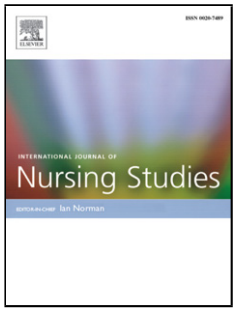
Evaluating the development and validation of empirically-derived prognostic models for pressure ulcer risk assessment: a systematic review

DOI:
10.1016/j.ijnurstu.2018.08.005

Document Version
Accepted author manuscript



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

PII: S0020-7489(18)30188-3
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

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 determine the 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 **CH**ecklist for critical **A**ppraisal and data extraction for systematic **R**eviews of prediction **M**odelling **S**tudies (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; Borlawsky 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 in 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, Fragmment 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 **PROG**nosis **RES**earch **S**trategy (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 proporties (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 Fragmment 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

Alba, A.C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P.J., McGinn, T., Guyatt, G., 2017. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA 318, 1377–1384.

Altman, D.G. (2001). Systematic reviews of evaluations of prognostic variables. The BMJ, 323, 224–228.

Altman, D.G., 2009. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest. 27, 235–243.

Anthony, D., Clark, M., Dallender, J. (2000). An optimization of the Waterlow score using regression and artificial neural networks. Clinical Rehabilitation, 14(1), 102-109.

Baldwin, K.M., Ziegler, S.M. (1998). Pressure ulcer risk following critical traumatic injury. Advances in Wound Care, 11(4), 168-173.

Bergquist, S. (2001). Subscales, subscores, or summative score: evaluating the contribution of Braden Scale items for predicting pressure ulcer risk in older adults receiving home health care. Journal of Wound, Ostomy, & Continence Nursing, 28(6), 279-289.

Berlowitz, D.R., Ash, A.S., Brandeis, G.H., Brand, H.K., Halpern, J.L., Moskowitz, M.A. (1996). Rating long-term care facilities on pressure ulcer development: importance of case-mix adjustment. Annals of Internal Medicine, 124(6), 557-563.

Berlowitz, D. R., Brandeis, G. H., Morris, J. N., Ash, A. S., Anderson, J. J., Kader, B., Moskowitz, M. A. (2001a). Deriving a risk-adjustment model for pressure ulcer development using the Minimum Data Set. Journal of the American Geriatrics Society, 49(7), 866-871.

Berlowitz, D. R., Brandeis, G. H., Anderson, J. J., Ash, A. S., Kader, B., Morris, J. N., Moskowitz, M. A. (2001b). Evaluation of a risk-adjustment model for pressure ulcer development using the Minimum Data Set. Journal of the American Geriatrics Society, 49(7), 872-876.

Borlawsky, T., Hripcsak, G. (2007). Evaluation of an automated pressure ulcer risk assessment model. Home Health Care Management & Practice, 19(4), 272-284.

Bouten, C.V., Oomens, C.W., Baaijens, F.P., Bader, D.L. (2003). The etiology of pressure ulcers: skin deep or muscle bound? Archives of Physical Medicine and Rehabilitation, 84, 616–619.

Coleman, S., Gorecki, C., Nelson, E.A., Closs, S.J., Defloor, T., Halfens, R., Farrin, A., Brown, J., Schoonhoven, L., Nixon, J., 2013. Patient risk factors for pressure ulcer development: systematic review. Int J Nurs Stud 50, 974–1003.

Collins, G.S., Mallett, S., Omar, O., Yu, L.-M., 2011. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine 9, 103.

Compton, F., Hoffmann, F., Hortig, T., Strauss, M., Frey, J., Zidek, W., Schafer, J. H. (2008). Pressure ulcer predictors in ICU patients: nursing skin assessment versus objective parameters.[Erratum appears in J Wound Care. 2008 Nov;17(11):493]. Journal of Wound Care, 17(10), 417-420, 422-414.

Cook, N.R., 2008. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin. Chem. 54, 17–23. https://doi.org/10.1373/clinchem.2007.096529

Corniello, A. L., Moyse, T., Bates, J., Karafa, M., Hollis, C., Albert, N. M. (2014). Predictors of pressure ulcer development in patients with vascular disease. Journal of Vascular Nursing, 32(2), 55-62.

Cullum, N., Buckley, H., Dumville, J., Hall, J., Lamb, K., Madden, M., Morley, R., O'Meara, S., Goncalves, P.S., Soares, M., Stubbs, N., 2016. Wounds research for patient benefit: a 5-year programme of research, Programme Grants for Applied Research. NIHR Journals Library, Southampton (UK).

Damen, J.A.A.G., Hooft, L., Schuit, E., Debray, T.P.A., Collins, G.S., Tzoulaki, I., Lassale, C.M., Siontis, G.C.M., Chiocchia, V., Roberts, C., Schlüssel, M.M., Gerry, S., Black, J.A., Heus, P., van der Schouw, Y.T., Peelen, L.M., Moons, K.G.M., 2016. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 353, i2416.

Debray, T.P.A., Damen, J.A.A.G., Snell, K.I.E., Ensor, J., Hooft, L., Reitsma, J.B., Riley, R.D., Moons, K.G.M., 2017. A guide to systematic review and meta-analysis of prediction model performance. BMJ 356, i6460.

Defloor, T., Grypdonck, M.F.H., 2004. Validation of pressure ulcer risk assessment scales: a critique. J Adv Nurs 48, 613–621.

DeJong, G., Ching-Hui, J. H., Brown, P., Smout, R. J., Horn, S. D., Ballard, P., Bouchard, T. (2014). Factors Associated with Pressure Ulcer Risk in Spinal Cord Injury Rehabilitation. American Journal of Physical Medicine & Rehabilitation, 93(11), 971-986.

Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M., 2006. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 59, 1087–1091.

García-Fernández, F.P., Pancorbo-Hidalgo, P.L., Agreda, J.J.S., 2014. Predictive capacity of risk assessment scales and clinical judgment for pressure ulcers: a meta-analysis. J Wound Ostomy Continence Nurs 41, 24–34.

Geersing, G.J., Bouwmeester, W., Zuithoff, P., Spijker, R., Leeflang, M., Moons, K., 2012. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One 7(2):e32844.

Groenwold, R.H.H., Moons, K.G.M., Pajouheshnia, R., Altman, D.G., Collins, G.S., Debray, T.P., Reitsma, J.B., Riley, R.D., Peelen, L.M., 2016. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. J. Clin. Epidemiol. 78, 90–100.

Harrell, F.E., 2015. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis, Second ed, Springer Series in Statistics. Springer International Publishing, Switzerland.

Hatanaka, N., Yamamoto, Y., Ichihara, K., Mastuo, S., Nakamura, Y., Watanabe, M., Iwatani, Y. (2008). A new predictive indicator for development of pressure ulcers in bedridden patients based on common laboratory tests results. Journal of Clinical Pathology, 61(4), 514-518.

He, W., Liu, P., Chen, H.-L., 2012. The Braden Scale cannot be used alone for assessing pressure ulcer risk in surgical patients: a meta-analysis. Ostomy Wound Manage 58, 34–40.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., Altman, D.G., 2003. Measuring inconsistency in meta-analyses. BMJ 327, 557–560.

Higgins, J.P.T., Green, S., 2011. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series. The Cochrane Collaboration and John Wiley & Sons Ltd., Chichester, West Sussex.

Ingui, B.J., Rogers, M.A., 2001. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc 8, 391–397.

Justice, A.C., Covinsky, K.E., Berlin, J.A., 1999. Assessing the generalizability of prognostic information. Ann. Intern. Med. 130, 515–524.

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.

Kottner, J., Dassen, T., Tannen, A., 2009. Inter- and intrarater reliability of the Waterlow pressure sore risk scale: a systematic review. Int J Nurs Stud 46, 369–379.

Kottner, J., Balzer, K., 2010. Do pressure ulcer risk assessment scales improve clinical practice? J Multidiscip Healthc 3, 103–111.

Kottner, J., Hauss, A., Schlüer, A.-B., Dassen, T., 2013. Validation and clinical impact of paediatric pressure ulcer risk assessment scales: A systematic review. Int J Nurs Stud 50, 807–818.

Lu, C.X., Chen, H.L., Shen, W.Q., Feng, L.P. (2017). A new nomogram score for predicting surgery-related pressure ulcers in cardiovascular surgical patients. International Wound Journal, 14(1), 226-232.

McInnes, E., Jammali-Blasi, A., Bell-Syer, S.E., Dumville, J.C., Middleton, V., Cullum, N., 2015. Support surfaces for pressure ulcer prevention. Cochrane Database Syst Rev 9, CD001735.

Moons, K.G.M., Kengne, A.P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D.G., Grobbee, D.E., 2012. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 98, 683–690.

Moons, K.G.M., de Groot, J.A.H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D.G., Reitsma, J.B., Collins, G.S., 2014. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 11, e1001744.

Moore, Z.E.H., Cowman, S., 2014. Risk assessment tools for the prevention of pressure ulcers. Cochrane Database Syst Rev 2, CD006471.

Mortenson, W.B., Miller, W.C., SCIRE Research Team, 2008. A review of scales for assessing the risk of developing a pressure ulcer in individuals with SCI. Spinal Cord 46, 168–175.

National Pressure Ulcer Advisory Panel, European Pressure Ulcer Advisory Panel and Pan Pacific Pressure Injury Alliance (NPUAP/EPUAP/PPPIA), 2014. Prevention and Treatment of Pressure Ulcers: Quick Reference Guide. Emily Haesler (Ed.). Cambridge Media: Osborne Park, Western Australia.

National Institute for Health and Care Excellence (NICE), 2014. Pressure ulcers: prevention and management (Clinical guideline No. CG179). http://guidance.nice.org.uk/CG179

Page, K. N., Barker, A. L., Kamar, J. (2011). Development and validation of a pressure ulcer risk assessment tool for acute hospital patients. Wound Repair & Regeneration, 19(1), 31-37.

Pancorbo-Hidalgo, P.L., Garcia-Fernandez, F.P., Lopez-Medina, I.M., Alvarez-Nieto, C., 2006. Risk assessment scales for pressure ulcer prevention: a systematic review. J Adv Nurs 54, 94–110.

Papanikolaou, P., Clark, M., Lyne, P. A. (2002). Improving the accuracy of pressure ulcer risk calculators: some preliminary evidence. International Journal of Nursing Studies, 39(2), 187-194.

Perneger, T. V., Gaspoz, J. M., Rae, A. C., Borst, F., Heliot, C. (1998). Contribution of individual items to the performance of the Norton pressure ulcer prediction scale. Journal of the American Geriatrics Society, 46(10), 1282-1286.

Perneger, T. V., Rae, A. C., Gaspoz, J. M., Borst, F., Vitek, O., Heliot, C. (2002). Screening for pressure ulcer risk in an acute care hospital: development of a brief bedside scale. Journal of Clinical Epidemiology, 55(5), 498-504.

Poss, J., Murphy, K. M., Woodbury, M. G., Orsted, H., Stevenson, K., Williams, G., MacAlpine, S., Curtin-Telegdi, N., Hirdes, J. P. (2010). Development of the interRAI Pressure Ulcer Risk Scale (PURS) for use in long-term care and home care settings. BMC Geriatrics, 10, 67.

Riley, R.D., Hayden, J.A., Steyerberg, E.W., Moons, K.G.M., Abrams, K., Kyzas, P.A., Malats, N., Briggs, A., Schroter, S., Altman, D.G., Hemingway, H., PROGRESS Group., 2013. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Medicine, 10, e1001380.

Rose P, Cohen R, Amsel, R. (2006). Development of a Scale to Measure the Risk of Skin Breakdown in Critically Ill Patients. American Journal of Critical Care, 15(3), 337.

Royston, P., Altman, D.G., 1994. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Journal of the Royal Statistical Society. Series C (Applied Statistics) 43, 429–467.

Royston, P., Altman, D.G., Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25, 127–141.

Saleh, M., Anthony, D., Parboteeah, S., 2009. The impact of pressure ulcer risk assessment on patient outcomes among hospitalised patients. Journal of Clinical Nursing 18(3), 1923-1929.

Samuriwo, R., Dowding, D., 2014. Nurses' pressure ulcer related judgements and decisions in clinical practice: A systematic review. International Journal of Nursing Studies 51, 1667–1685.

Schoonhoven, L., Haalboom, J. R., Bousema, M. T., Algra, A., Grobbee, D. E., Grypdonck, M. H., Buskens, E.; prePURSE study group. The prevention and pressure ulcer risk score evaluation study. (2002). Prospective cohort study of routine use of risk assessment scales for prediction of pressure ulcers. BMJ, 325(7368), 797.

Schoonhoven, L., Van Kol, E., Buskens, E., Van Achterberg, T., 2005. Pressure ulcers: development and validation of a prediction rule. 16th International Nursing Research Congress, Hawaii Big Island.

Schoonhoven, L., Grobbee, D.E., Donders, A.R.T., Algra, A., Grypdonck, M.H., Bousema, M.T., Schrijvers, A.J.P., Buskens, E., 2006. Prediction of pressure ulcer development in hospitalized patients: a tool for risk assessment. Quality & Safety in Health Care 65-70.

Schue, R. M., Langemo, D. K. (1998). Pressure ulcer prevalence and incidence and a modification of the Braden Scale for a rehabilitation unit. Journal of Wound, Ostomy, & Continence Nursing, 25(1), 36-43.

Shi, C., Dumville, J.C., Cullum, N., 2018. Support surfaces for pressure ulcer prevention: A network meta-analysis. PLoS One 13, e0192707.

Siregar, S., Groenwold, R.H.H., Heer, F. de, Bots, M.L., Graaf, Y. van der, Herwerden, L.A. van, 2012. Performance of the original EuroSCORE. Eur J Cardiothorac Surg 41, 746–54.

Siontis, G.C.M., Tzoulaki, I., Siontis, K.C., Ioannidis, J.P.A., 2012. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ 344, e3318.

Slowikowski, G.C., Funk, M. (2010). Factors associated with pressure ulcers in patients in a surgical intensive care unit. Journal of Wound, Ostomy, & Continence Nursing, 37(6), 619-626.

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

Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J., Kattan, M.W., 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 21, 128–138.

Steyerberg, E.W., Moons, K.G.M., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., Riley, R.D., Hemingway, H., Altman, D.G., PROGRESS Group, 2013. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 10, e1001381.

Steyerberg, E.W., Uno, H., Ioannidis, J.P.A., van Calster, B., Ukaegbu, C., Dhingra, T., Syngal, S., Kastrinos, F., 2017. Poor performance of clinical prediction models: the harm of commonly applied methods. Journal of Clinical Epidemiology. https://doi.org/10.1016/j.jclinepi.2017.11.013

Suriadi, Sanada, H., Sugama, J., Thigpen, B., Subuh, M. (2008). Development of a new risk assessment scale for predicting pressure ulcers in an intensive care unit. Nursing in Critical Care, 13(1), 34-43.

van der Heijden, G.J.M.G., Donders, A.R.T., Stijnen, T., Moons, K.G.M., 2006. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. Journal of Clinical Epidemiology 59, 1102–1109.

van Houwelingen, H.C., Arends, L.R., Stijnen, T., 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statist. Med. 21, 589–624. https://doi.org/10.1002/sim.1040

Webster, J., Coleman, K., Mudge, A., Marquart, L., Gardner, G., Stankiewicz, M., Kirby, J., Vellacott, C., Horton-Breshears, M., McClymont, A., 2011. Pressure ulcers: effectiveness of risk-assessment tools. A randomised controlled trial (the ULCER trial). BMJ Quality and Safety 20(4), 297-306.

Walker-Dilks, C., Wilczynski, N.L., Haynes, R.B., 2008. Cumulative Index to Nursing and Allied Health Literature search strategies for identifying methodologically sound causation and prognosis studies . Applied Nursing Research 21(2):98-103.

Wilczynski, N.L., Haynes, R.B., 2004. Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. BMC Medicine 2:23.

Williams, D.J., Zhu, Y., Grijalva, C.G., Self, W.H., Harrell, F.E., Reed, C., Stockmann, C., Arnold, S.R., Ampofo, K.K., Anderson, E.J., Bramley, A.M., Wunderink, R.G., McCullers, J.A., Pavia, A.T., Jain, S., Edwards, K.M., 2016. Predicting Severe Pneumonia Outcomes in Children. Pediatrics 138. https://doi.org/10.1542/peds.2016-1019

Wolff R, Collins G.S., Kleijnen, J., Mallett, S., Reitsma, J.B., Riley, R., Westwood, M., Whiting, P., Moon, K.G., 2016. PROBAST: a risk of bias tool for prediction modelling studies. Cochrane Colloquium, Seoul.

Zlotnik, A., Abraira, V., 2015. A general-purpose nomogram generator for predictive logistic regression models. Stata Journal 15, 537–546.

Fig 1. PRISMA flow chart of study selection

Figr-1

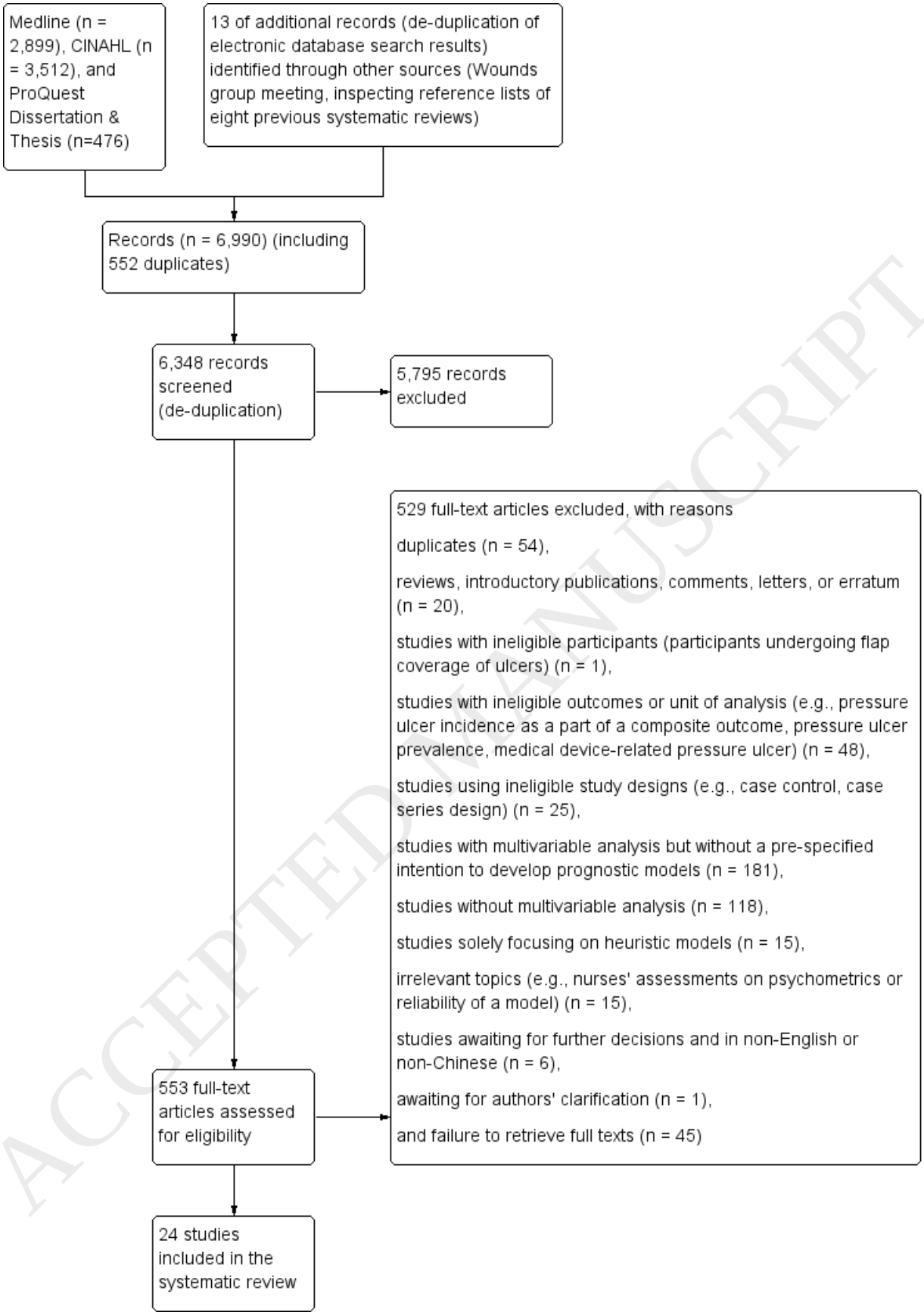


Table 1 Summary of empirically-derived models included in this review

Models by settings (numbers of models)	Development *		Validation		Comparisons	c-statistics ¶	O:E ratios ¶
	Data source	Risk of bias/ applicability	Data source	Risk of bias/ applicability	Numbers of comparisons – heuristic tools in comparisons (data source)	Numbers of data sources	Numbers of data sources
General (acute care) hospital settings (7)							
5-item Waterlow logistic model	Dev (Anthony et al., 2000)	High/unclear					
Clinical laboratory data-based equation	Dev (Hatanaka et al., 2008)	High/high			One – Braden (Hatanaka et al., 2008)	One (Dev)	
TNH-PUPP model	Dev (Page et al., 2011)	High/unclear	Val (Page et al., 2011)	High/unclear		Two (Dev–Val)	
Papanikolaou simplified Waterlow scale	Dev (Papanikolaou et al., 2002)	High/unclear			One – Waterlow (Papanikolaou et al., 2002)		
Perneger simplified Norton scale	Dev (Perneger et al., 1998)	High/unclear			One – Norton (Perneger et al., 1998)		
Fragmment scale	Dev (Perneger et al., 2002)	High/low			Two – Norton/ Braden (Perneger et al., 2002)	One (Dev)	
prePURSE study tool	Dev (Schoonhoven et al., 2006)	High/low	Val (Schoonhoven et al., 2005)	Unclear/unclear	Three – Norton/ Braden/ Waterlow (Schoonhoven et al., 2002)	One (Dev)	One (Dev)
Specific acute care (e.g., ICU) settings (4)							
Compton ICU model	Dev (Compton et al., 2008)	High/unclear	Val (Compton et al., 2008)	High/unclear	One – Waterlow (Compton et al., 2008)	Two (Dev–Val)	
Rose ICU scale	Dev (Rose et al., 2006)	High/unclear			One – Braden (Rose et al., 2006)		
Surgical ICU risk assessment scale	Dev (Slowikowski and Funk, 2010)	High/unclear			One – Braden (Slowikowski and Funk, 2010)		
Suriadi and Sanada (S.S.) scale	Dev (Suriadi et al., 2008)	High/unclear	Val (Suriadi et al., 2008)	High/unclear		One (Dev)	

Cardiovascular surgery settings (2)							
Vascular surgery PU risk score	Dev (Corniello et al., 2014)	High/unclear				One (Dev)	
Cardiovascular surgical PU score	Dev (Lu et al., 2017)	High/high				One (Dev)	One (Dev)
Trauma and burn centres (1)							
Baldwin two-item Braden scale	Dev (Baldwin and Ziegler, 1998)	High/high			One – Braden (Baldwin and Ziegler, 1998)		
Long-term care settings (5)							
Bergquist two-item Braden scale	Dev (Bergquist, 2001)	High/unclear			One – Braden (Bergquist, 2001)		
Bergquist three-item Braden scale	Dev (Bergquist, 2001)	High/unclear			One – Braden (Bergquist, 2001)		
Berlowitz 11-item model	Dev (Berlowitz et al., 1996)	High/unclear	Val (Berlowitz et al., 1996)	Unclear/unclear		Two (Dev–Val)	Two (Dev–Val)
Berlowitz MDS risk-adjustment model	Dev (Berlowitz et al., 2001a)	Unclear/unclear	Val (Berlowitz et al., 2001b)	High/unclear		Two (Dev–Val)	Two (Dev–Val)
interRAI PURS	Dev (Poss et al., 2010)	Unclear/high	Val 1 (Poss et al., 2010)	Unclear/high		Three (Dev–Val)	
			Val 2 (Poss et al., 2010)	Unclear/high			
Rehabilitation units (2)							
Admission PU-FIM model	Dev (DeJong et al., 2014)	High/unclear				One (Dev)	
Schue modified Braden scale	Dev (Schue and Langemo, 1998)	High/unclear			One – Braden (Schue and Langemo, 1998)		
Unclear setting (1)							
Electronic pressure ulcer prediction	Dev (Borlawsky and Hripcsak, 2007)	High/unclear					One (Dev)
Summary							

22 data-driven models	22 models (20 data sources)	7 models (8 data sources)	15 comparisons vs heuristic tools (12 models)	12 evaluations	5 evaluations

** 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

Models by settings (References)	Data sources and participants (total number, average age in years, gender in male/ female, and baseline skin status)	Outcomes (pressure ulcer grade and timing of risk prediction)	Follow-up duration/ length of stay	Candidate predictors	Modelling methods	Predictor selection prior to modelling	Predictor selection in modelling	Missing data	Continuous predictors	Number of events (EPVs)	Internal validation	Model presentation format
General (acute care) hospital settings (n = 7)												
5-item Waterlow logistic model (Anthony et al., 2000)	Prospective; 422; mean 64.83 (SD 17.85); 200/222; no existing pressure ulcers	Any grade	14 days	11	Logistic regression and artificial neural network	No pre-selection	Stepwise (forward)	Not reported	No continuous predictor	69 (EPV <10)	Data splitting	Formula equation
Clinical laboratory data-based equation (Hatanaka et al., 2008)	Prospective; 149; mean 71.6 (SD 11.3); 104/45; no existing pressure ulcers	Grade II or above (predicting three-month risk)	3 months	15	Logistic and Cox regression	Multivariable analysis	Full model approach	Not reported	Log transfer	38 (EPV <10)	Not reported	Formula equation

TNH-PUPP 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 et al., 2002)	Prospective; 213; mean 76.7 (SD 8.05); 200/222; no existing ulcers	Any grade	14 days	24	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)
Perneger simplified Norton scale (Perneger et al., 1998)	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)
Fragmment scale (Perneger et al., 2002)	Prospective; 1,190; 61.4(19.1); 650/540; not reported	Any grade (predicting five-day risk)	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
prePURSE study tool (Schoonhoven et al., 2006)	Prospective; 1,229; 60.1 (16.7); 556/673; no existing ulcers	Grade II or above (predicting one-week 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 model (Compton et al., 2008)	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	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

Surgical ICU risk assessment scale (Slowikowski and Funk, 2010)	Prospective; 369; 58.3 (19.3); 208/161; not reported	Any grade	Not reported	Not reported	Logistic regression	Univariate analysis (p = 0.10)	Stepwise	Not reported	Categorising or dichotomising	88 (EPV 10 – 20)	Not reported	Formula equation
Suriadi and Sanada (S.S.) scale (Suriadi et al., 2008)	Prospective; 105; 48.63 (17.48); 72/33; no existing pressure ulcers	Any grade	Mean 5.9 (SD 3.49) days	Not reported	Logistic regression and discriminant analysis	Multivariable analysis	Full model approach	Not reported	Categorising or dichotomising	35 (EPV 10 – 20)	Not reported	Score system
Cardiovascular surgery settings (n = 2)												
Vascular surgery PU risk score (Corniello et al., 2014)	Retrospective; 849; 68.66 (13.01); 389/186; not reported	Any grade	Mean 7.08 (SD 7.44) days	31	Logistic regression	Univariate analysis (p-value unclear)	Data reduction approach	MICE	Restricted cubic spline	101 (EPV < 10)	Re-sampling	Nomogram scale
Cardiovascular surgical PU score (Lu et al., 2017)	Prospective; 149; 49.8 (17.7); 79/70; not reported	Any grade	Not reported	13	Logistic regression	No pre-selection	Stepwise (backward)	Not reported	Not reported	37 (EPV < 10)	Not reported	Nomogram scale
Trauma and burn centres (n = 1)												
Baldwin two-item Braden scale (Baldwin and Ziegler, 1998)	Prospective; 36; 31.8 (10.9); 26/10; not reported	Any grade	26.5 days	Not reported	Logistic regression	No pre-selection	Stepwise (forward)	Not reported	No continuous predictor	11 (EPV < 10)	Not reported	Score system based on original scores (e.g., Braden scale scores)
Long-term care settings (n = 5)												
Bergquist two-item Braden scale (Bergquist, 2001)	Retrospective; 1,684; 76.4 (8.6); 633/1,051; no existing ulcers	Any grade	Mean 60.6 (SD 94.3) days	25	Cox regression	Univariate analysis (p = 0.01)	Stepwise (backward)	Censoring in Cox regression	No continuous predictor	107 (EPV < 10)	Not reported	Score system based on original scores (e.g., Braden scale score)

Bergquist three-item Braden scale (Bergquist, 2001)	Retrospective; 1,684; 76.4 (8.6); 633/1,051; no existing ulcers	Any grade	Mean 60.6 (SD 94.3) days	25	Cox regression	Univariate analysis (p = 0.01)	Stepwise (backward)	Censoring in Cox regression	No continuous predictor	107 (EPV < 10)	Not reported	Score system based on original scores (e.g., Braden scale score)
Berlowitz 11-item model (Berlowitz et al., 1996)	Retrospective; 31,150; 70.0 (11.6); 30,215/ 935; not reported	Grade II or above (predicting six-month risk)	6 months	Not reported	Logistic regression	Univariate analysis (p = 0.10)	Not reported	Not reported	Not reported	1,350 (EPV > 20)	Not reported	Not reported
Berlowitz MDS risk-adjustment model (Berlowitz et al., 2001a)	Retrospective; 14,607; 82.5 (10.9); 3,593/ 11,014; no existing ulcers	Grade II or above (predicting three-month risk)	3 months	Not reported	Logistic regression	Univariate analysis in combination with other evidence	Not reported	Imputation using another data set	Smoothing curves using LOESS	902 (EPV > 20)	Not reported	Not reported
interRAI PURS (Poss et al., 2010)	Retrospective; 14,083;82.2 (10.2); 4,338/ 9,745; no existing pressure ulcers	Any grade	91 days	Not reported	Logistic regression and decision tree tool	Not reported	Not reported	Not reported	No continuous predictor	Approx. 503 (EPV > 20)	Not reported	Score system
Rehabilitation units (n = 2)												
Admission PU-FIM model (DeJong et al., 2014)	Prospective; 159; 46.9 (19.1); 124/35; allowing for existing ulcers	Grade II or above	Mean 36.5 (SD 31.4) days	16	Logistic regression & recursive partitioning analysis	Multivariable analysis	Full model approach	Not reported	Not reported	21 (EPV 10 – 20)	Not reported	Not reported
Schue modified Braden scale (Schue and Langemo, 1998)	Retrospective; 170; 69.2 (10.9); 170/0; no existing pressure ulcers	Any grade	Not reported	6	Logistic regression	No pre-selection	Stepwise (backward)	Not reported	No continuous predictor	9 (EPV < 10)	Not reported	Score system based on original scores (e.g., Braden scale score)
Unclear (n = 1)												
Electronic pressure ulcer prediction	Retrospective; 3,300; age, gender and baseline skin status not reported	Any grade	Not reported	Not reported	C4.5 machine learning	Not reported	Not reported	Not reported	Not reported	Not reported	"tree pruning" of the	Not reported

(Borlawsky and Hripcsak, 2007)											decision tree tool	
Summary												
22 empirically-derived models	50% (11/22) using prospective design; 73,175 (median 560, range 36 to 31,150); median of average age 63.9 (range 31.8 to 82.5); 43,115/ 26,375; 54.5% (12/22) of models for people without pressure ulcers	22.7% (5/22) of models developed for risk at specific time points	The median follow-up duration 26.5 (range 6 to 180) days	The median number of predictors 16 (range 6 to 46)	77.3% (17/22) involving logistic regression	31.8% (7/22) using univariate analysis alone	40.9% (9/22) using stepwise selection methods	68.2% (15/22) not reporting information on missing data; 13.6% (3/22) using imputation	22.7% (5/22) categorising continuous data; 13.6% (3/22) transforming continuous data appropriately	4,007 events (median: 94.5; range 9 to 1,350); 54.5% (12/22) having EPVs < 10	22.7% (5/22) using methods for internal validation	27.3% (6/22) probably using original scale scores to derive score systems; 31.8% (7/22) using regression coefficients to derive nomograms or score systems

Table 3: Summary of eight model validation studies

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

Specific acute care (e.g., ICU) settings (n = 4)								
Compton ICU model (Compton et al., 2008)	Retrospective	329; median 67; 181/ 148; no existing pressure ulcers	9 days	Logistic regression	Not reported	56 (number of events < 100; but number of non-events > 100)	Temporal data	Not reported
Suriadi and Sanada (S.S.) scale (Suriadi et al., 2008)	Prospective	253; mean 51.3 (SD 19.4); 158/95; no existing pressure ulcers	Mean 8.19 (SD 5.11) days	Logistic regression	Not reported	72 (number of events < 100; but number of non-events > 100)	External data	Not reported
Long-term care settings (n = 5)								
Berlowitz 11-item model (Berlowitz et al., 1996)	Retrospective	17,946; mean 71.0 (SD 11.4); 17,408/538; not reported	6 months	Logistic regression	Not reported	Approx. 556 (numbers of events and non-events > 100)	Temporal data	Not reported
Berlowitz MDS risk-adjustment model (Berlowitz et al., 2001b)	Retrospective	13,457; age, gender and baseline skin status not reported	3 months	Logistic regression	Not reported	608 observations (numbers of events and non-events > 100)	Temporal data	No recalibration
interRAI PURS (Poss et al., 2010)	Retrospective	13,062; age, gender and baseline skin status not reported	Not reported	Logistic regression	Not reported	Approx. 1,267 (numbers of events and non-events > 100)	External data	Not reported
interRAI PURS (Poss et al., 2010)	Retrospective	73,183; age, gender and baseline skin status not reported	Not reported	Logistic regression	Not reported	Approx. 1,903 (numbers of events and non-events > 100)	External data	Not reported
Summary								
Seven empirically-derived models	37.5% (3/8) data sources using prospective design	119,835 (median 7,251, range 165 to 73,183); median of average age 67.5 (51.3 to 71); 17,834/ 859	The median follow-up duration 15 (8 to 180) days	87.5% (7/8) validations involving logistic regression	All not reported	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	50% (4/8) using temporal data and another 50% (4/8) using external data	87.5% (7/8) not reporting recalibration; 12.5% (1/8) not doing recalibration

Table 4: Results of discrimination and calibration estimation by settings

Discrimination					
Prognostic models (number of data sources)	**c-statistic (95% CI [*]) in each data source**	**References of data sources [†]**	**Risk of bias**	**Concern about applicability**	**Original reported results/ meta-analyses**
General (acute care) hospitals					
Clinical laboratory data-based predictive equation (n = 1)	0.79 (95% CI 0.66 to 0.88)	Hatanaka et al. (2008) (Dev)	High	High	0.79 (95% CI 0.66 to 0.88)
The Northern Hospital Pressure Ulcer Prevention Plan (TNH-PUPP) (n = 2)	0.86 (95% CI 0.81 to 0.90)	Page et al. (2011) (Dev)	High	Unclear	Pooled 0.86 (95% CI 0.81 to 0.90) Fixed-effect; two data sources; I^2 = 0.0%, p-value = 0.637
	0.90 (95% CI 0.66 to 0.98)	Page et al. (2011) (Val)	High	Unclear	
Fragmment scale (n = 1)	0.80 (95% CI 0.77 to 0.84) for model development; 0.79 (0.75 to 0.82) for internal validation	Perneger et al. (2002) (Dev)	High	Low	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
prePURSE study tool (n = 1)	0.71 (95% CI 0.66 to 0.75)	Schoonhoven et al. (2006) (Dev)	High	Low	0.71 (95% CI 0.66 to 0.75)
Long-term settings					
Berlowitz 11-item model (n = 2)	0.75 (95% CI 0.74 to 0.76)	Berlowitz et al. (1996) (Dev)	High	Unclear	Pooled 0.75 (95% CI 0.74 to 0.76) Fixed-effect; two data sources; I^2 = 0.0%, p-value = 0.419
	0.76 (95% CI 0.74 to 0.78)	Berlowitz et al. (1996) (Val)	Unclear	Unclear	

Berlowitz MDS risk-adjustment model (n = 2)	0.73 (95% CI 0.71 to 0.75)	Berlowitz et al. (2001a) (Dev)	Unclear	Unclear	Pooled 0.73 (95% CI 0.72 to 0.74) Fixed-effect; two data sources; I^2 = 0.0%, p-value = 1.000
	0.73 (95% CI 0.71 to 0.75)	Berlowitz et al. (2001b) (Val)	High	Unclear	
interRAI PURS (Logistic) (n = 3)	0.71 (95% CI 0.68 to 0.73)	Poss et al. (2010) (Dev)	Unclear	Unclear	Pooled 0.65 (95% CI 0.60 to 0.69) Random-effect; three data sources; I^2 = 95.8%, p-value = 0.000
	0.61 (95% CI 0.59 to 0.62)	Poss et al. (2010) (Val1)	Unclear	Unclear	
	0.63 (95% CI 0.62 to 0.64)	Poss et al. (2010) (Val2)	Unclear	Unclear	
Rehabilitation units					
Admission PU-FIM model (n = 1)	0.77 (95% CI 0.65 to 0.86)	DeJong et al. (2014) (Dev)	High	Unclear	0.77 (95% CI 0.65 to 0.86)
Cardiovascular surgery settings					
Vascular surgery pressure ulcer risk score (n = 1)	0.85 (95% CI 0.81 to 0.89)	Corniello et al. (2014) (Dev)	High	Unclear	0.85 (95% CI 0.81 to 0.89)
Cardiovascular surgical PU nomogram score (n = 1)	0.73 (95% CI 0.62 to 0.81)	Lu et al. (2017) (Dev)	High	High	0.73 (95% CI 0.62 to 0.81)
Acute care and ICU settings					
Compton ICU model (n = 2)	0.82 (95% CI 0.78 to 0.85)	Compton et al. (2008) (Dev)	High	Unclear	Pooled 0.81 (95% CI 0.78 to 0.84) Fixed-effect; two data sources; I^2 = 0.0%, p-value = 0.594
	0.80 (95% CI 0.73 to 0.86)	Compton et al. (2008) (Val)	High	Unclear	
Suriadi and Sanada (S.S.) scale (n = 1)	0.89 (95% CI 0.83 to 0.93)	Suriadi et al. (2008) (Dev)	High	Unclear	0.89 (95% CI 0.83 to 0.93)
Calibration					
Prognostic models	**Total O: E ratio (95% CI [*]) in each data source**	**Ref. No. of data sources [†]**	**Risk of bias**	**Concern about applicability**	**Original reported results/ meta-analyses**

General (acute care) hospitals					
prePURSE study tool (n = 1)	1.00 (95% CI 0.84 to 1.19)	Schoonhoven et al. (2006) (Dev)	High	Low	1.00 (95% CI 0.84 to 1.19)
Long-term settings					
Berlowitz 11-item model (n = 2)	1.00 (95% CI 0.95 to 1.05)	Berlowitz et al. (1996) (Dev)	High	Unclear	Pooled 0.99 (95% CI 0.95 to 1.04) Fixed-effect; two data sources; I^2 = 0.0%, p-value = 0.615
	0.97 (95% CI 0.90 to 1.06)	Berlowitz et al. (1996)(Val)	Unclear	Unclear	
Berlowitz MDS risk-adjustment model (n = 2)	0.97 (95% CI 0.91 to 1.04)	Berlowitz et al. (2001a) (Dev)	Unclear	Unclear	Pooled 0.94 (95% CI 0.88 to 1.01) Fixed-effect; two data sources; I^2 = 48.8%, p-value = 0.162
	0.91 (95% CI 0.84 to 0.98)	Berlowitz et al. (2001b) (Val)	High	Unclear	
Cardiovascular surgery settings					
Cardiovascular surgical PU nomogram score (n = 1)	1.00 (95% CI 0.76 to 1.32)	Lu et al. (2017) (Dev)	High	High	1.00 (95% CI 0.76 to 1.32)