Evaluating the development and validation of empirically-derived prognostic models for pressure ulcer risk assessment: a systematic review

DOI:
10.1016/j.ijnurstu.2018.08.005

Document Version
Accepted author manuscript

Citation for published version (APA):

Published in:
International Journal of Nursing Studies

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.
Accepted Manuscript

Title: Evaluating the development and validation of empirically-derived prognostic models for pressure ulcer risk assessment: a systematic review

Authors: Chunhu Shi, Jo C. Dumville, Nicky Cullum

DOI: https://doi.org/10.1016/j.ijnurstu.2018.08.005

Reference: NS 3201

To appear in:

Received date: 12-4-2018
Revised date: 9-8-2018
Accepted date: 13-8-2018

Please cite this article as: Shi C, Dumville JC, Cullum N, Evaluating the development and validation of empirically-derived prognostic models for pressure ulcer risk assessment: a systematic review, International Journal of Nursing Studies (2018), https://doi.org/10.1016/j.ijnurstu.2018.08.005

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Evaluating the development and validation of empirically-derived prognostic models for pressure ulcer risk assessment: a systematic review

Chunhu Shi 1*, Jo C Dumville 1, Nicky Cullum 1, 2

1 Division of Nursing, Midwifery & Social Work, School of Health Sciences, Faculty of Biology, Medicine & Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK, M13 9PL

2 Research and Innovation Division, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, 1st Floor, Nowgen Building, 29 Grafton Street, Manchester, UK, M13 9WU

* Corresponding author

Mobile: +44 (0) 7421 239 138
E-mail: chunhu.shi@postgrad.manchester.ac.uk; shichunhu2014@hotmail.com (CS)

Abstract
Background: Researchers advocate developing empirically-derived prognostic models to predict pressure ulcer risk. However, there remains a scarcity of evidence about the performance and clinical value of these models.

Objectives: To identify and describe empirically-derived models for predicting pressure ulcer risk; to assess the predictive performance of these models; and to evaluate their clinical impact in reducing pressure ulcer incidence.

Methods: We performed a comprehensive database search up to February 2017 and searched other resources to identify longitudinal studies that developed and/or validated prognostic models for predicting pressure ulcer risk and studies evaluating the clinical effects of such models. There were no language or publication date restrictions. Two reviewers independently conducted study selection. Using a pre-prepared data extraction form, one reviewer collected data on the characteristics and performance of the included models and assessed study risk of bias. A second reviewer checked all the data. Using narrative synthesis, we summarised the characteristics of the included studies and models. Using meta-analysis, we combined performance (discrimination and calibration) measurement statistics for relevant models.
**Results:** We included 24 studies with 28 data sources in the review and identified 22 models that were developed using these data. Of the 22 models, only seven had further external validations (one model was validated twice). In development, a third of models used univariate analysis alone to identify statistically significant predictors for subsequent multivariable analysis; and nine of the 16 developed models were formed using stepwise selection processes in multivariable analysis. Missing data were often incompletely reported, and continuous predictors were correctly handled in only two models (e.g., using restricted cubic spline). Sample sizes of the model development studies were small with 13 models involving fewer than 10 events per variable. The risk of bias associated with the development of all 22 models and eight validations was judged as high or unclear. The predictive performance was reported as: c-statistic point estimates ranging from 0.65 to 0.89, and total Observed:Expected risk ratios between 0.94 and 1.00. Compared with heuristic tools, relevant included models had better discrimination and calibration. No eligible study was identified that evaluated the clinical impact of any included model.

**Conclusions:** Whilst many prognostic models for predicting ulcer risk have been developed few have been validated. The methods used for model development are generally flawed which reduces the potential for using these models in practice. Future research should address these weaknesses.

**Keywords:** Pressure Ulcer; Systematic Review; Meta-Analysis; Prognostic Model
What is already known about the topic?

- Pressure ulcer management guidelines recommend the use of prognostic models (usually called ‘risk assessment tools’) to predict pressure ulcer risk.
- Seven previous systematic reviews identified 57 heuristic tools (e.g., Braden and Waterlow scales), that is, models that were developed without the support of advanced statistical methods. These tools are regarded as having limited predictive ability and have not been shown to reduce pressure ulcer incidence.
- It is now widely acknowledged that prognostic models should be developed using advanced statistical techniques (termed “empirically-derived models”). It was necessary to evaluate the available empirically-derived models for pressure ulcer risk in terms of methodological issues, predictive performance and their clinical impact.

What this paper adds

- This systematic review identifies 22 empirically-derived prognostic models for pressure ulcer risk but finds only seven of them are externally validated.
- None of the 22 models was developed using current best methods because of sub-optimal predictor selection strategies, insufficient sample sizes, and inappropriate methods for handling missing data and continuous predictors.
- Included studies show that empirically-derived models have better discrimination and calibration abilities than heuristic tools but no study has evaluated the effects of these models in reducing pressure ulcer incidence.
1. Introduction

Pressure ulcers (also known as pressure injuries, pressure sores, bedsores, and decubitus ulcers) are localised injuries to skin and/or underlying tissue (NPUAP/EPUAP/PPPIA, 2014). Pressure ulcers have a point prevalence of approximately 3.1 per 10,000 in the United Kingdom, which represents a serious health burden (Cullum et al., 2016). The current United Kingdom guideline states that all patients are potentially at risk of pressure ulcer development and that patients in secondary care or care homes should have this risk assessed and documented. Nurses can consider using a prognostic model to support their clinical judgement in assessing pressure ulcer risk and people determined to be at risk of ulceration then should be provided with prevention measures (NICE, 2014). A prognostic model is a combination of two or more patient-level or other characteristics (e.g., age or sex) which are able to predict the risk of an outcome occurring. In this context such characteristics are termed prognostic factors or predictors (Steyerberg et al., 2013). Prognostic models have been developed to assess individual pressure ulcer risk. The development of such models involves two key steps (1) the identification of potentially important prognostic factors for predicting pressure ulcer risk and then (2) the combination of some (or all) of these factors in a model that provides the most accurate risk assessment.

Several conventional approaches to model development have been applied in practice: e.g., the use of expert consensus, Delphi technique, or literature review (Pancorbo-Hidalgo et al., 2006). Using these approaches, experts may arbitrarily decide which (potentially) prognostic factors should be included in the final model (termed “final model predictors”) and the score (or weight) that should be allocated to the predictors to reflect their estimated prognostic strength for pressure ulcer incidence. Models developed using this approach are termed “heuristic tools” hereafter. Seven systematic reviews have identified existing heuristic tools and synthesised their predictive performance (mainly diagnostic accuracy performance) (García-Fernández et al., 2014; He et al., 2012; Kottner et al., 2009; Kottner et al., 2013; Moore and Cowman, 2014; Mortenson et al., 2008; Pancorbo-Hidalgo et al., 2006). Together these reviews identified 57 heuristic tools from longitudinal and case-control studies (the numbers of included studies in reviews ranged from three to 73). These reviews suggest that reviewed heuristic tools have limited ability to distinguish between people at risk of pressure ulcers and those not at risk. Moore and Cowman (2014) assessed the impact of using these heuristic tools as part of a wider “predict and prevent strategy”, where the results of the risk assessment dictated prevention practice and subsequent outcomes were assessed. Moore and Cowman concluded that there is a lack of evidence that using heuristic tools to assess pressure ulcer risk reduces in pressure ulcer incidence.

Rather than relying on ad hoc selection of potentially prognostic factors and arbitrary allocation of weights to factors that are finally included in a model, the development of prognostic models should be supported by advanced statistical techniques (e.g., multivariable analysis; termed “empirically-derived models” hereafter) (Steyerberg et al., 2013). The development of empirically-derived models usually starts with an informed list of prognostic factors that have hypothesised associations with pressure ulcer
risk (termed “candidate predictors”) (Moons et al., 2012). Important factors for risk prediction are then
selected from these candidates using multivariable analysis (e.g., logistic regression) and their
contributions to the risk are statistically weighted to build a final model (‘‘model development’’) (Moons
et al., 2012). Before the use of a model in practice, further validation is needed to assess whether the
model: (1) is predictive in other populations (and settings) beyond those directly involved in its
development (‘‘external validation’’), (2) performs better than alternative models (e.g., existing heuristic
tools, other empirically-derived models) in the same cohort (‘‘prognostic model comparisons’’), and (3)
has actual clinical impact with pressure ulceration reduced via the effective provision of preventive
measures once the prognostic model has identified those at risk (Steyerberg et al., 2013).

The development and validation of empirically-derived models often involves many methodological
and statistical issues (e.g., how decisions are made about selecting predictors in a model) that determinethe quality of the model (Steyerberg et al., 2013). A perfect model would always correctly identify
every individual who will develop the outcome and those who will not. To evaluate model performance
researchers commonly measure: the extent to which a model can predict a higher risk for individuals
who go on to develop the outcome versus those who do not (“discrimination”), and how well the
outcome risk predicted by a model (pressure ulcer incidence in this case) agrees with the actual
(observed) outcome incidence in individual groups of different risk categories (“calibration”) (Alba et
al., 2017). Whilst the methodological issues and evidence for predictive performance and the clinical
impact in managing pressure ulcers have been evaluated extensively for heuristic tools of pressure ulcer
risk, this is not the case for empirically-derived models.

2. Objectives
This review was registered with the International Prospective Register of Systematic Reviews
(PROSPERO) (CRD42016042151) and follows the established systematic review methodology for
prognostic models (Debray et al., 2017). Its objectives are:

- To identify and describe available empirically-derived models for predicting pressure ulcer risk in
  any population (model development and validation) and describe the data sources used to develop
  these models;
- To assess the predictive performance of individual empirically-derived models across model
development and/or validation studies, and compare the performance of empirically-derived models
and heuristic tools where they have been compared in the same cohort;
- To evaluate the effects of empirically-derived models in reducing pressure ulcer risk in clinical
  practice (clinical impact).
3. Methods

3.1. Eligibility criteria

To investigate the development and validation of empirically-derived models (objective 1) and model performance and comparison (objective 2), we included longitudinal studies with the a priori intention of developing and/or validating a prognostic model (using multivariable analysis) for predicting pressure ulcer risk in any populations (Wolff et al., 2016). Longitudinal studies can reflect the prognostic relationship between factors of interest at baseline and a future outcome (Riley et al., 2013). Such studies can be retrospective in that outcomes have occurred and baseline data are collected from records that precede the event, or prospective in that participants have not had the outcome event when they enter the study. In both cases, longitudinal studies incorporate a temporal gap between baseline measures and outcome occurrence. We note that one study may include multiple longitudinal data sources, e.g., one for developing a model, and one or more for validating it. Eligible empirically-derived models had to combine at least two different variables (Altman, 2009). If multiple models were developed using a single data source we focused only on the final model(s). For model comparison, heuristic tools were also included only if they were compared with an eligible empirically-derived model in the same cohort. For investigation of the clinical impact of empirically-derived models (objective 3), we included before-after studies and randomised or non-randomised trials that evaluated the clinical effects of the models in reducing pressure ulcer incidence.

We excluded case-control studies, cross-sectional studies, case series, case reports, reviews, qualitative studies, comments, and animal studies (Steyerberg et al., 2013). Though they could be used to identify potential associations, a case-control design is at high risk of recall and selection bias. Cross-sectional studies were not included because prognostic factors and outcomes are measured simultaneously meaning the design is sub-optimal for determining the temporal relationship between the factor and outcomes (i.e., we cannot know which came first) (Altman, 2001; Riley et al., 2013). We also excluded studies in wheelchair users, in those undergoing flap coverage of pressure ulcers, and studies focusing on medical device-related ulcers (e.g., cervical collar-related ulcers).

3.2. Search strategy

We developed search strategies by combining pressure ulcer terms developed by Cochrane Wounds (McInnes et al., 2015) with validated database-specific prognosis search filters: (1) filters in Ingui and Rogers (2001) and Geersing (2012) for searching Ovid MEDLINE (1946 to 14 February 2017) (see Appendix 1); and (2) the filter of Walker-Dilks (2008) for searching EBSCO CINAHL Plus (1937 to 14 February 2017). All these search strategies were validated prior to a formal search. There was no restriction on the basis of language or publication status.
We also searched ProQuest (searched 14 February 2017) to identify potentially relevant doctoral theses in English and Chinese using the filter detailed by Wilczynski and Haynes (2004). Bibliography searches were undertaken using the reference lists of seven previously published systematic reviews.

3.3. Selection of studies

All citations were screened by two reviewers. One reviewer screened all citations using the title and abstract information. The search results were also divided into six batches with each batch independently screened against the eligibility criteria by one of six second reviewers (one reviewer per batch of citations). The full texts of potentially relevant studies were obtained. One reviewer inspected the full texts of all potentially eligible studies and a single second reviewer independently judged the eligibility of 10% of potentially eligible studies (random sample). Disagreements were resolved by discussion between the two reviewers and involvement of a third reviewer if necessary.

3.4. Data Extraction

One reviewer extracted data from all included studies using a pre-specified and piloted data extraction form. A second reviewer independently checked all data extraction. Any disagreements were resolved by discussion and, if necessary, with the involvement of a third reviewer. Where necessary, the authors of included studies were contacted to collect and/or clarify data.

The data extraction form included key information suggested by the CHChecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) (Moons, 2014): general information, data sources (i.e., study designs, participants, and follow-up period), study settings, participant details (number of participants, baseline skin status, average age in years, and gender), candidate predictors, outcomes to be predicted, sample sizes, model development methods, predictor selection strategy, missing data and methods for addressing missing data, methods for addressing continuous predictors, model validation, model comparisons, model performance measures and statistics, and authors’ conclusions.

Whilst statistics on possible performance measures were all extracted, we primarily focused on evaluating model discrimination using the c-statistic (ranging from 0.5 to 1, the closer to one indicating the better discrimination ability) and calibration using the total Observed:Expected risk (O:E) ratio (a value of one indicating a perfect calibration). Where c-statistic and total O:E ratio statistics were not reported, we calculated them where possible using methods suggested in Debray et al. (2017).

3.5. Risk of bias assessment

One reviewer applied the working version of the prognostic model risk of bias assessment tool (PROBAST) to assess risk of bias and applicability for each final model development–data source pair, as well as each model validation–data source pair described in the included studies (Wolff et al., 2016). Another reviewer was involved in checking all assessments independently. Any discrepancy between the two reviewers was resolved by discussion, and a third reviewer was involved where necessary.
The PROBAST tool considers five domains to judge the risk of biased estimates of predictive performance of a model in its target population: participant selection, predictors, outcome, sample size and participant flow, and analysis. We judged domains as being at low, unclear or high risk of bias. Using these assessments we made an overall risk of bias judgement for each model development–data source pair and model validation–data source pair separately. We gave an overall high risk judgement if more than one domain was high risk; as low if all domains were judged as low risk and as unclear if it was neither these cases (Wolff et al., 2016).

We also used the PROBAST tool to assess the applicability of each included model, that is, whether the model was applicable for its intended use in the target population (Moons et al., 2014). We assessed the extent to which a prognostic model matched our review question regarding the target population, predictors and outcome to be predicted (Wolff et al., 2016). For example, if a model included a final predictor (e.g., length of hospital stay) that can only be calculated after hospital discharge; because of this, the model has reduced applicability for predicting pressure ulcer risk in those not yet discharged from hospital.

3.6. Data synthesis

3.6.1. Description of data sources and models

We narratively summarised the characteristics of included (1) data sources (i.e., study designs, participants, and follow-up period), as well as (2) models (i.e., outcomes, candidate and final model predictors, sample sizes, the number of events, model development methods, predictor selection strategy, missing data, the handling of continuous predictors, model validation, and model presentation).

3.6.2. Predictive performance and model comparison

To evaluate the predictive performance of a specific model when assessed using different data sources we meta-analysed available c-statistics (and variances) for discrimination and the O:E ratio (and variances) for calibration. Analyses were conducted in STATA (Version 14.0, StataCorp, College Station, Texas). We present meta-analysis results (measured statistics with 95% confidence intervals (CIs)) for each model.

Before meta-analysis we explored clinical and methodological heterogeneity across data sources in terms of study design, participants, follow-up period, outcome, and risk of bias assessments. Meta-analysis was only conducted for a model when the various data sources used in its evaluation were considered clinically homogeneous. We also measured statistical heterogeneity using the $I^2$ measure (Higgins et al., 2003). Heterogeneity was regarded as low, moderate, substantial, or considerable if $I^2$ was less than 40%, 30% to 60%, 50% to 90%, or 75% to 100%, respectively (Higgins and Green, 2011).
We used a fixed-effect model if clinical heterogeneity was minimal and $I^2$ was less than 50%. If the clinical heterogeneity was judged as low but $I^2 \geq 75\%$, we used a random-effects model to evaluate the average performance, and we explored heterogeneity further via subgroup analysis, as detailed below.

We pre-specified subgroup analysis in terms of: study design (retrospective versus prospective designs), outcomes (pressure ulcers at grade II or above versus at any grade), study quality (e.g., high/unclear versus low risk of bias), and settings. If meta-analysis combined more than 10 data sources per independent variable (van Houwelingen et al., 2002) and with an $I^2 \geq 75\%$, we planned to perform meta-regression analysis following the methods in Siregar et al. (2012) to adjust for the above factors in each case.

When the predictive performance of two or more models was compared using a single data source, we narratively summarised each pairwise comparison between models, and presented corresponding discrimination and calibration evidence (Siontis et al., 2012).

3.6.3. Clinical impact evaluation
We planned to meta-analyse similar studies to evaluate the effects of specific models in reducing pressure ulcer incidence, where possible, following the meta-analytical method previously specified. If meta-analysis was impossible, we planned to summarise and present the relevant evidence narratively.

4. Results
4.1. Search results
We retrieved 6,990 citations from electronic searching and other search sources. Full-text screening of 553 potentially eligible studies led to the inclusion of 24 studies reporting model development, validation and/or comparison of 22 final prognostic models (Fig. 1; Anthony et al., 2000; Baldwin and Ziegler, 1998; Bergquist, 2001; Berlowitz et al., 1996, 2001a, 2001b; Borlawsky and Hripcsak, 2007; Compton et al., 2008; Corniello et al., 2014; DeJong et al., 2014; Hatanaka et al., 2008; Lu et al., 2017; Page et al., 2011; Papanikolaou et al., 2002; Perneger et al., 1998, 2002; Poss et al., 2010; Rose et al., 2006; Schoonhoven et al., 2002, 2005, 2006; Schue and Langemo, 1998; Slowikowski and Funk, 2010; Suriadi et al., 2008). The 24 studies used 28 different data sources.

We report on the development of 22 models (Anthony et al., 2000; Baldwin and Ziegler, 1998; Bergquist, 2001; Berlowitz et al., 1996, 2001a; Borlawsky and Hripcsak, 2007; Compton et al., 2008; Corniello et al., 2014; DeJong et al., 2014; Hatanaka et al., 2008; Lu et al., 2017; Page et al., 2011; Papanikolaou et al., 2002; Perneger et al., 1998, 2002; Poss et al., 2010; Rose et al., 2006; Schoonhoven et al., 2006; Schue and Langemo, 1998; Slowikowski and Funk, 2010; Suriadi et al., 2008); eight external validations of seven (of the 22) models (with two validations for one model) (Berlowitz et al.,
1996; Compton et al., 2008; Page et al., 2011; Poss et al., 2010; Schoonhoven et al., 2005; Suriadi et al., 2008); and 15 comparisons of 12 of the 22 models in comparisons with heuristic tools (Baldwin and Ziegler, 1998; Bergquist, 2001; Compton et al., 2008; Hatanaka et al., 2008; Papanikolaou et al., 2002; Rose et al., 2006; Schoonhoven et al., 2002; Slowikowski and Funk, 2010). We did not identify studies evaluating the clinical impact of models (Table 1).

4.2. Model development and validation

4.2.1. Summary of approaches to model development in included studies

We summarise details of the development of the 22 models and 20 data sources used in Table 2 (see Appendix Table 1 for details).

In total 50.0% (11/22) of the models were developed using prospective longitudinal data. The 22 models were developed using participant data from a variety of settings, with general hospitals (acute care) (seven models; Anthony et al., 2000; Hatanaka et al., 2008; Page et al., 2011; Papanikolaou et al., 2002; Perneger et al., 1998, 2002; Schoonhoven et al., 2006), long-term care (five models; Bergquist, 2001; Berlowitz et al., 1996, 2001a; Poss et al., 2010), and specific acute care settings (including intensive care units) (four models; Compton et al., 2008; Rose et al., 2006; Slowikowski and Funk, 2010; Suriadi et al., 2008) being the most frequent.

The 20 data sources contained data from 73,175 participants in total (median of samples 560, range 36 to 31,150). Eighteen data sources provided the following details: the average age of participants ranged from 31.8 to 82.5 years (median: 63.9); by gender, 43,115 participants being male and 26,375 female. In total, 54.5% (12/22) of models were developed in people without existing pressure ulcers (Anthony et al., 2000; Berlowitz et al., 2001a; Compton et al., 2008; Bergquist, 2001; Hatanaka et al., 2008; Papanikolaou et al., 2002; Perneger et al., 1998; Poss et al., 2010; Schoonhoven et al., 2006; Schue and Langemo, 1998; Suriadi et al., 2008). Follow-up duration was reported in 16 data sources and ranged from 6 to 180 days (median: 16). Across the 18 data sources where the number of outcome events was estimated (Anthony et al., 2000; Baldwin and Ziegler, 1998; Bergquist, 2001; Berlowitz et al., 1996, 2001a; Compton et al., 2008; Corniello et al., 2014; DeJong et al., 2014; Hatanaka et al., 2008; Lu et al., 2017; Page et al., 2011; Papanikolaou et al., 2002; Perneger et al., 1998, 2002; Poss et al., 2010; Schoonhoven et al., 2006; Schue and Langemo, 1998; Slowikowski and Funk, 2010; Suriadi et al., 2008), numbers of participants with incident ulcers ranged from 9 to 1,350 (median 94.5).

In total 22.7% (5/22) of the models were used for predicting pressure ulcer risk at the following specific time points from admission to the particular care setting: five days (Perneger et al., 2002), one week (Schoonhoven et al., 2006), three months (Berlowitz et al., 2001a; Hatanaka et al., 2008), and six months (Berlowitz et al., 1996). Of the models, 68.2% (15/22) were designed to predict the risk of
pressure ulcer at any grade, and 31.8% (7/22) were designed to assess the risk of pressure ulceration at grade II or above. The remaining 77.3% (17/22) models did not specify time points. Candidate predictors were specified in the development of 13 models (median number of predictors: 16, range 6 to 46) (Anthony et al., 2000; Bergquist, 2001; Corniello et al., 2014; DeJong et al., 2014; Hatanaka et al., 2008; Lu et al., 2017; Page et al., 2011; Papanikolaou et al., 2002; Perneger et al., 2002; Rose et al., 2006; Schoonhoven et al., 2006; Schue and Langemo, 1998). In the reports related to the 13 models, only two (Berlowitz et al., 1996; Slowikowski and Funk, 2010) specified that they considered pressure ulcer preventive measures (e.g., repositioning, and support surfaces) as candidate predictors in modelling (see Appendix Table 2).

Though all included studies reported the use of multivariable analysis methods (as per the inclusion criteria), one study (Rose et al., 2006) did not state the specific method used. Where reported logistic regression was the approach most frequently applied (52.4%, 11/21) (Anthony et al., 2000; Baldwin and Ziegler, 1998; Berlowitz et al., 1996, 2001a; Compton et al., 2008; Corniello et al., 2014; DeJong et al., 2014; Hatanaka et al., 2008; Lu et al., 2017; Page et al., 2011; Papanikolaou et al., 2002; Perneger et al., 2002; Poss et al., 2010; Schoonhoven et al., 2006; Schue and Langemo, 1998; Slowikowski and Funk, 2010; Suriadi et al., 2008). To determine final model predictors during multivariable analysis, stepwise selection methods (e.g., forward method) were commonly used (40.9%, 9/22) (Anthony et al., 2000; Baldwin and Ziegler, 1998; Bergquist, 2001; Compton et al., 2008; Lu et al., 2017; Page et al., 2011; Schue and Langemo, 1998; Slowikowski and Funk, 2010); the full model approach was used for predictor selection in 27.3% (6/22) of the models (DeJong et al., 2014; Hatanaka et al., 2008; Papanikolaou et al., 2002; Perneger et al., 1998, 2002; Suriadi et al., 2008). In total there were 55 distinct final model predictors used across models (a median of 4 predictors per model), of which 31 predictors were included in only one model. Of the two models involving preventive measures as candidate predictors, only Berlowitz et al. (1996) included them as a final predictor (see Appendix Table 2).

No information on missing data was reported for 68.2% (15/22) of model developments, therefore we assumed that a complete case analysis had been used (Anthony et al., 2000; Baldwin and Ziegler, 1998; Berlowitz et al., 1996, 2001b; Borlasky and Hripcsak, 2007; Compton et al., 2008; DeJong et al., 2014; Hatanaka et al., 2008; Lu et al., 2017; Perneger et al., 1998, 2002; Poss et al., 2010; Rose et al., 2006; Schoonhoven et al., 2002, 2005; Schue and Langemo, 1998; Slowikowski and Funk, 2010; Suriadi et al., 2008). Imputation of missing data was reported in the development of one model (using Multiple Imputation using Chained Equations) (Corniello et al., 2014). Of the 14 models including continuous predictors, five categorised or dichotomised continuous data (Page et al., 2011; Perneger et al., 2002; Schoonhoven et al., 2006; Slowikowski and Funk, 2010; Suriadi et al., 2008); three used smoothing curves via the local regression (LOESS) function (Berlowitz et al., 2001a), restricted cubic spline (Corniello et al., 2014), and logarithm transformation (Hatanaka et al., 2008).
Data on events per variable (EPV) were available for 20 models (EPVs ranging from 1.5 to 27.2, median 8.5), of which 12 had fewer than 10 EPVs, representing inadequate sample sizes and a possibility of over-fitting (i.e., when a model fits the specific data set used to generate it extremely well but is not adequately prognostic when applied to other data) (Anthony et al., 2000; Baldwin and Ziegler, 1998; Bergquist, 2001; Compton et al., 2008; Corniello et al., 2014; Hatanaka et al., 2008; Lu et al., 2017; Page et al., 2011; Papanikolaou et al., 2002; Perneger et al., 2002; Schoonhoven et al., 2006; Schue and Langemo, 1998). To avoid possible over-fitting, the development of 22.7% (5/22) of the models employed additional statistical techniques (e.g., bootstrapping) (Anthony et al., 2000; Borlawsky and Hripcsak, 2007; Corniello et al., 2014; Perneger et al., 2002; Schoonhoven et al., 2006).

In total 72.7% (16/22) model development processes specified how final models were presented, with 22.7% (5/22) building new scoring systems based on regression coefficients (Page et al., 2011; Perneger et al., 2002; Poss et al., 2010; Schoonhoven et al., 2006; Suriadi et al., 2008); 4.5% (1/22) developing a nomogram scale using the method of Steyerberg et al. (2010) (Corniello et al., 2014); 4.5% (1/22) building a nomogram scale using the method of Zlotnik et al. (2015) (Lu et al., 2017); and 13.6% (3/22) using formula equations (Anthony et al., 2000; Hatanaka et al., 2008; Slowikowski and Funk, 2010).

4.2.2. PROBAST assessment for model development

As a result of the above issues, using the PROBAST tool, we judged the overall risk of biased estimates of predictive performance to be unclear for two models (Berlowitz et al., 2001a; Poss et al., 2010) and high for the remaining 20 models (Anthony et al., 2000; Baldwin and Ziegler, 1998; Bergquist, 2001; Berlowitz et al., 1996; Borlawsky and Hripcsak, 2007; Compton et al., 2008; Corniello et al., 2014; DeJong et al., 2014; Hatanaka et al., 2008; Lu et al., 2017; Page et al., 2011; Papanikolaou et al., 2002; Rose et al., 2006; Perneger et al., 1998, 2002; Schoonhoven et al., 2006; Schue and Langemo, 1998; Slowikowski and Funk, 2010; Suriadi et al., 2008) (see Table 1). In summary, the reasons for downgrading were mainly: inappropriate definitions or measurements of candidate predictors, non-validated measurements of outcomes (e.g., self-reported pressure ulcer outcome), inadequate sample size (e.g., EPV < 10), and inappropriate analysis (e.g., sub-optimal approaches to selecting predictors, poor handling of missing data, sub-optimal approaches to analysing continuous predictors, and high risk of over-fitting).

Only two of the 22 models matched the review question in terms of target populations, predictors and outcomes of interest and thus were judged as applicable (Perneger et al., 2002; Schoonhoven et al., 2006). The applicability was judged as unclear for 16 models (Anthony et al., 2000; Bergquist, 2001; Berlowitz et al., 1996, 2001a; Borlawsky and Hripcsak, 2007; Compton et al., 2008; Corniello et al., 2014; DeJong et al., 2014; Page et al., 2011; Papanikolaou et al., 2002; Perneger et al., 1998; Rose et al., 2006; Schue and Langemo, 1998; Slowikowski and Funk, 2010; Suriadi et al., 2008) and low for
the remaining four (Baldwin and Ziegler, 1998; Hatanaka et al., 2008; Lu et al., 2017; Poss et al., 2010), mainly because of the use of inappropriate predictors, and vague definitions of predictors and outcomes (see Appendix Table 3).

4.2.3. Summary of model validation approaches in included studies

We summarise details of the validation of seven models using eight data sources in Table 3 (see Appendix Table 1 for details).

Of the eight data sources, 37.5% (3/8) used a prospective design. The eight data sources contained 119,835 participants (median of samples 7,251, range 165 to 73,183); the median of reported average participant age (per study) was 67.5 years (range 51.3 to 71). There were more male participants (n = 17,834) than female (n = 859). Five data sources provided details on follow-up duration, with a median of 15 days (range 8 to 180). Numbers of participants with incident ulcers was estimable in seven data sources (Berlowitz et al., 1996, 2001b; Compton et al., 2008; Page et al., 2011; Poss et al., 2010; Suriadi et al., 2008), ranging from 7 to 1,903 (median 556). Of the eight data sets, 37.5% (3/8) had fewer than 100 events (Compton et al., 2008; Page et al., 2011; Suriadi et al., 2008).

In total, 87.5% (7/8) of validation exercises used logistic regression. None of the eight validations reported missing data issues. The authors of Berlowitz et al. (2001b) concluded that their model required recalibration because the predicted pressure ulcer risk using the model had poor agreement with the observed risk; however, the model was not recalibrated.

4.2.4. PROBAST assessment for the model validation

Using the PROBAST tool, we judged four of the eight validations as being unclear in terms of overall risk of bias (Berlowitz et al., 1996; Poss et al., 2010; Schoonhoven et al., 2005), and the remaining four as being at high risk of bias (Berlowitz et al., 2001b; Compton et al., 2008; Page et al., 2011; Suriadi et al., 2008).

There was insufficient information to judge the applicability for six of the eight validations (Berlowitz et al., 1996, 2001b; Compton et al., 2008; Page et al., 2011; Schoonhoven et al., 2005; Suriadi et al., 2008) and the remaining two were not likely to be applicable (Poss et al., 2010) (see Table 1). The main reasons for downgrading judgements were the same as the PROBAST assessment for model development.

4.3. Predictive performance of empirically-derived models

4.3.1. Discrimination of empirically-derived models

C-statistics (measuring model discrimination, c-statistic = 1.00 if a model predicts risk perfectly and c-statistic = 0.50 if the model predicts risk no better than chance) were available for 12 of the 22 (54.5%)
models (using a total of 17 data sources). Five models had discrimination data from two or more data sources (Berlowitz et al., 1996, 2001a, 2001b; Compton et al., 2008; Page et al., 2011; Poss et al., 2010); one model had one c-statistic for its development and another for one internal validation based on the same data source (Perneger et al., 2002); and we meta-analysed these data leading to six separate meta-analyses for six models. We also present discrimination results for the remaining seven models from single data sets (see Table 4).

* CI: Confidence interval; † Dev: model development; Val: model validation

The c-statistics across the 12 models ranged from 0.65 to 0.89 (with a median of 0.78; eight of 12 (66.7%) models had c-statistics greater than 0.75). The c-statistic, for example, of the Fragmment scale implied that, if we ran a hypothetical study and randomly chose pairs of participants, the pressure ulcer risk predicted by the model would be higher in participants who went on to develop pressure ulcers than in those who would not for 79 (95% CI 77 to 82) of 100 chosen pairs.

No formal subgroup analysis was conducted because neither clinical nor methodological heterogeneity were noted in the six analyses with two or more data in each case (Berlowitz et al., 1996, 2001a, 2001b; Compton et al., 2008; Page et al., 2011; Perneger et al., 2002; Poss et al., 2010). Considering findings by settings two models aimed for use in intensive care unit (ICU) settings had higher c-statistics (0.81 and 0.89) than models for other settings. Models judged at unclear risk of bias had a lower c-statistics (0.65, 0.73 and 0.75 for three distinct models) than those with high judgement.

4.3.2. Calibration of empirically-derived models

We evaluated the calibration abilities of five models using total O:E ratios (a ratio of actual risk to predicted risk, indicating a perfect calibration when it equals to one; Berlowitz et al., 1996, 2001a, 2001b; Borlawsky and Hripcsak, 2007; Lu et al., 2017; Schoonhoven et al., 2006). Of the five models, Borlawsky and Hripcsak (2007) had a seemingly incorrect O:E ratio (of 206:18) and was excluded from further analysis. Two models individually had enough data to pool and two corresponding meta-analyses for total O:E ratios were conducted (Berlowitz et al., 1996, 2001a, 2001b) (see Table 4). We also presented O:E ratios from single studies for the remaining two models (Lu et al., 2017; Schoonhoven et al., 2006).

Analyses showed that three of the four models were well calibrated: the prePURSE study tool (total O:E ratio 1.00, 95% CI 0.84 to 1.19; high risk of bias); the Berlowitz 11-item model (0.99, 95% CI 0.95 to 1.04; unclear risk of bias); and the cardiovascular surgical PU nomogram score (1.00, 95% CI 0.76 to 1.32; high risk of bias). The Berlowitz MDS risk-adjustment model was poorly calibrated (total O:E ratio 0.94, 95% CI 0.88 to 1.01; unclear risk of bias). Again, no subgroup analysis was conducted for the same reasons as above.
No meta-analysis was performed for any of the other reported calibration measure (e.g., Hosmer-Lemeshow and/or Pearson Chi-squared test statistics, likelihood ratio test statistics) as corresponding meta-analytical methods are not available (see Appendix Table 1 for results). We note that all original authors claimed that corresponding models were well calibrated.

4.3.3. Model performance of empirically-derived models in comparisons with heuristic tools

There were fifteen pairwise comparisons comparing 12 different empirically-derived models with various heuristic tools (Baldwin and Ziegler, 1998; Bergquist, 2001; Compton et al., 2008; Hatanaka et al., 2008; Papanikolaou et al., 2002; Perneger et al., 1998, 2002; Rose et al., 2006; Schoonhoven et al., 2002; Schue and Langemo, 1998; Slowikowski and Funk, 2010) (Table 1). Eleven of the 15 comparisons (covering eight different data driven models) reported model discrimination and/or calibration data: all eight empirically-derived models (i.e., Compton ICU model, clinical laboratory data-based predictive equation, empirically-derived simplified Waterlow model, Fragment scale, prePURSE study tool, Bergquist two-item Braden scale, the Bergquist three-item Braden scale, and surgical ICU pressure ulcer risk assessment scale) had better discrimination and/or calibration abilities than heuristic tools (see Appendix Table 4).

4.4. Clinical impact evaluation

We did not identify any studies evaluating the clinical impact of empirically-derived models.

5. Discussion

5.1. Main findings

We have conducted the first systematic review to identify, synthesise and summarise all available evidence on the characteristics and predictive performance of available empirically-derived prognostic models for predicting pressure ulcer risk.

We found 22 empirically-derived models for pressure ulcer risk assessment and whilst these models were reported to have better performance than the Braden, Norton, and Waterlow scales, most (15 of 22) had not been externally validated. The reported c-statistics (for 12 models) ranged from 0.65 to 0.89, and total O:E ratios were between 0.94 and 1.00. The modelling process of most models (20 of 22) was judged to be of low quality, mainly due to inappropriate definitions or measurements of candidate predictors, invalidated measurements of outcomes, insufficient sample size, and inappropriate analysis. In particular, models with unclear risk of bias appeared to have lower discrimination ability than those with high risk of bias in the same setting. No research was identified that evaluated the impact of specific models on reducing pressure ulcer incidence.

5.2. Common issues found in model development approaches

Effective pressure ulcer preventive measures (e.g., repositioning and support surface use) can reduce pressure ulcer incidence (McInnes et al., 2015; Shi et al., 2018), that is, modify pressure ulcer risk at
baseline. Therefore, preventive measures that are planned or delivered at baseline should be considered as predictors when developing a model to ensure reliable model performance (Groenwold et al., 2016; Defloor and Grypdonck, 2004). This review highlights that such consideration within modelling is seldom undertaken: part of a wider problem with the lack of appropriate consideration of candidate predictors to maximise unbiased model performance. Omission of preventive measures during the development of a prognostic model may also affect further model validation. To illustrate, because the use of a pressure ulcer prognostic model may result in the provision of preventive measures, if these are effective then individuals have a reduced risk of ulcer development. Without explicit consideration of the preventive measures in the model it is difficult to evaluate whether the model correctly identifies individuals at risk or not.

The included models were designed to predict the risk of pressure ulcers of (a) any grades or (b) grade 2 or above. Models did not differentiate superficial ulcers (i.e., grade 1 or 2) from deep ulcers (grade 3 or 4). Whilst pathophysiological evidence suggests that the pathway from healthy skin to superficial ulcers is probably different from that for deep ones (Bouten et al. 2003), the current lack of epidemiological evidence on specific predictors of superficial and deep ulcers limits further exploration this issue.

In terms of statistical analysis, the high risk of bias inherent in the development and validation of most of the models identified resulted from sub-optimal approaches to predictor selection, poor handling of missing data and continuous predictors, small sample sizes and high risk of over-fitting.

In this review, univariate analysis was usually used to identify predictors for inclusion in multivariable analysis. This strategy is regarded as sub-optimal for model development because it is associated with high risk of rejecting important predictors which may appear insignificant in univariate analysis but could explain pressure ulcer risk at the multivariable analysis stage (i.e., predictor selection bias) (Collins et al., 2011). To avoid this bias, multiple sources of information (e.g., expert opinion, literature evidence) are recommended for a priori selection of predictors at this stage (Kattan and Harrell, 2016; Justice et al., 1999) however only two models were developed using this approach (Berlowitz et al., 2001a; Schoonhoven et al., 2006).

Predictor selection bias may also occur in multivariable analysis if sub-optimal selection strategies are applied. For example, the forward selection approach may lead to the exclusion of potentially important predictors because it does not simultaneously examine the effects of all candidate predictors (Moons et al., 2012). Steyerberg et al. (2017) have recently shown that the commonly used selection methods cannot select a consistent set of final model predictors if the model development process is repeated in simulations, and they could produce exaggerated predictor associations. To avoid predictor selection bias in multivariable analysis, the full model approach could be used after a set of promising
predictors has been determined (Steyerberg et al., 2017; Kattan and Harrell, 2016); an approach which was applied in the development of six of the 22 models included in this review.

As with other types of epidemiological research, missing data are likely to be a common issue in model development and validation studies but details are often incompletely reported (van der Heijden et al., 2006). Because of this, for this review we were required to assume that complete case analysis had been conducted in 15 of 22 model developments and all eight external validations. In the presence of missing data, complete case analysis may produce biased estimates of prognostic associations because of a decrease in the effective sample sizes (Donders et al., 2006). Ideally, advanced techniques (e.g., multiple imputations) should be used but this was seldom the case for the model developments included in this review (Donders et al., 2006). This is consistent with findings of prognostic model reviews in other clinical areas, for example for predicting Type-2 diabetes risk (Collins et al., 2011).

The handling of continuous predictors is another problematic area in multivariable analysis. Of the available methods, categorising (or dichotomising) continuous predictors leads to the loss of information and statistical power (Steyerberg et al., 2017) however the use of specific cut-off points for categorising is also associated with biased estimates (Steyerberg et al., 2017; Royston et al., 2006). The limitations associated with certain approaches to handling continuous predictors have been acknowledged for many years but these approaches were employed by half the studies involving continuous predictors included in this review. It is widely suggested that the continuity of continuous data should be retained, for example by using regression splines or LOESS functions (Royston and Altman, 1994). Williams et al. (2016) modelled continuous predictors using restricted cubic spline transformations to allow for nonlinear associations. In this review there were two cases where model development was supported by the appropriate transformation of continuous predictors using smoothing curves (Berlowitz et al., 2001a), or restricted cubic spline (Corniello et al., 2014).

Small sample sizes are a recurring problem in model development and validation and can lead to biased effect estimates, poor predictive performance, and over-fitting (Steyerberg et al., 2017). Whilst larger sample sizes can mitigate most of these issues, over-fitting remains a risk as it can also occur when sample sizes are adequate but when highly data-dependent predictor selection methods are used (Harrell, 2015). Statistical methods (e.g. bootstrapping) are commonly suggested to avoid over-fitting and improve model validity (Altman, 2009). However, the application of statistical methods to ensure appropriate model fit forms only part of model validity assessment. It is also necessary to know whether the model will perform well for new populations/settings; therefore external validation is required (Altman, 2009). In this review, 13 of 22 model developments had inadequate sample sizes and just one third (5/12) applied additional analysis to address over-fitting. Only one of the five models was reported to consider an external validation (Schoonhoven et al., 2005).
5.3. **Strengths in comparison with previous reviews and limitations of this review**

A comprehensive search identified seven previous, related systematic reviews (García-Fernández et.
al., 2014; He et al., 2012; Kottner et al., 2009; Kottner et al., 2013; Moore and Cowman, 2014; Mortenson et.
al., 2008; Pancorbo-Hidalgo et al., 2006). However the previous reviews focused only on heuristic tools. By focusing on empirically-derived models, our review adds to the current evidence base to provide a complete picture on the evidence for pressure ulcer risk prediction models. We found a variety of prognostic models with little validation. In addition, previous reviews reported that the available heuristic tools had poor discrimination while our review suggests that empirically-derived tools perform better than the heuristic tools.

For this first systematic review of empirically-derived prognostic models for pressure ulcer risk we employed the PROGnosis RESearch Strategy (PROGRESS) framework for synthesising prognosis evidence (Steyerberg et al., 2013; Debray et al., 2017). Firstly, we defined a specific question for prognostic topics using the CHARMS checklist (Moons et al., 2014), and considered only longitudinal studies as Steyerberg et al. (2013) suggested. Previous reviews focusing on heuristic tools (e.g., Defloor and Grypdonck (2004)) also included cross-sectional studies.

Secondly, the work is also strengthened by the use of established search filters to identify prognostic model studies; this ensures a comprehensive literature search. Despite this, we noted that it is still a challenge to identify prognostic model studies comprehensively and a great deal of sifting of irrelevant material is required. Geersing et al. (2012) stated that no single search filter (including those used in this review) identifies model studies adequately. Because of this, we used multiple filters and databases to conduct a search that would identify the maximum potentially relevant citations. However in order to manage this search within available resources we had to exclude some databases, e.g., Embase. Interestingly previous reviews identified few empirically-derived models, perhaps due to poor search strategies. Thirdly, we based our data extraction on the CHARMS checklist (Moons et al., 2014). We also used the PROBAST tool for a formal prognostic model risk of bias assessment (Wolff et al., 2016), whereas previous reviews did not use structured, prognostic model-specific, appraisal tools. Finally, we followed the methodology of Debray et al. (2017) for data synthesis, evaluated two important prognostic model properties (calibration and discrimination) (Alba et al., 2017), and considered important characteristics (settings and risk of bias) to explain findings. Previous reviews, for example Defloor and Grypdonck (2004), mostly evaluated diagnostic accuracy measures (e.g., sensitivity) and did not incorporate risk of bias or quality/certainty assessment into the interpretation of findings. Additionally, the use of diagnostic accuracy methods and measures has important limitations when considering pressure ulcer risk prediction since diagnostic methods and outputs are intended to evaluate how well a tool discriminates people with a disease/condition status (clearly defined by using a reference standard) from those without the status (Cook, 2008). As there is no approved reference standard for defining the “true” risk of pressure ulcer strictly we cannot measure diagnostic accuracy (Kottner and Balzer, 2010).
The presence of a pressure ulcer is commonly used as an alternative reference standard however really this can only reflect the stochastic nature of pressure ulcer risk in longitudinal studies rather than in diagnostic accuracy studies (Cook, 2008).

Our review has some limitations. Firstly although we conducted very wide searches resulting in the assessment of 553 full texts we cannot rule out the possibility that we missed potentially eligible studies. Additionally, we excluded six potentially eligible studies published in languages other than English and Chinese and 45 studies with hard-to-access full text which may have led to publication bias. We were unable to formally assess publication bias using funnel plots because we had fewer than ten included studies in meta-analyses of specific models. Despite the number of unavailable studies (a large proportion of which we are likely to have excluded), our findings are probably credible: prognostic models are frequently developed but seldom validated, and their development process is usually flawed. These methodological problems with prognostic models occur in many fields e.g., models for predicting cardiovascular disease and Type 2 diabetes (Damen et al., 2016; Collins et al., 2011).

Secondly, whilst many studies have examined heuristic tools, a summary of this evidence was beyond the scope of our review. Thirdly, as the use of meta-analytic techniques improves the evaluation of average performance measures across studies, we aimed to maximally use the available data. In some cases this required further manipulation of reported data using recommended approaches (Debray et al., 2017) in order to produce the estimate type required. We recognise that such statistics may not be exactly the same as the true data (Debray et al., 2017).

Finally, we did not focus in detail on which information sources (e.g., electronic population registry database, clinical records) were used as data sources for model development or validation. We acknowledge that data from different sources may be of different quality. For example, data from population registry and clinical records may be of poorer quality than bespoke data collected by trained researchers and the data collected specifically for research might not reflect routine care. Data quality was however considered as part of the study quality appraisal.

5.4. Recommendations and implications

There are many potential predictors for pressure ulcer development (Coleman et al., 2013) meaning future model research may not need to identify new predictors but rather develop good quality models using a small set of promising predictors determined by literature review and clinical relevance. When a new model is developed, researchers should consider pressure ulcer prevention measures that are planned or delivered at baseline as potential candidate predictors (Groenwold et al., 2016), apply the appropriate predictor selection strategies, appropriately handle missing data and continuous predictors, and avoid over-fitting. Future research should pay equal attention to development and validation. Two models judged in this review as applicable (the Fragmment scale and the prePURSE study tool) could be validated in the future. The CHARMS checklist should be fully complied with to ensure the complete
reporting of candidate predictors, outcomes, the model development process, and the regression coefficients of included predictors (Moons et al., 2014), which determine the applicability of models. Finally, because we have only considered empirically-derived tools, high-quality evidence on heuristic tools evidence is still required.

In terms of clinical practice, international clinical practice guidelines generally recommend considering risk assessment scales as an aid to clinical judgement regarding pressure ulcer risk (NPUAP/EPUAP/PPPIA, 2014; NICE, 2014). However, only 27% to 32% of nurses routinely applied prediction models (mainly heuristic tools) in their risk assessment practice (Samuriwo and Dowding, 2014). This contradiction could be explained by main findings of the previous eight systematic reviews: heuristic tools often have poor performance in predicting pressure ulcer risk. Additionally, there is no clear evidence that heuristic tools improve patient clinical outcomes in one Cochrane review: Moore and Cowman (2014) reviewed two randomised controlled trials (Saleh et al., 2009; Webster et al., 2011) of comparing heuristic tool-aided risk assessment with clinical judgement alone – both followed with the provision of preventive measures – in reducing pressure ulcer incidence. Whilst we identified abundant empirically-derived tools and found these tools performed better than the heuristic tools, the extent to which these empirically-derived tools are currently used in practice is not clear. We did not find research evaluating the clinical impact of the tools in reducing pressure ulcer risk. As with other risk assessment tools there is uncertainty about the value of these tools in clinical practice (Kottner and Balzer, 2010).

6. Conclusions

Many prognostic models have been developed for predicting the risk of pressure ulcer development, but the methods by which most models were developed are flawed by current standards and consequently model performance cannot be regarded as satisfactory despite such claims in the original studies. Most models have not been validated for use in new populations and have seldom been compared with alternative heuristic models, and their clinical effects are still unknown. Therefore new models should be developed following international methodological standards for model development and validation. The Fragment scale and the prePURSE tool could be validated as they are judged to have low applicability concern.

Acknowledgements

This research was funded by the President’s Doctoral Scholar award of the University of Manchester (CS) and supported by the NIHR Manchester Biomedical Research Centre. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
The authors would like to acknowledge the contribution of Ross Atkinson, Maggie Westby, and Gillian Norman who conducted independent screening of search results for eligible studies, as well as Zhenmi Liu who carried out independent study selection and independent data checking.
References


23


Kattan, M.W., Harrell, F.E., 2016. We should not be so quick to abandon the use of domain experts and full models (letter commenting: J Clin Epidemiol. 2015;71C:76–85.). J Clin Epidemiol 75, 131.


StataCorp. 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP.


Fig 1. PRISMA flow chart of study selection
13 of additional records (de-duplication of electronic database search results) identified through other sources (Wounds group meeting, inspecting reference lists of eight previous systematic reviews)

Records (n = 6,990) (including 552 duplicates)

6,348 records screened (de-duplication)

5,795 records excluded

529 full-text articles excluded, with reasons:
- duplicates (n = 54),
- reviews, introductory publications, comments, letters, or erratum (n = 20),
- studies with ineligible participants (participants undergoing flap coverage of ulcers) (n = 1),
- studies with ineligible outcomes or unit of analysis (e.g., pressure ulcer incidence as a part of a composite outcome, pressure ulcer prevalence, medical device-related pressure ulcer) (n = 49),
- studies using ineligible study designs (e.g., case control, case series design) (n = 25),
- studies with multivariable analysis but without a pre-specified intention to develop prognostic models (n = 181),
- studies without multivariable analysis (n = 118),
- studies solely focusing on heuristic models (n = 15),
- irrelevant topics (e.g., nurses’ assessments on psychometrics or reliability of a model) (n = 15),
- studies awaiting for further decisions and in non-English or non-Chinese (n = 6),
- awaiting for authors’ clarification (n = 1),
- and failure to retrieve full texts (n = 45)

563 full-text articles assessed for eligibility

24 studies included in the systematic review
Table 1 Summary of empirically-derived models included in this review

<table>
<thead>
<tr>
<th>Models by settings (numbers of models)</th>
<th>Development</th>
<th>Validation</th>
<th>Comparisons</th>
<th>c-statistics</th>
<th>O:E ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data source</td>
<td>Risk of bias/ applicability</td>
<td>Data source</td>
<td>Risk of bias/ applicability</td>
<td>Numbers of comparisons – heuristic tools in comparisons (data source)</td>
</tr>
<tr>
<td>General (acute care) hospital settings (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-item Waterlow logistic model</td>
<td>Dev (Anthony et al., 2000)</td>
<td>High/unclear</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical laboratory data-based equation</td>
<td>Dev (Hatanaka et al., 2008)</td>
<td>High/high</td>
<td></td>
<td>One – Braden (Hatanaka et al., 2008)</td>
<td>One (Dev)</td>
</tr>
<tr>
<td>TNH-PUPP model</td>
<td>Dev (Page et al., 2011)</td>
<td>High/unclear</td>
<td>Val (Page et al., 2011)</td>
<td>High/unclear</td>
<td>Two (Dev–Val)</td>
</tr>
<tr>
<td>Papanikolaou simplified Waterlow scale</td>
<td>Dev (Papanikolaou et al., 2002)</td>
<td>High/unclear</td>
<td></td>
<td>One – Waterlow (Papanikolaou et al., 2002)</td>
<td></td>
</tr>
<tr>
<td>Perneger simplified Norton scale</td>
<td>Dev (Perneger et al., 1998)</td>
<td>High/unclear</td>
<td></td>
<td>One – Norton (Perneger et al., 1998)</td>
<td></td>
</tr>
<tr>
<td>Fragment scale</td>
<td>Dev (Perneger et al., 2002)</td>
<td>High/low</td>
<td>Val (Schoonhoven et al., 2005)</td>
<td>Unclear/unclear</td>
<td>Three – Norton/ Braden/ Waterlow (Schoonhoven et al., 2002)</td>
</tr>
<tr>
<td>prePURSE study tool</td>
<td>Dev (Schoonhoven et al., 2006)</td>
<td>High/low</td>
<td>Val (Schoonhoven et al., 2005)</td>
<td>Unclear/unclear</td>
<td>Three – Norton/ Braden/ Waterlow (Schoonhoven et al., 2002)</td>
</tr>
<tr>
<td>Specific acute care (e.g., ICU settings) (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compton ICU model</td>
<td>Dev (Compton et al., 2008)</td>
<td>High/unclear</td>
<td>Val (Compton et al., 2008)</td>
<td>High/unclear</td>
<td>One – Waterlow (Compton et al., 2008)</td>
</tr>
<tr>
<td>Rose ICU scale</td>
<td>Dev (Rose et al., 2006)</td>
<td>High/unclear</td>
<td></td>
<td></td>
<td>One – Braden (Rose et al., 2006)</td>
</tr>
<tr>
<td>Surgical ICU risk assessment scale</td>
<td>Dev (Slowikowski and Funk, 2010)</td>
<td>High/unclear</td>
<td></td>
<td></td>
<td>One – Braden (Slowikowski and Funk, 2010)</td>
</tr>
<tr>
<td>Suriadi and Sanada (S.S.) scale</td>
<td>Dev (Suriadi et al., 2008)</td>
<td>High/unclear</td>
<td>Val (Suriadi et al., 2008)</td>
<td>High/unclear</td>
<td>One (Dev)</td>
</tr>
<tr>
<td>Setting</td>
<td>Model Description</td>
<td>Dev (Year)</td>
<td>Val (Year)</td>
<td>Setting Description</td>
<td>Outcome</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>------------------------------------</td>
<td>--------------------------------</td>
<td>--------------------------------</td>
<td>--------------------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>Cardiovascular surgery settings (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vascular surgery PU risk score</td>
<td>Dev (Corniello et al., 2014)</td>
<td>High/unclear</td>
<td></td>
<td>One (Dev)</td>
<td></td>
</tr>
<tr>
<td>Cardiovascular surgical PU score</td>
<td>Dev (Lu et al., 2017)</td>
<td>High/high</td>
<td></td>
<td>One (Dev)</td>
<td>One (Dev)</td>
</tr>
<tr>
<td>Trauma and burn centres (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baldwin two-item Braden scale</td>
<td>Dev (Baldwin and Ziegler, 1998)</td>
<td>High/high</td>
<td></td>
<td>One – Braden (Baldwin and Ziegler, 1998)</td>
<td></td>
</tr>
<tr>
<td>Long-term care settings (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergquist two-item Braden scale</td>
<td>Dev (Bergquist, 2001)</td>
<td>High/unclear</td>
<td></td>
<td>One – Bergquist (Bergquist, 2001)</td>
<td></td>
</tr>
<tr>
<td>Bergquist three-item Braden scale</td>
<td>Dev (Bergquist, 2001)</td>
<td>High/unclear</td>
<td></td>
<td>One – Bergquist (Bergquist, 2001)</td>
<td></td>
</tr>
<tr>
<td>Berlowitz 11-item model</td>
<td>Dev (Berlowitz et al., 1996)</td>
<td>High/unclear</td>
<td>Val (Berlowitz et al., 1996)</td>
<td>Unclear/unclear</td>
<td>Two (Dev–Val)</td>
</tr>
<tr>
<td>Berlowitz MDS risk-adjustment model</td>
<td>Dev (Berlowitz et al., 2001a)</td>
<td>Unclear/unclear</td>
<td>Val (Berlowitz et al., 2001b)</td>
<td>High/unclear</td>
<td>Two (Dev–Val)</td>
</tr>
<tr>
<td>interRAI PURS</td>
<td>Dev (Poss et al., 2010)</td>
<td>Unclear/high</td>
<td>Val 1 (Poss et al., 2010)</td>
<td>Unclear/high</td>
<td>Three (Dev–Val)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Val 2 (Poss et al., 2010)</td>
<td>Unclear/high</td>
<td></td>
</tr>
<tr>
<td>Rehabilitation units (2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission PU-FIM model</td>
<td>Dev (DeJong et al., 2014)</td>
<td>High/unclear</td>
<td></td>
<td></td>
<td>One (Dev)</td>
</tr>
<tr>
<td>Schue modified Braden scale</td>
<td>Dev (Schue and Langemo, 1998)</td>
<td>High/unclear</td>
<td></td>
<td>One – Braden (Schue and Langemo, 1998)</td>
<td></td>
</tr>
<tr>
<td>Unclear setting (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic pressure ulcer prediction</td>
<td>Dev (Borlawsky and Hripcsak, 2007)</td>
<td>High/unclear</td>
<td></td>
<td></td>
<td>One (Dev)</td>
</tr>
<tr>
<td>Summary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Models by settings (References)</td>
<td>Data sources and participants (total number, average age in years, gender in male/female, and baseline skin status)</td>
<td>Outcomes (pressure ulcer grade and timing of risk prediction)</td>
<td>Follow-up duration/length of stay</td>
<td>Candidate predictors</td>
<td>Modelling methods</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>General (acute care) hospital settings (n = 7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-item Waterlow logistic model (Anthony et al., 2000)</td>
<td>Prospective; 422; mean 64.83 (SD 17.85); 200/222; no existing pressure ulcers</td>
<td>Any grade</td>
<td>14 days</td>
<td>11</td>
<td>Logistic regression and artificial neural network</td>
</tr>
<tr>
<td>Clinical laboratory data-based equation (Hatanaka et al., 2008)</td>
<td>Prospective; 149; mean 71.6 (SD 11.3); 104/45; no existing pressure ulcers</td>
<td>Grade II or above (predicting three-month risk)</td>
<td>3 months</td>
<td>15</td>
<td>Logistic and Cox regression</td>
</tr>
</tbody>
</table>

* Dev = data for model development; Val = data for model validation; ¶ These columns show the number of data sources containing c-statistics (or O:E ratios) for specific models. For example, the “Two (Dev–Val)” means there are two data sources (one for model development, and another for validation) having c-statistics of the “TNH-PUPP model.”

Table 2: Summary of 22 included model development studies
| Score system | Study Tool | Setting (n = 4) | Study Tool | Setting (n = 4) | Model (Page et al., 2011) | Retrospective; 342; mean 63 (19.82); 156/186; not reported | Grade II or above | Mean 15.42 (SD 22.29) days | 12 | Logistic regression | Univariate analysis (p = 0.10) | Stepwise (backward) | None missing | Categorising or dichotomising | 67 (EPV <10) | Not reported | Score system |
|--------------|------------|----------------|------------|----------------|--------------------------|-------------------------------------------------------------|-------------------|----------------------------|----------------|-----------------------------|----------------|-----------------------------|----------------|----------------|----------------|
| Papanikolaou | simplified | Waterlow scale | (Papanikolaou | Retrospective; 213; mean 76.7 (SD 8.05); 200/222; no existing ulcers | Any grade | 14 days | 124 | Logistic regression | Univariate analysis (p-value unclear) | Full model approach | None missing | No continuous predictor | 47 (EPV <10) | Not reported | Score system based on original scores (e.g., Braden scale score) |
| Norton scale | (Perneger et | Retrospective; 2,373; 63(19); 1,210/1,163; no existing ulcers | Any grade | 9 days | Not reported | Cox regression | Multivariable analysis | Full model approach | Not reported | No continuous predictor | 245 (EPV >20) | Not reported | Score system based on original scores (e.g., Braden scale score) |
| Fragnment | scale (Perneger | Prospective; 1,190; 61.4(19.1); 650/540; not reported | Any grade | 3 weeks | 11 | Logistic and Cox regression | No pre-selection | Full model approach | Not reported | Categorising or dichotomising | 182 (EPV <10) | Cross-validation | Score system |
| Score system | prePURSE study tool (Schoonhoven et al., 2006) | Prospective; 1,229; 60.1 (16.7); 556/673; no existing ulcers | Grade II or above (predicting five-day risk) | 12 weeks | 46 | Logistic regression | Univariate analysis in combination with other evidence | Not reported | Complete case analysis (MCAR claimed) | Categorising or dichotomising | 121 (EPV <10) | Re-sampling | Score system |
| Specific acute care (e.g., ICU settings (n = 4) | Compton ICU | Retrospective; 698; median 66 (IQR 56 to 75.25); 392/306; no existing pressure ulcers | Grade II or above | Median 6 (IQR 3 to 14) days | Not reported | Logistic regression | Not reported | Stepwise | 8% of missing data, analysis methods not reported | Not reported | 121 (EPV <10) | Not reported |
| Rose ICU scale (Rose et al., 2006) | Prospective; 111; age, gender and baseline skin status not reported | Any grade | 8 days | 36 | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported

32
<table>
<thead>
<tr>
<th>Surgical ICU risk assessment scale (Slowikowski and Funk, 2010)</th>
<th>Prospective; 369; 58.3 (19.3); 208/161; not reported</th>
<th>Any grade</th>
<th>Not reported</th>
<th>Not reported</th>
<th>Logistic regression</th>
<th>Univariate analysis (p = 0.10)</th>
<th>Stepwise</th>
<th>Not reported</th>
<th>Categorising or dichotomising</th>
<th>88 (EPV 10 – 20)</th>
<th>Not reported</th>
<th>Formula equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suriadi and Sanada (S.S.) scale (Suriadi et al., 2008)</td>
<td>Prospective; 105; 48.63 (17.48); 72/33; no existing pressure ulcers</td>
<td>Any grade</td>
<td>Mean 5.9 (SD 3.49) days</td>
<td>Not reported</td>
<td>Logistic regression and discriminant analysis</td>
<td>Multivariable analysis</td>
<td>Full model approach</td>
<td>Not reported</td>
<td>Categorising or dichotomising</td>
<td>35 (EPV 10 – 20)</td>
<td>Not reported</td>
<td>Score system</td>
</tr>
<tr>
<td>Cardiovascular surgery settings (n = 2)</td>
<td>Retrospective; 849; 68.66 (13.01); 389/186; not reported</td>
<td>Any grade</td>
<td>Mean 7.08 (SD 7.44) days</td>
<td>31</td>
<td>Logistic regression</td>
<td>Univariate analysis (p-value unclear)</td>
<td>Data reduction approach</td>
<td>MICE</td>
<td>Restricted cubic spline</td>
<td>101 (EPV &lt; 10)</td>
<td>Re-sampling</td>
<td>Nomogram scale</td>
</tr>
<tr>
<td>Cardiovascular surgical PU score (Lu et al., 2017)</td>
<td>Prospective; 149; 49.8 (17.7); 79/70; not reported</td>
<td>Any grade</td>
<td>Not reported</td>
<td>13</td>
<td>Logistic regression</td>
<td>No pre-selection</td>
<td>Stepwise (backward)</td>
<td>Not reported</td>
<td>Categorising or dichotomising</td>
<td>37 (EPV &lt; 10)</td>
<td>Not reported</td>
<td>Nomogram scale</td>
</tr>
<tr>
<td>Trauma and burn centres (n = 1)</td>
<td>Prospective; 36; 31.8 (10.9); 26/10; not reported</td>
<td>Any grade</td>
<td>26.5 days</td>
<td>Not reported</td>
<td>Logistic regression</td>
<td>No pre-selection</td>
<td>Stepwise (forward)</td>
<td>Not reported</td>
<td>No continuous predictor</td>
<td>11 (EPV &lt; 10)</td>
<td>Not reported</td>
<td>Score system based on original scores (e.g., Braden scale scores)</td>
</tr>
<tr>
<td>Long-term care settings (n = 5)</td>
<td>Retrospective; 1,684; 76.4 (8.6); 633/1,051; no existing ulcers</td>
<td>Any grade</td>
<td>Mean 60.6 (SD 94.3) days</td>
<td>25</td>
<td>Cox regression</td>
<td>Univariate analysis (p = 0.01)</td>
<td>Stepwise (backward)</td>
<td>Censoring in Cox regression</td>
<td>No continuous predictor</td>
<td>107 (EPV &lt; 10)</td>
<td>Not reported</td>
<td>Score system based on original scores (e.g., Braden scale score)</td>
</tr>
</tbody>
</table>
| Study | Design | Score System | Ulcers Status | Ulcers Grade | Duration | Method | Predictor Selection | Censoring | Continuous Predictor | Age | Gender | Skin Status | Predictor Type | Model Type | EPV | Score System
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bergquist (Berlowitz et al., 2001)</td>
<td>Retrospective</td>
<td>Any grade</td>
<td>60.6 (SD 94.3) days</td>
<td>25</td>
<td>Cox regression</td>
<td>Univariate analysis (p = 0.01)</td>
<td>Stepwise (backward)</td>
<td>Not reported</td>
<td>No</td>
<td>Mean</td>
<td>91 days</td>
<td>Logistic regression and decision tree tool</td>
<td>Not reported</td>
<td>107 (EPV &lt; 10)</td>
<td>Score system based on original scores (e.g., Braden scale)</td>
<td></td>
</tr>
<tr>
<td>Berglowitz 11-item model (Berlowitz et al., 1996)</td>
<td>Retrospective</td>
<td>Grade II or above</td>
<td>Not reported</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>Not reported</td>
<td>No</td>
<td>Univariate analysis (p = 0.10)</td>
<td>Not reported</td>
<td>Not reported</td>
<td>6 months</td>
<td>Logistic regression in combination with other evidence</td>
<td>Not reported</td>
<td>1,350 (EPV &gt; 20)</td>
<td>Not reported</td>
<td></td>
</tr>
<tr>
<td>Berlowitz MDS risk-adjustment model (Berlowitz et al., 2001a)</td>
<td>Retrospective</td>
<td>Grade II or above</td>
<td>Not reported</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>Imputation using another data set</td>
<td>Smoothing curves using LOESS</td>
<td>Not reported</td>
<td>No</td>
<td>3 months</td>
<td>Logistic regression in combination with other evidence</td>
<td>Not reported</td>
<td>902 (EPV &gt; 20)</td>
<td>Not reported</td>
<td></td>
<td></td>
</tr>
<tr>
<td>interRAI PURS (Poss et al., 2010)</td>
<td>Retrospective</td>
<td>Any grade</td>
<td>Not reported</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>Not reported</td>
<td>No</td>
<td>No continuous predictor</td>
<td>Appro. 503 (EPV &gt; 20)</td>
<td>Not reported</td>
<td>91 days</td>
<td>Logistic regression</td>
<td>Score system</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rehabilitation units (n = 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission PU-FIM model (DeJong et al., 2014)</td>
<td>Prospective</td>
<td>Grade II or above</td>
<td>Mean 36.5 (SD 31.4) days</td>
<td>16</td>
<td>Logistic regression &amp; recursive partitioning analysis</td>
<td>Multivariable analysis</td>
<td>Full model approach</td>
<td>Not reported</td>
<td>Not reported</td>
<td>159; 46.9 (19.1); 124/35; allowing for existing ulcers</td>
<td>Logistic regression</td>
<td>Stepwise (backward)</td>
<td>No continuous predictor</td>
<td>21 (EPV 10 – 20)</td>
<td>Not reported</td>
<td></td>
</tr>
<tr>
<td>Schue modified Braden scale (Schue and Langemo, 1998)</td>
<td>Retrospective</td>
<td>Any grade</td>
<td>Not reported</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>Stepwise (backward)</td>
<td>No continuous predictor</td>
<td>9 (EPV &lt; 10)</td>
<td>No</td>
<td>170/69.2 (10.9); 170/0; no existing pressure ulcers</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>Score system based on original scores (e.g., Braden scale)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclear (n = 1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic pressure ulcer prediction</td>
<td>Retrospective</td>
<td>Any grade</td>
<td>Not reported</td>
<td>C4.5 machine learning</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>3,300; age, gender and baseline skin status not reported</td>
<td>“tree pruning” of the</td>
<td>Not reported</td>
<td>Score system</td>
<td>Score system</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Summary of eight model validation studies

<table>
<thead>
<tr>
<th>Models by settings (Refs)</th>
<th>Data sources</th>
<th>Participants (total number, average age in years, male/ female, &amp; baseline skin status)</th>
<th>Follow-up duration/ length of stay</th>
<th>Modelling methods</th>
<th>Missing data</th>
<th>Number of events (EPVs)</th>
<th>External validation</th>
<th>Recalibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>General (acute care) hospital settings (n = 7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNH-PUPP model (Page et al., 2011)</td>
<td>Prospective</td>
<td>165; mean 68 (SD 18.4); 87/ 78; not reported</td>
<td>Mean 14.97 (SD 22.29) days</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>7 (both numbers of events and non-events &lt; 100)</td>
<td>Temporal data</td>
<td>Not reported</td>
</tr>
<tr>
<td>prePURSE study tool (Schoonhoven et al., 2005)</td>
<td>Prospective</td>
<td>1,440; age, gender and baseline skin status not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>External data</td>
<td>Not reported</td>
</tr>
</tbody>
</table>
## Specific acute care (e.g., ICU) settings (n = 4)

<table>
<thead>
<tr>
<th>Study Description</th>
<th>Study Design</th>
<th>Follow-up</th>
<th>Model</th>
<th>Event Count</th>
<th>Temporal Data</th>
<th>Recalibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compton ICU model (Compton et al., 2008)</td>
<td>Retrospective</td>
<td>9 days</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>56 (number of events &lt; 100; but number of non-events &gt; 100)</td>
<td>Temporal data</td>
</tr>
<tr>
<td>Suriadi and Sanada (S.S.) scale (Suriadi et al., 2008)</td>
<td>Prospective</td>
<td>Mean 8.19 (SD 5.11) days</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>72 (number of events &lt; 100; but number of non-events &gt; 100)</td>
<td>External data</td>
</tr>
</tbody>
</table>

## Long-term care settings (n = 5)

<table>
<thead>
<tr>
<th>Study Description</th>
<th>Study Design</th>
<th>Follow-up</th>
<th>Model</th>
<th>Event Count</th>
<th>Temporal Data</th>
<th>Recalibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlowitz 11-item model (Berlowitz et al., 1996)</td>
<td>Retrospective</td>
<td>6 months</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>Approx. 556 (numbers of events and non-events &gt; 100)</td>
<td>Temporal data</td>
</tr>
<tr>
<td>Berlowitz MDS risk-adjustment model (Berlowitz et al., 2001b)</td>
<td>Retrospective</td>
<td>3 months</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>608 observations (numbers of events and non-events &gt; 100)</td>
<td>Temporal data</td>
</tr>
<tr>
<td>interRAI PURS (Poss et al., 2010)</td>
<td>Retrospective</td>
<td>Not reported</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>Approx. 1,267 (numbers of events and non-events &gt; 100)</td>
<td>External data</td>
</tr>
<tr>
<td>interRAI PURS (Poss et al., 2010)</td>
<td>Retrospective</td>
<td>Not reported</td>
<td>Logistic regression</td>
<td>Not reported</td>
<td>Approx. 1,903 (numbers of events and non-events &gt; 100)</td>
<td>External data</td>
</tr>
</tbody>
</table>

## Summary

<table>
<thead>
<tr>
<th>Study Description</th>
<th>Event Count</th>
<th>Temporal Data</th>
<th>Recalibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seven empirically-derived models using prospective design</td>
<td>4,469 events in total, with a median of event numbers 556 (range 7 to 1,903); 37.5% (3/8) validations using fewer than 100 events</td>
<td>50% (4/8) using temporal data and another 50% (4/8) using external data</td>
<td>87.5% (7/8) not reporting recalibration; 12.5% (1/8) not doing recalibration</td>
</tr>
</tbody>
</table>
Table 4: Results of discrimination and calibration estimation by settings

<table>
<thead>
<tr>
<th>Prognostic models (number of data sources)</th>
<th>c-statistic (95% CI) (^\dagger) in each data source</th>
<th>References of data sources (^\dagger)</th>
<th>Risk of bias</th>
<th>Concern about applicability</th>
<th>Original reported results/ meta-analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General (acute care) hospitals</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical laboratory data-based predictive equation (n = 1)</td>
<td>0.79 (95% CI 0.66 to 0.88)</td>
<td>Hatanaka et al. (2008) (Dev)</td>
<td>High</td>
<td>High</td>
<td>0.79 (95% CI 0.66 to 0.88)</td>
</tr>
<tr>
<td>The Northern Hospital Pressure Ulcer Prevention Plan (TNH-PUPP) (n = 2)</td>
<td>0.86 (95% CI 0.81 to 0.90)</td>
<td>Page et al. (2011) (Dev)</td>
<td>High</td>
<td>Unclear</td>
<td>Pooled 0.86 (95% CI 0.81 to 0.90)</td>
</tr>
<tr>
<td>Fragment scale (n = 1)</td>
<td>0.80 (95% CI 0.77 to 0.84) for model development; 0.79 (0.75 to 0.82) for internal validation</td>
<td>Perneger et al. (2002) (Dev)</td>
<td>High</td>
<td>Low</td>
<td>Pooled 0.79 (95% CI 0.77 to 0.82) Fixed-effect; one data source but having one data for model development and another for internal validation; (I^2 = 0.0%), p-value = 0.695</td>
</tr>
<tr>
<td>prePURSE study tool (n = 1)</td>
<td>0.71 (95% CI 0.66 to 0.75)</td>
<td>Schoonhoven et al. (2006) (Dev)</td>
<td>High</td>
<td>Low</td>
<td>0.71 (95% CI 0.66 to 0.75)</td>
</tr>
<tr>
<td><strong>Long-term settings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berlowitz 11-item model (n = 2)</td>
<td>0.75 (95% CI 0.74 to 0.76)</td>
<td>Berlowitz et al. (1996) (Dev)</td>
<td>High</td>
<td>Unclear</td>
<td>Pooled 0.75 (95% CI 0.74 to 0.76) Fixed-effect; two data sources; (I^2 = 0.0%), p-value = 0.419</td>
</tr>
<tr>
<td></td>
<td>0.76 (95% CI 0.74 to 0.78)</td>
<td>Berlowitz et al. (1996) (Val)</td>
<td>Unclear</td>
<td>Unclear</td>
<td></td>
</tr>
<tr>
<td>Setting</td>
<td>Model</td>
<td>Risk of bias</td>
<td>Concern about applicability</td>
<td>Ref.</td>
<td>No. of data sources</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------</td>
<td>--------------</td>
<td>------------------------------</td>
<td>------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Rehabilitation units</td>
<td>Admission PU-FIM model (n = 1)</td>
<td>High</td>
<td>Unclear</td>
<td>De Jong et al. (2014) (Dev)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Vascular surgery pressure ulcer risk score (n = 1)</td>
<td>High</td>
<td>Unclear</td>
<td>Corniello et al. (2014) (Dev)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cardiovascular surgical PU nomogram score (n = 1)</td>
<td>High</td>
<td>High</td>
<td>Lu et al. (2017) (Dev)</td>
<td>1</td>
</tr>
<tr>
<td>Acute care and ICU settings</td>
<td>Compton ICU model (n = 2)</td>
<td>High</td>
<td>Unclear</td>
<td>Compton et al. (2008) (Dev)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Suriadi and Sanada (S.S.) scale (n = 1)</td>
<td>High</td>
<td>Unclear</td>
<td>Suriadi et al. (2008) (Dev)</td>
<td>1</td>
</tr>
<tr>
<td>Calibration</td>
<td>Prognostic models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total O: E ratio (95% CI *) in each data source</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ref. No. of data sources</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Risk of bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Concern about applicability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General (acute care) hospitals</td>
<td>1.00 (95% CI 0.84 to 1.19)</td>
<td>Schoonhoven et al. (2006) (Dev)</td>
<td>High</td>
<td>Low</td>
<td>1.00 (95% CI 0.84 to 1.19)</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------</td>
<td>----------------------------------</td>
<td>------</td>
<td>-----</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Long-term settings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berlowitz 11-item model (n = 2)</td>
<td>1.00 (95% CI 0.95 to 1.05)</td>
<td>Berlowitz et al. (1996) (Dev)</td>
<td>High</td>
<td>Unclear</td>
<td>Pooled 0.99 (95% CI 0.95 to 1.04)</td>
</tr>
<tr>
<td></td>
<td>0.97 (95% CI 0.90 to 1.06)</td>
<td>Berlowitz et al. (1996)(Val)</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Fixed-effect; two data sources; I² = 0.0%, p-value = 0.615</td>
</tr>
<tr>
<td>Berlowitz MDS risk-adjustment model (n = 2)</td>
<td>0.97 (95% CI 0.91 to 1.04)</td>
<td>Berlowitz et al. (2001a) (Dev)</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Pooled 0.94 (95% CI 0.88 to 1.01)</td>
</tr>
<tr>
<td></td>
<td>0.91 (95% CI 0.84 to 0.98)</td>
<td>Berlowitz et al. (2001b) (Val)</td>
<td>High</td>
<td>Unclear</td>
<td>Fixed-effect; two data sources; I² = 48.8%, p-value = 0.162</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cardiovascular surgery settings</th>
<th>1.00 (95% CI 0.76 to 1.32)</th>
<th>Lu et al. (2017) (Dev)</th>
<th>High</th>
<th>High</th>
<th>1.00 (95% CI 0.76 to 1.32)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Cardiovascular surgical PU nomogram score (n = 1)</th>
<th>1.00 (95% CI 0.76 to 1.32)</th>
<th>Lu et al. (2017) (Dev)</th>
<th>High</th>
<th>High</th>
<th>1.00 (95% CI 0.76 to 1.32)</th>
</tr>
</thead>
</table>