replace type=1 if _n>300+n1101 & _n<=750-n0001 & partno<=comp23
replace type=2 if _n>300+n1101 & _n<=750-n0001 & partno>comp23
/*choosing number compliant when H=1,R=0: n from 751 to 850 */
replace comp23=n1010/(n1010+n0110)
replace type=3 if _n>750 & _n<=750+n1110
replace type=0 if _n>850-n0010 & _n<=850
replace type=1 if _n>750+n1110 & _n<=850-n0010 & partno<=comp23
replace type=2 if _n>750+n1110 & _n<=850-n0010 & partno>comp23
/*choosing number compliant when H=R=1: n from 851 to 1000 */
replace comp23=n1011/(n1011+n0111)
replace type=3 if _n>850 & _n<=850+n1111
replace type=0 if _n>1000-n0011 & _n<=1000
replace type=1 if _n>850+n1111 & _n<=1000-n0011 & partno<=comp23
replace type=2 if _n>850+n1111 & _n<=1000-n0011 & partno>comp23
.....
/* specifying heterogeneous hazard rates for each stratum*/
/*HR-heteg-(AB)-11:(0.009,0.006),01:(0.008,0.006),10:(0.009,0.007),00:(0.010,0.008)*/
/*HR-homog-(AB)-11:(0.009,0.006),01:(0.009,0.006),10:(0.009,0.006),00:(0.009,0.006)*/
gen haza11=(3/4)*haz0;gen hazb11=haz0/2;gen haza01=(2/3)*haz0;gen hazb01=haz0/2;
gen haza10=(3/4)*haz0;gen hazb10=(7/12)*haz0;gen haza00=(5/6)*haz0;gen hazb00=(2/3)*haz0
/*gen haza11=(3/4)*haz0;gen hazb11=haz0/2;gen haza01=haza11;gen hazb01=hazb11;*/
/*gen haza10=haza11;gen hazb10=hazb11;gen haza00=haza11;gen hazb00=hazb11*/

/* linking baseline hazard rates to compliance types and arm-trt0=A,trt1=B*/
gen haz=haz0
replace haz=haza00 if trt==0 & type==0
replace haz=hazb00 if trt==1 & type==0
replace haz=haza10 if trt==0 & type==1
replace haz=hazb10 if trt==1 & type==1
replace haz=haza01 if trt==0 & type==2
replace haz=hazb01 if trt==1 & type==2
replace haz=haza11 if trt==0 & type==3
replace haz=hazb11 if trt==1 & type==3
/* selecting overall noncompliers for treat A and B */
gen noncompA=1 if trt==0 & type==0 | trt==0 & type==2
recode noncompA .=0 if trt==0
```

```
gen noncompB=1 if trt==1 & type==0 | trt==1 & type==1
recode noncompB .=0 if trt==1
/* mean compliance: noncomplyA and noncomplyB */
return scalar noncomplyA=‘noncomplyA’
return scalar noncomplyB=‘noncomplyB’
.....
/* correlations between baseline hazards and noncompliance for A and B: */
return scalar corhazA=‘corhazA’ /* corr(noncomplyA,haz0) */
return scalar corhazB=‘corhazB’ /* corr(noncomplyB,haz0) */
.....
gen death=0
gen rand1=runiform()
gen event1=0
replace event1 =1 if rand1<(1-exp(-haz))
recode death 0=1 if event1==1
gen time=1 if event1==1
.....
gen rand24=runiform()
gen event24=0 if death!=1
replace event24 =1 if death!=1 & rand24<(1-exp(-haz))
recode death 0=1 if event24==1
recode time .=24 if event24==1
recode time .=24 if death==0

/* itt analysis - no correction for noncompliance*/
stset time death, id(id)
stcox trt   /* effitt */
/* ITT effects in each stratum (stratum analysis)-comparing efficacy of B to A*/
stset time death, id(id)
stcox trt if type==3    /* eff11 */
stcox trt if type==2    /* eff01 */
stcox trt if type==1    /* eff10 */
stcox trt if type==0    /* eff00 */
/*logistic model predicting arm-specific compliance*/
gen complyA=0 if noncompA==1
replace complyA=1 if noncompA==0
logit complyA hyst risk
gen complyB=0 if noncompB==1
replace complyB=1 if noncompB==0
logit complyB hyst risk
keep complyA complyB trt hyst risk type death
save "C:Roy08simsdata.dta",replace
end
simulate mn_haz0=r(mn_haz0) sd_haz0=r(sd_haz0) n00_00=r(n00_00) /*
*/ n10_00=r(n10_00) n01_00=r(n01_00) n11_00=r(n11_00) n00_10=r(n00_10) /*
*/ n10_10=r(n10_10) n01_10=r(n01_10) n11_10=r(n11_10) n00_01=r(n00_01) /*
```

```
*/ n10_01=r(n10_01) n01_01=r(n01_01) n11_01=r(n11_01) n00_11=r(n00_11) /*
*/ n10_11=r(n10_11) n01_11=r(n01_11) n11_11=r(n11_11) enotype00=r(enotype00) /*
*/ enotype10=r(enotype10) enotype01=r(enotype01) enotype11=r(enotype11) /*
*/ tot00=r(tot00) tot10=r(tot10) tot01=r(tot01) tot11=r(tot11) /*
*/ haz_00=r(haz_00) haz_10=r(haz_10) haz_01=r(haz_01) haz_11=r(haz_11) /*
*/ mn_trt1=r(mn_trt1) noncomplyA=r(noncomplyA) noncomplyB=r(noncomplyB) /*
*/ nod1trt0=r(nod1trt0) nod1trt1=r(nod1trt1) corhazA=r(corhazA) /*
*/ corhazB=r(corhazB) consA=r(consA) compAh=r(compAh) compAr=r(compAr) /*
*/ consB=r(consB) compBh=r(compBh) compBr=r(compBr) orcAhyst=r(orcAhyst) /*
*/ orcArisk=r(orcArisk) probcA=r(probcA) orcBhyst=r(orcBhyst) /*
*/ orcBrisk=r(orcBrisk) probcB=r(probcB) effitt=r(effitt) eff11=r(eff11) /*
*/ eff01=r(eff01) eff10=r(eff10) eff00=r(eff00), reps(2000):Roy08simthesis2010C8


model Roy08sim{
# WinBUGS+Stata: predicting compliance in placebo and active arms
for (i in 1:1000) {
complyA[i]~dbern(pA[i])
logit(pA[i])<-(lambda0[1]+lambda0[2]*hyst[i]+lambda0[3]*risk[i])*equals(trt[i],0)
complyB[i]~dbern(pB[i])
logit(pB[i])<-(lambda1[1]+lambda1[2]*hyst[i]+lambda1[3]*risk[i])*equals(trt[i],1)
# compliance probability for each arm
muA[i]<-pow((1+exp(-(lambda0[1]+lambda0[2]*hyst[i]+lambda0[3]*risk[i]))),-1)
muB[i]<-pow((1+exp(-(lambda1[1]+lambda1[2]*hyst[i]+lambda1[3]*risk[i]))),-1)
# likelihoods - risks of death for individual stratum
  # L(pi,beta|Y=1,A=1,Z=1,X) - u(0,1) prior for pis
death3[i]<-death[i]*equals(type[i],3)
death3[i] ~ dbern(pi11_1[i])
    pi11_1[i]<-pi1[1]*mu01[i]+pi1[2]*mu11[i]
# L(pi,beta|Y=1,A=0,Z=1,X) - u(0,1) prior for pis
death2[i]<-death[i]*equals(type[i],2)
death2[i] ~ dbern(pi10_1[i])
pi10_1[i]<-pi1[3]*mu00[i]+pi1[4]*mu10[i]
  # L(pi,beta|Y=1,A=1,Z=0,X) - u(0,1) prior for pis
death1[i]<-death[i]*equals(type[i],1)
death1[i] ~ dbern(pi11_0[i])
    pi11_0[i]<-pi1[5]*mu10[i]+pi1[6]*mu11[i]
# L(pi,beta|Y=1,A=0,Z=0,X) - u(0,1) prior for pis
death0[i]<-death[i]*equals(type[i],0)
death0[i] ~ dbern(pi10_0[i])
       pi10_0[i]<-pi1[3]*mu00[i]+pi1[7]*mu01[i]
# ratio of compliance probability: B/A
u_x[i]<-min(1,muB[i]/muA[i])
# compliance probabilities for each stratum -phi=0,0.2,0.5,0.8
mu11[i]<-muA[i]*muB[i]+0.5*muA[i]*(u_x[i]-muB[i])
mu01[i]<-muB[i]-muA[i]*muB[i]-0.5*muA[i]*(u_x[i]-muB[i])
```

```
mu10[i]<-muA[i]-muA[i]*muB[i]-0.5*muA[i]*(u_x[i]-muB[i])
mu00[i]<-1-muA[i]-muB[i]+muA[i]*muB[i]+0.5*muA[i]*(u_x[i]-muB[i])
}
# Also considered general phi without pre-specification: phi~dunif(0,1.0)
# non-informative priors for parameters
   for (k in 1:3) {
lambda0[k] ~ dnorm(0.0,1.0E-6)
lambda1[k] ~ dnorm(0.0,1.0E-6)
}
# uniform priors for the pis=risk/probabilities of event
 for (p in 1:7) {
pi1[p] ~ dunif(0,1)
}
# mean posterior estimates
m_muA<-mean(muA[]);m_muB<-mean(muB[]);m_ux<-mean(u_x[]);m_mu11<-mean(mu11[]);
m_mu01<-mean(mu01[]);m_mu10<-mean(mu10[]);m_mu00<-mean(psi00[]);
tau11<-pi1[2]/pi1[6];tau01<-pi1[1]/pi1[7];tau10<-pi1[4]/pi1[5]
}
# script calling WinBUGS
display('log');check('C:/Roy08C8/model.txt');data('C:/Roy08C8/data.txt')
compile(3);inits(1,'C:/Roy08C8/inits1.txt');inits(2,'C:/Roy08C8/inits2.txt')
inits(3,'C:/Roy08C8/inits3.txt');gen.inits();update(1000);set('lambda0')
set('lambda1');set('m_muA');set('m_muB');set('m_ux');set('m_mu11') set('m_mu01')
set('m_mu10');set('m_mu00');set('pi1');set('tau11');set('tau01') set('tau10')
dic.set();update(10000);dic.stats();coda(*,'C:/Roy08C8/coda')
save('C:/Roy08C8/log.txt');quit()

*Stata calling WinBUGS
use "C:\RoyChp8\Data4Stata-05.dta", clear
* transforming data to stata format
wbarray complyA complyB trt hyst risk type death, /*
*/ format(%3.0f %3.0f %3.0f %3.0f %3.0f %3.0f %3.0f) /*
*/ noprint saving(C:\RoyChp8\data.txt,replace)
* Model fitting
wbscript, sav(C:\RoyChp8\script.txt,replace) model(C:\RoyChp8\model.txt) /*
*/ data(C:\RoyChp8\data.txt) inits(C:\RoyChp8\inits.txt) coda(C:\RoyChp8\coda) /*
*/ burn(1000) update(10000) chain(3); set(lambda0 lambda1 m_muA m_muB m_ux m_mu11 /*
*/ m_mu01 m_mu10 m_mu00 pi1 tau11 tau01 tau10) dic log(C:\RoyChp8\log.txt) quit
* running the bugs program
wbrun, script(C:\RoyChp8\script.txt) winbugs(c:\Program Files\WinBUGS14\winbugs14.exe)
* Reading the MCMC results
wbcoda,root(C:\RoyChp8\coda) clear multichain chain(3) id(chain)
* summarising the results
wbstats lambda0* lambda1* m_muA m_muB m_ux /*
*/ m_mu11 m_mu01 m_mu10 m_mu00 pi1* tau11 tau01 tau10
```