medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Accuracy and clinical effectiveness of risk prediction tools for pressure injury occurrence: An umbrella review

Bethany Hillier[1,2]

Katie Scandrett[1,2]

April Coombe[1,2]

Tina Hernandez-Boussard[3]

Ewout Steyerberg[4]

Yemisi Takwoingi[1,2]

Vladica Velickovic[5,6]

Jacqueline Dinnes[1,2*]

Affiliations

[1] Biostatistics, Evidence Synthesis, Test Evaluation And Prediction Modelling (BESTEAM), Institute of Applied Health Research, University of Birmingham, Edgbaston, Birmingham, UK

[2] NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK

[3] Department of Medicine, Stanford University, Stanford, CA USA

[4] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[5] Evidence Generation Department, HARTMANN GROUP, Heidenheim, Germany

[6] Institute of Public Health, Medical, Decision Making and Health Technology Assessment, UMIT, Hall, Tirol, Austria

* Corresponding author:

E-mail: j.dinnes@bham.ac.uk (JD)

Keywords

Sensitivity, specificity, AUC, AUROC, prognostic model, clinical scale, pressure injury, pressure ulcer, incidence, umbrella review, overview

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

ABSTRACT

Background

Pressure injuries (PIs) pose a substantial healthcare burden and incur significant costs worldwide. Several risk prediction models to allow timely implementation of preventive measures and potentially reduce healthcare system burden are available and in use. The ability of risk prediction tools to correctly identify those at high risk of PI (prognostic accuracy) and to have a clinically significant impact on patient management and outcomes (effectiveness) is not clear.

We aimed to evaluate the prognostic accuracy and clinical effectiveness of risk prediction tools for PI, and to identify gaps in the literature.

Methods and Findings

The umbrella review was conducted according to Cochrane guidance. MEDLINE, Embase, CINAHL, EPISTEMONIKOS, Google Scholar and reference lists were searched to identify relevant systematic reviews. Risk of bias was assessed using adapted AMSTAR-2 criteria. Results were described narratively.

We identified 16 reviews that assessed prognostic accuracy and 10 that assessed clinical effectiveness of risk prediction tools for PI. The 16 reviews of prognostic accuracy evaluated 63 tools (39 scales and 24 machine learning models), with the Braden, Norton, Waterlow, Cubbin-Jackson scales (and modifications thereof) the most evaluated tools. Meta-analyses from a focused set of included reviews showed that the scales had sensitivities and specificities ranging from 53%-97% and 46%-84%, respectively. Only 2/16 reviews performed appropriate statistical synthesis and quality assessment. One review assessing machine learning based algorithms reported high prognostic accuracy estimates, but some of which were sourced from the same data within which the models were developed, leading to potentially overoptimistic results.

Two randomised trials assessing the effect of PI risk assessment tools on incidence of PIs were identified from the 10 systematic reviews of clinical effectiveness; both were included in a Cochrane review and assessed as high risk of bias. Both trials found no evidence of an effect on PI incidence.

Conclusions

Our findings underscore the lack of high-quality evidence for the accuracy of risk prediction tools for PI. There is no reliable evidence to suggest that using existing risk prediction tools effectively reduces the incidence of PIs. Further research is needed on their clinical effectiveness, but only once promising prediction tools have been developed and appropriately validated.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

INTRODUCTION

Pressure injuries (PI), also known as pressure ulcers or decubitus ulcers, have an estimated global prevalence of 12.8% among hospitalised adults,[1] and place a significant burden on healthcare systems (estimated at $26.8 billion per year in the US alone[2]). PIs are most common in individuals with reduced mobility, limited sensation, poor circulation, or compromised skin integrity, and can affect those in community settings and long-term care as well as hospital settings. Effective prevention of PI requires multicomponent preventive strategies such as mattresses, overlays, and other support systems, nutritional supplementation, repositioning, dressings, creams, lotions, and cleansers.[3] It is therefore important to correctly identify those most at risk of PI to allow timely and targeted implementation of preventive measures, to reduce harm and consequently burden to healthcare systems.[4]

Numerous clinical assessment scales (e.g. Braden[5 6], Norton[7] and Waterlow[8]) and statistical risk prediction models for assessing the risk of PI are available however, many are limited by reliance on subjective clinical judgment and do not appear to meet basic standards for the development or reporting of risk prediction models.[9] Nevertheless, many such tools are in routine clinical usage. For example, in certain hospitals and long-term care settings in the US, healthcare professionals must conduct mandatory risk assessments for PI for all patients for the purposes of risk stratification and clinical triage.

Despite the apparent lack of sound methods for development and validation (including external validation) of available risk prediction tools, there is a considerable body of evidence evaluating their clinical utility, much of which has been synthesised in systematic reviews and meta-analyses.[9] Clinical utility includes both prognostic accuracy and clinical effectiveness. Prognostic accuracy is estimated by applying a numeric threshold above (or below) which there is a greater risk of PI, with study results presented using accuracy metrics such as sensitivity, specificity or the area under the receiver operating characteristic (ROC) curve.[10] Resulting accuracy is driven not only by the nominated threshold for defining participants as at low or high risk for PI but by other study factors including population and setting.[11] Clinical effectiveness, or the ability of a tool to impact on health outcomes such as the incidence or severity of PI, is related both to the accuracy of the tool (or its ability to correctly identify those most likely to develop PI) and to the uptake and implementation of the tool in practice. Demonstrating a change in health outcomes as a result of use of a risk prediction tool is vital to encourage implementation.[12]

Using an umbrella review approach, we aimed to provide a comprehensive overview of available systematic reviews that consider the prognostic accuracy and clinical effectiveness of PI risk prediction tools.

METHODS

Protocol registration and reporting of findings

We followed Cochrane guidance for conducting umbrella reviews[13], and 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies' (PRISMA-DTA) reporting guidelines[14] (see Appendix 1). The protocol was registered on Open Science Framework (https://osf.io/tepyk).

Literature search

Electronic searches of MEDLINE, Embase via Ovid and CINAHL Plus EBSCO from inception to January 2023 were developed and conducted by an experienced information specialist (AC), employing well-

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

established systematic review and prognostic search filters,[15-17] combined with appropriate keywords related to PIs. Simplified supplementary searches in EPISTEMONIKOS and Google Scholar were also undertaken (see Appendix 2 for further details). Screening of search results and full texts were conducted independently and in duplicate by two reviewers (BH, JD, YT, KS), with disagreements resolved by a third reviewer.

Eligibility criteria for this umbrella review

Published English-language systematic reviews of risk prediction tools developed for adult patients at risk of PI in any setting were included. Clinical risk assessment scales and models developed using statistical or machine learning (ML) methods were eligible (models exclusively using pressure sensor data were not considered). Risk prediction tools could be applied by any healthcare professional using any threshold for classifying patients as high or low risk and using any PI classification system[18-21] as a reference standard. For prognostic accuracy, we required accuracy metrics, such as sensitivity and specificity, to be presented but did not require full 2x2 classification tables to be reported. Reviews on diagnosing or staging suspected or existing PIs were excluded.

To be considered 'systematic', reviews were required to report a thorough search of at least two electronic databases and at least one other indication of systematic methods (e.g. explicit eligibility criteria, formal quality assessment of included studies, adequate data presentation for reproducibility of results, or review stages (e.g. search screening) conducted independently in duplicate).

Data extraction and quality assessment

Data extraction forms (Appendix 3) were informed by the CHARMS checklist (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) and Cochrane Prognosis group template.[22 23] Data extraction items included review characteristics, number of studies and participants, study quality and results.

The methodological quality of included systematic reviews was assessed using AMSTAR-2 (A Measurement Tool to Assess Systematic Reviews)[24], adapted for systematic reviews of risk prediction models (Appendix 4). Our adapted AMSTAR-2 contains six critical items, and limitations in any of these items reduces the overall validity of a review.[24] Quality assessment and data extraction were conducted by one reviewer and checked by a second (BH, JD, KS), with disagreements resolved by consensus.

Synthesis methods

Reviews about prognostic accuracy and clinical effectiveness of risk prediction tools were considered separately. Review methods and results were tabulated and a narrative synthesis provided. Prognostic accuracy results from reviews including a statistical synthesis were tabulated according to risk prediction tool.

Considerable overlap in risk prediction tools and included primary studies was noted between reviews. For risk prediction tools that were included in multiple meta-analyses, we focused our synthesis on the review(s) with the most recent search date or most comprehensive (based on number of included studies) and most robust estimate of prognostic accuracy (judged according to the appropriateness of the meta-analytic method used, e.g. use of recommended hierarchical approaches for test accuracy data[25]). The prognostic accuracy of risk prediction tools that were included in three or fewer reviews, was reported only if an appropriate method of statistical synthesis[13] was used.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

For clinical effectiveness results, reviews with the most recent search date or most comprehensive overview of available studies and that at least partially met more of the AMSTAR-2 criteria[24] were prioritised for narrative synthesis.

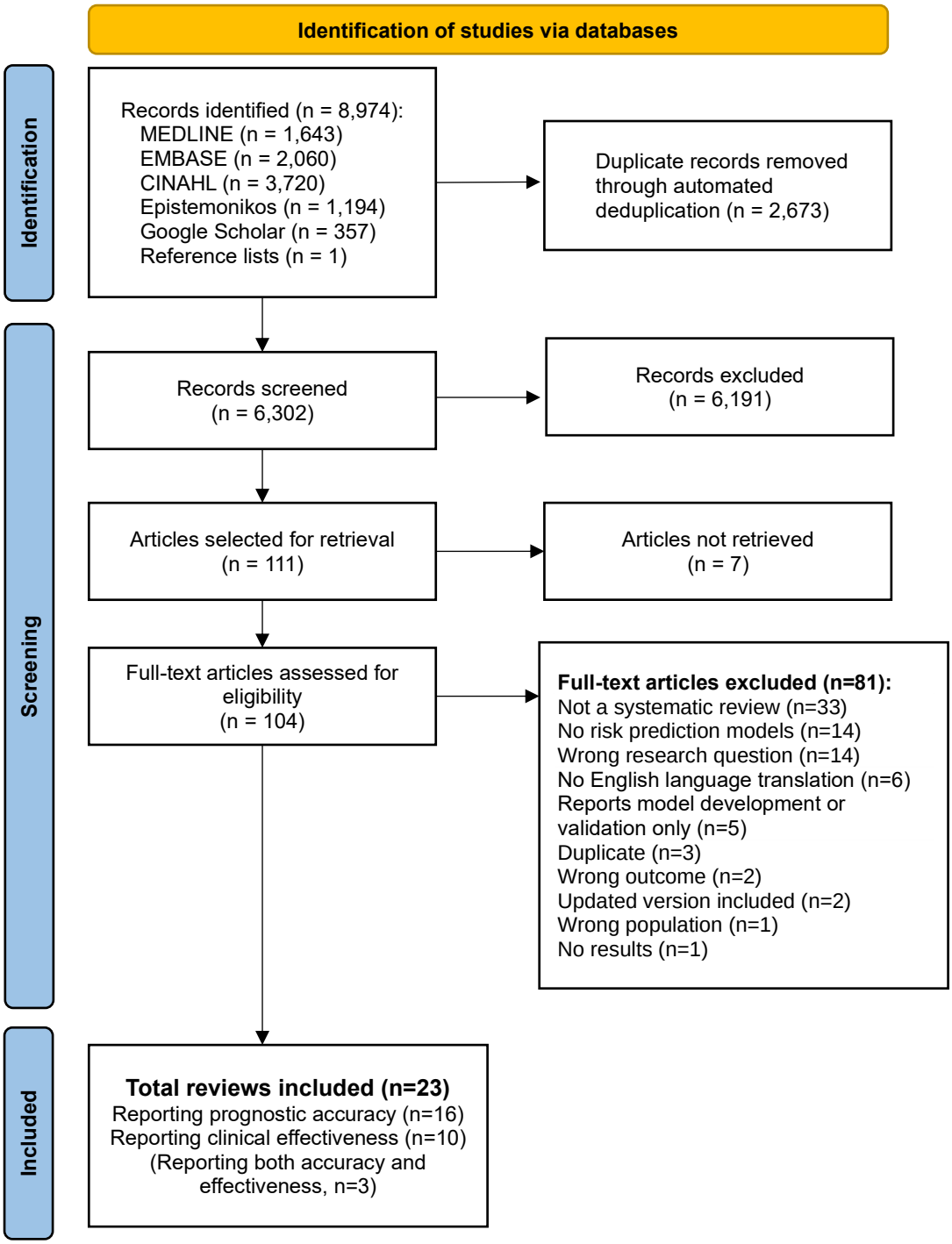
RESULTS

Characteristics of included reviews

A total of 110 records were selected for full-text assessment from 6302 unique records. We could obtain the full text of 104 publications, of which 23 reviews met all eligibility criteria (Figure 1), 16 reported accuracy data[24-39] and 10 reported clinical effectiveness data[25 29 34 40-46] (three reported both accuracy and effectiveness data[25 29 34]). Table 1 and Figure 2 provide an overview of the characteristics, methods and methodological quality of all 23 reviews (see Appendix 5 for full details).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figure 1. PRISMA[26] flowchart: identification, screening and selection process



List of full-text articles excluded, with reasons, is given in Appendix 5.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

167 *Table 1. Summary of included systematic review characteristics*

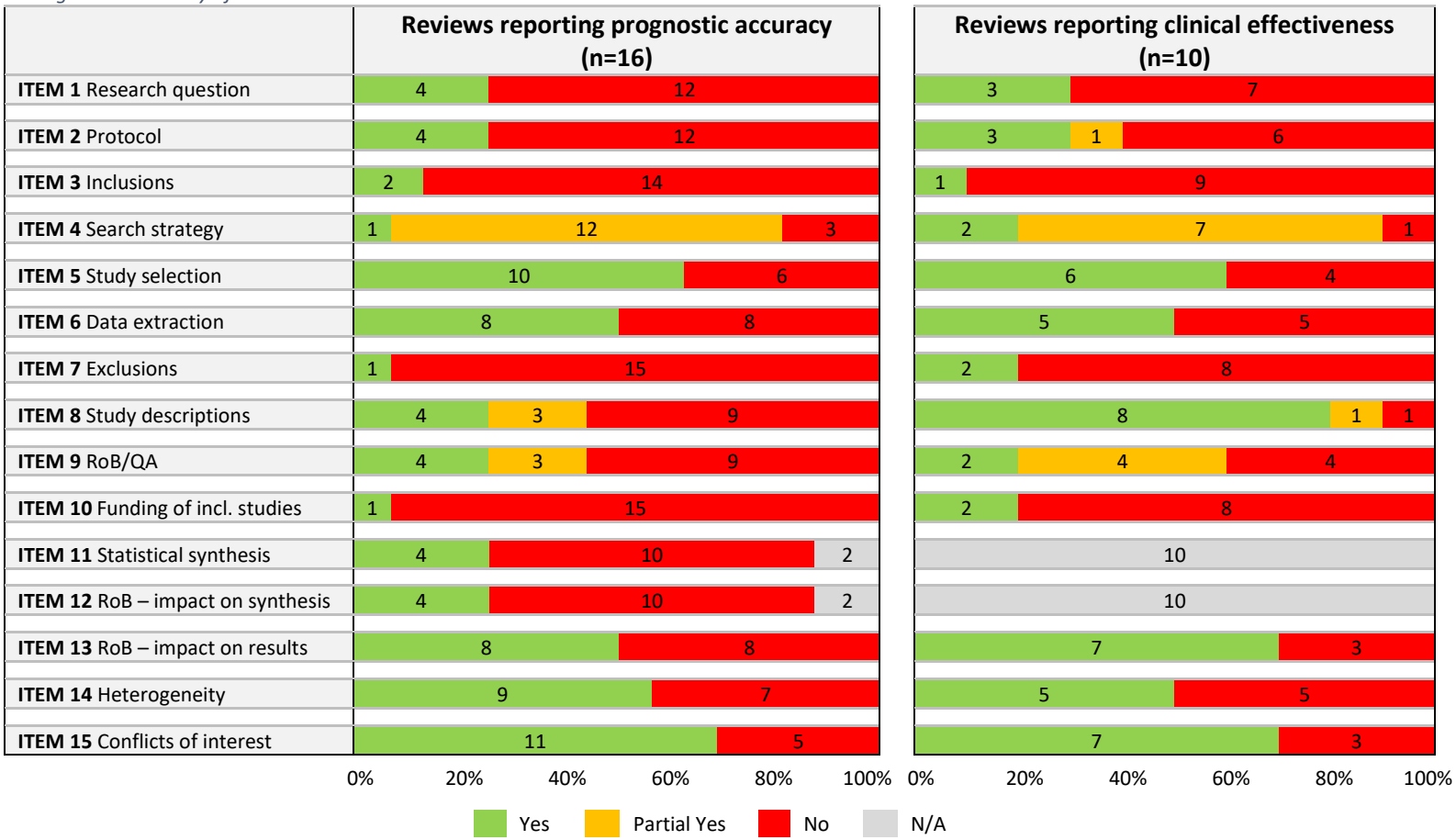
Review characteristic	Reviews on prognostic accuracy of risk prediction tools (N=16)	Reviews on clinical effectiveness of risk prediction tools (N=10)	All included reviews (N=23)
Median (range) year of publication	2016 (2006 – 2022)	2014 (2006 – 2021)	2016 (2006 – 2022)
Eligibility criteria			
Participants			
Adults only	9 (56) [A]	5 (50)	13 (57) [A]
Any age	2 (13)	1 (10)	3 (13)
No age restriction reported	5 (31)	4 (40)	7 (30)
Presence of PI at baseline			
Excluded (no PI at baseline)	5 (31)	2 (20)	6 (26)
NS	11 (69)	7 (70)	17 (74)
Setting			
Any healthcare setting	6 (38)	5 (50)	10 (43)
Hospital	2 (13)	2 (20)	4 (22)
Long-term care	2 (13)	0 (0)	2 (9)
Acute care (incl. surgical and ICU)	5 (31)	2 (20)	6 (26)
Long-term, acute or community settings	1 (6)	1 (10)	1 (4)
Risk assessment tools			
Any prediction tool or scale	5 (31)	6 (60)	9 (39)
Specified clinical scale(s)	8 (50)	2 (20)	10 (43)
ML-based prediction models	1 (6)	0 (0)	1 (4)
PI prevention strategies	0 (0)	1 (10)	1 (4)
NS	2 (13)	1 (10)	2 (9)
PI classification system			
Any	1 (6)	0 (0)	1 (4)
Accepted standard classifications	1 (6)	1 (10)	2 (9)
Several specified classification systems (NPUAP, EPUAP, AHCPR or TDCPS)	3 (19)	1 (10)	3 (13)
PI stage predefined[27]/defined by study authors	1 (6)	0 (0)	1 (4)
NS	10 (63)	8 (80)	16 (70)
Source of data			
Prospective only	4 (25)	2 (20)	4.5 (20) [B]
Prospective or retrospective	1 (6)	2 (20)	2.5 (7) [B]
NS	11 (69)	6 (60)	16 (60)
Study design restrictions			
Yes	7 (44)	6 (60)	12 (52)
No	2 (13)	2 (20)	3 (13)
NS	7 (44)	2 (20)	8 (39)
Phase of development/evaluation of tools			
External evaluations only	1 (6)	N/A	N/A
Validation studies (internal or external NS)	1 (6)	N/A	N/A
NS	14 (88)	N/A	N/A
Review methods			
Median (range) no. sources[C] searched	6 (2 – 14)	5 (3 – 14)	6 (2 – 14)
Publication restrictions:			
End date (year)			
2000-2009	1 (6)	3 (30)	3 (13)
2010-2019	12 (75)	6 (60)	16 (70)
2020-2023	3 (19)	1 (10)	4 (17)
Language			

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

English only	6 (38)	6 (60)	9 (39)
2 languages	0 (0)	2 (20)	2 (9)
>2 languages	2 (13)	2 (20)	3 (13)
No restrictions	3 (19)	1 (10)	4 (17)
NS	5 (31)	0 (0)	5 (22)
Quality assessment tool [D]			
PROBAST	0 (0) [E]	N/A	0 (0) [E]
QUADAS	2 (13)	N/A	2 (9)
QUADAS-2	7 (44)	N/A	7 (30)
JBI tools	1 (6)	2 (20)	3 (13)
CASP	2 (13)	1 (10)	2 (9)
Cochrane RoB tool	0 (0)	1 (10)	1 (4)
Other	1 (6)	4 (40)	5 (22)
None	3 (19)	2 (20)	4 (17)
Meta-analysis included	13 (81)	0 (0)	13 (57)
Method of meta-analysis (% of reviews incl. meta-analysis)			
Univariate RE/FE model (depending on heterogeneity assessment)	2 (15) [F]	N/A	N/A
Univariate RE model	5 (38) [F]	N/A	N/A
Hierarchical model (for DTA studies)	2 (15)	N/A	N/A
Unclear/NS	4 (31) [F]	N/A	N/A
Volume of evidence			
Median (range) no. studies	19 (2 – 70)	4 (1 – 20)	13 (1 – 70)
Median (range) no. participants	11,729 (609 – 221,541)	1,910 (528 – 2,273)	6,106 (528 – 221,541)
Median (range) no. tools	4 (1 – 28)	3 (2 – 9)	3 (1 – 28)

Figures are number (%) of reviews, unless otherwise specified. [A] one review[28] restricted to aged >60 years; [B] one review[29] states either prospective or retrospective data eligible for Research Question 1, but prospective only for Research Question 2, hence 0.5 added to each category; [C] including databases, bibliographies or registries; [D] reviews may fall into multiple categories, therefore total number within domain not necessarily equal to N (100%); [E] one review[30] reported use of PROBAST in methods, but did not present any PROBAST results; [F] one review conducts univariate meta-analysis for single estimate, e.g. AUC[31], RR[32] or OR[33]; AHCPR – Agency for Health Care Policy and Research; AUC – area under the curve; CASP – Critical Appraisal Skills Programme; DTA – diagnostic test accuracy; EPUAP – European Pressure Ulcer Advisory Panel; FE – fixed effects; ICU – intensive care unit; JBI – Joanna Briggs Institute; ML – machine learning; NPUAP – National Pressure Ulcer Advisory Panel; NS – not stated; OR – odds ratio; PI – pressure injury; PROBAST – Prediction model Risk of Bias Assessment; QUADAS (2) – Quality Assessment of Diagnostic Accuracy Studies (Version 2); RE – random effects; RR – risk ratio; TDCPS – Torrance Developmental Classification of Pressure Sore.

Figure 2. Summary of AMSTAR-2 assessment results.



Item 1 – Adequate research question/ inclusion criteria?; Item 2 – Protocol and justifications for deviations?; Item 3 – Reasons for study design inclusions?; Item 4 – Comprehensive search strategy?; Item 5 – Study selection in duplicate?; Item 6 – Data extraction in duplicate?; Item 7 – Excluded studies list (with justifications)?; Item 8 – Included studies description adequate?; Item 9 – Assessment of RoB/quality satisfactory?; Item 10 – Studies' sources of funding reported?; Item 11 – Appropriate statistical synthesis method?; Item 12 – Assessment of impact of RoB on synthesised results?; Item 13 – Assessment of impact of RoB on review results?; Item 14 – Discussion/investigation of heterogeneity?; Item 15 – Conflicts of interest reported?; N/A – Not Applicable; RoB – Risk of Bias; QA – quality assessment. Further details on AMSTAR items are given in Appendix 4, and results per review are given in Appendix 5.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Reviews were published between 2006 and 2022. Approximately half (12/23, 52%) restricted inclusion to adult populations (Table 1), three included any age group, and nine (39%) did not report any age restrictions. Six reviews (6/23, 26%) specified only populations without PIs at baseline for inclusion. Acute care was the most common setting across both review questions, 5/16 (31%) and 4/10 (40%) for accuracy and effectiveness reviews, respectively. Quality assessment tools varied, with QUADAS-2 (n=7) or QUADAS (n=2) being most common for reviews of accuracy (9/16, 56%). One accuracy review[30] reported use of both QUADAS-2 and PROBAST tools in their methods, but only reported QUADAS-2 results.

Reviews of accuracy predominantly focused on studies using any (5/16, 31%) or pre-specified (8/16, 50%) risk assessment tools or scales, one included only ML-based prediction models.[30] A total of 63 risk prediction tools were reported across the reviews, including 24 ML models. The number of included risk prediction tools in a single review ranged from one[34-39] to 28[32]. Only two reviews reported eligibility criteria related to the development or validation of the risk prediction tools. One[33] (6%) excluded evaluation studies that used the same data that was used to develop the tool and the other[29] included only "validated risk assessment instruments", however this was not further defined and the review included studies reporting the original development of risk prediction tools.

The majority (13/16, 81%) of accuracy reviews conducted a statistical synthesis of data, however only two utilised currently recommended hierarchical approaches for the meta-analysis of test accuracy data,[36 40] seven conducted univariate meta-analysis of individual accuracy measures (e.g. sensitivity and specificity separately, or AUC[31], RR[32] or odds ratio[33]) and four did not clearly report the type of analysis approach used.

Of the 10 systematic reviews evaluating the clinical effectiveness of risk prediction tools, two only considered the reliability of risk assessment scales[41 42] and eight considered effects on patient outcomes (one of which also considered tool reliability[43]). More than half of reviews (6, 60%) compared use of PI risk assessment scales to clinical judgement alone or 'standard care'. The number of included studies ranged from one[44] to 20[45] and the sample sizes of primary studies ranged from one (one subject and 110 raters, in an inter-rater reliability study[46]) to 3,027 patients. Reported outcomes included the incidence of PIs (7/10), preventative interventions prescribed (5/10) and interrater reliability (3/10) (reported in Appendix 5). One (Cochrane) review used the Cochrane RoB tool for quality assessment of included studies and three used JBI (n=2) or CASP (n=1) tools. Due to heterogeneity in study design, risk prediction tools and outcomes evaluated, none of the included reviews provided any form of statistical synthesis of study results.

Methodological quality of included reviews

The quality of included reviews was generally poor (Figure 2; Appendix 5). The AMSTAR-2 items that were most consistently met (yes or partial yes) were: comprehensiveness of the search (19/23, 83%), study selection independently in duplicate (15/23, 65%), and conflicts of interest reported (17/23, 74%).

Of the 16 accuracy reviews, four (25%)[30 35 36 40] used an appropriate method of quality assessment of included studies (i.e. QUADAS or QUADAS-2 dependent on publication year) and presented judgements per study. Of the 10 effectiveness reviews, two (20%)[29 47] used an appropriate method of quality assessment (the Cochrane tool for assessing risk of bias[48] and a criteria consistent with AHRQ Methods Guide for Effectiveness and Comparative Effectiveness Reviews[49], respectively) and provided judgements per study. Four reviews either reported quality assessment results per study (n=3[41 45 50]) or were considered to use an appropriate quality assessment tool (n=1[33]) (AMSTAR-2 criterion partially met).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Of the accuracy reviews that included a statistical synthesis, 31% (4/13)[31 32 36 40] used an appropriate method of meta-analysis and investigated sources of heterogeneity. Two reviews[36 40] used recommended hierarchical approaches to meta-analysis of test accuracy data (the bivariate model[36] and hierarchical summary ROC (HSROC) model[40]) and one[31] calculated summary AUC using random effects meta-analysis.[32]

Compared to the reviews of accuracy, reviews of effectiveness more commonly provided adequate descriptions of primary studies (8/10, 80% vs 4/16, 25%) (Figure 2). No other major differences across review questions were noted.

Results from reviews evaluating the prognostic accuracy of risk prediction models

Five of 16 accuracy reviews were prioritised for narrative synthesis (Tables 2-3) and are reported below according to risk prediction tool. Four of the five reviews did not include development study estimates within their meta-analyses, but this information could not be ascertained for the review[30] of ML-based models. None of these reviews assessed the quality of development methods for the prediction tools considered in their statistical syntheses.

Braden, and modified Braden scales

The most recent and largest review[36] of the Braden scale (60 studies, including 49,326 patients), which used hierarchical bivariate meta-analysis, reported an overall summary sensitivity of 0.78 (95% CI 0.74, 0.82; 15,241 patients) and specificity of 0.72 (95% CI 0.66, 0.78; 34,085 patients) across all reported thresholds (range ≤10 to ≤20). Summary sensitivities and specificities ranged from 0.79 (95% CI 0.76, 0.82) and 0.66 (95% CI 0.55, 0.75) at the lowest cut-offs for identification of high-risk patients (≤15 in 15 studies) to 0.82 (95% CI 0.73, 0.89) and 0.70 (95% CI 0.62, 0.77) using a cut-off of 18 (15 studies), respectively. Heterogeneity investigations suggested higher accuracy for predicting pressure injury risk in patients with a mean age of 60 years or less, in hospitalised patients (compared to long-term care facility residents) and in Caucasian populations (compared to Asian populations).[36] The review noted a high risk of bias for the 'index test' section of the QUADAS-2 assessment in approximately a third of included studies, but failed to provide details or reasons for this assessment.

Two modified versions of the Braden scale[51 52] were included by Park and colleagues.[53] Summary sensitivities were 0.97 (95% CI 0.92, 0.99; 125 patients from four studies)[51] and 0.89 (95% CI 0.71, 0.98; 27 patients from two studies)[52], and summary specificities were 0.70 (95% CI 0.66, 0.73; 563 patients)[51] and 0.71 (95% CI 0.67, 0.75; 599 patients).[52] The review was rated critically low on the AMSTAR-2 assessment, with only 3/15 (20%) criteria fulfilled. Despite reporting the use of QUADAS-2 for their risk of bias assessment, QUADAS-2 results were not reported, except that none of the included studies were estimated to be at high risk.

Cubbin & Jackson scale

Zhang and colleagues[40] included six studies evaluating the original Cubbin & Jackson scale[54] (800 patients). Summary sensitivity and specificity were both reported as 0.84 (95% CIs 0.59, 0.95 and 0.66, 0.93, respectively) [40] suggesting that this represents the point on the HSROC curve where sensitivity equals specificity, particularly as reported thresholds ranged from 24 to 34. The review authors concluded that although the accuracy of the Cubbin & Jackson scale was higher than the EVARUCI scale and the Braden scale, low quality of evidence and significant heterogeneity limit the strength of conclusions that can be drawn.

273 *Table 2. Findings related to prognostic accuracy, by model: Characteristics and quality of studies included within reviews*

Review author (publication year)	**Tool(s) evaluated** n studies; N participants	**Brief description of included studies**	**Brief description of included study quality**	**Method of meta-analysis**
Huang[36] (2021)	**Braden**[5 6] n = 60; N = 49,326	Setting: hospital (n=45; includes 22 in ICU or other acute units), LTCF (n=15) Sample size: 25 to 10,098 Mean age: range 31.7±10.9 to 84.6±7.9 Design: 47 prospective, 13 retrospective Braden cut-off (out of 23): range ≤10 to ≤20 Development study not included.	QUADAS-2: Patient selection: low RoB in 11/60 (18%), unclear RoB (>50%); low concern about applicability in 44/60 (73%), high concern for applicability in 16; Index test: low RoB in 39/60 (65%), high RoB (approx. 33%); low concern about applicability in 51/60 (85%); Reference standard: low RoB in 58/60 (97%), unclear RoB in 2 (3%); low concern about applicability in 51/60 (85%), unclear concern in 9 (15%); Flow and timing: low RoB in 50/60 (83%; unclear RoB in 10 (17%).	Bivariate meta-analysis; SROC constructed; and subgroup/stratified analyses to explore heterogeneity
Zhang[40] (2021)	**4 tools evaluated in meta-analyses**	Studies not described according to prediction tool. All prospective	Studies not described according to prediction tool QUADAS-2: Overall judgement was "not so satisfactory".	HSROC model for >3 studies, or univariate fixed- or random-effects models if ≤3 studies; meta-regression heterogeneity investigation
	Braden[5 6] n = 18; N = 11,167	Cut-offs used range from 10.5 to 20. Development study not included.		
	Cubbin & Jackson[54] n = 6; N = 800	Cut-offs used range from 24 to 34. Development study not included.		
	EVARUCI[55] n = 3; N = 3,063	Cut-offs: >11, >11.5, NS Development study not included.		'Inconsistency' (I^2 statistic) of studies was found to be 0%, therefore univariate fixed-effects models used
	Waterlow[8] n = 4; N = 1,000	Cut-offs: 12, 16, <25, 20.5 Development study not included.		
Park[28] (2016b)	**3 tools evaluated**	Studies not described according to prediction tool Cut-offs selected "by following the one which the study researcher(s) indicated to be the most effective".		DTA meta-analysis (random effects) using MetaDiSc; no further details

	Braden[5 6] n = 25; N = 10,547	cut-off: 13 (n=2); 16 (n=8); 17 (n=2); 18 (n=9); 19 (n=3); 20 (n=1) Development study not included.		
	Norton[7] n = 5; N = 2,408	cut-offs: 14 (n=2); 16 (n=3) Development study not included.		
	Waterlow[8] n = 5; N = 1,406	15 (n=1); 16 (n=2); 17 (n=1); NS (n=1) Development study not included.		
Park[53] (2016a)	**5 tools evaluated**	Described below according to prediction tool	QUADAS-2: Studies not described according to prediction tool "None had 'high risk'"	DTA meta-analysis (random-effects) using MetaDiSc; Cochrane Handbook (2010)[56] and Walter 2002[57] cited
	Braden – modified by Song & Choi[51] n = 4; N = 688	Prospective (4/4), recruiting patients with no PI at baseline (hospital ward (n=2) or ICU (n=3); mean age in the 50s (n=2), 60s (n=2). Classification used: AHCPR (n=3), Bergstrom (n=1). Braden scale cut-off used: <21 (n=1), <23 (n=1), <24 (n=2) Development study not included.		
	Braden – modified by Pang & Wong[52] n = 2; N = 626	Prospective (2/2), recruiting patients with no PI at baseline (OS ward (n=1) or NS (n=1); mean age 79.4 and 54.1. Classification used: NPUAP (n=2) Braden scale cut-off used: <19 (n=1), <14 (n=1) Development study not included.		
	Cubbin & Jackson[54] n = 4; N = 662	Prospective (4/4); ICU patients for all studies (1 in surgical ICU), with no PI at baseline (n=3); mean age in the 50s (n=2), 60s (n=2). Classification used: AHCPR (n=2), NPUAP (n=1), Lowthian (n=1). C&J scale cut-off used: <24 (n=2), <26 (n=1), <28 (n=1) Development study not included.		
	Norton[7] n = 7; N = 2,899	Prospective (6/7); inpatients with no PI at baseline (1 LTC, 2 'hospital', 1 ICU, 1 ICU & wards); mean age in the 50s (n=1), 60s (n=3), or 80s (n=1), or NS (n=2). Classification used: AHCPR (n=3), NPUAP (n=2), EPUAP (n=1), TDCPS (n=1). Norton scale cut-off used: <14 (n=2, but reported as 3 in paper), <15 (n=2), <16 (n=3) Development study not included.		
	Waterlow[8] n = 6; N = 1,268	Prospective (6/6); all male* inpatients aged over 60 on average with no PI at baseline (3 included ICU patients).		

		Classification used: AHCPR (n=2), NPUAP (n=2), EPUAP (n=1), TDCPS (n=1). Waterlow scale cut-off used: <9 (n=1), <15 (n=1), <16 (n=2), <17 (n=1), NS (n=1) Development study not included.		
Qu[30] (2022)	Models by ML algorithm type	Characteristics only reported overall, not by algorithm type. Conducted in: hospital patients (n= 13); surgical patients (n=3), ICU (n=5), CVD patients (n=2), cancer patients (n=1), LTC (n=1) Unclear whether development, internal validation or external validation studies included.	QUADAS-2 details reported overall, not by algorithm type.	RevMan (Moses & Littenberg method[58 59]) for analysis of quantitative data (SROC plot presented) and Bayesian DTA-NMA
	Decision Tree n = 14; N = 118,292		2/14 high RoB; 10/14 low RoB; 2/14 unclear RoB	
	Logistic Regression n = 14; N = 195,927		4/14 high RoB; 9/14 low RoB; 1/14 unclear RoB	
	Neural Network n = 9; N = 97,815		1/9 high RoB; 7/9 low RoB; 1/9 unclear RoB	
	Random Forest n = 7; N = 161,334		1/7 high RoB; 6/7 low RoB.	
	Support Vector Machine n = 9; N = 152,068		1/9 high RoB; 8/9 low RoB.	

* as reported in review's text. However, the table reports a mixture of female and male participants for all studies, with a mean female proportion of 50.73%.

AHCPR – Agency for Health Care Policy and Research; CI – confidence interval; CVD – cardiovascular disease; DTA – diagnostic test accuracy; EPUAP – European Pressure Ulcer Advisory Panel; (H)SROC – (hierarchical) summary receiver operating characteristic curve; ICU – intensive care unit; LTC(F) – long-term care (facility); ML – machine learning; N – number of participants; n – number of studies; NMA – network meta-analysis; NS – not stated; NPUAP – National Pressure Ulcer Advisory Panel; PI – pressure injury; PPPU – Panel for the Prediction and Prevention of Pressure Ulcers; QUADAS – Quality Assessment of Diagnostic Accuracy Studies; RCT – randomised controlled trial; RoB – risk of bias; TDCPS – Torrance Developmental Classification of Pressure Sore.

280 *Table 3. Summary estimates of accuracy parameters (main results from statistical syntheses), by prediction tool*

Review author (publication year)	**n studies; N participants**	**Sensitivity (95% CI)** (N = no. participants with PI)	**Specificity (95% CI)** (N = no. participants without PI)	**Likelihood ratios (95% CI)**	**DOR (95% CI)**	**AUROC (95% CI)**
TOOL: Braden[5 6] (1987)						
Huang[36] (2021)	n = 60; N = 49,326	**0.78 (0.74, 0.82)** [A] N=15,241 By cut-off: *≤15* (n=15): 0.79 (0.76, 0.82) *16* (n=19): 0.75 (0.67, 0.82) *17* (n=4): 0.69 (0.61, 0.76) *18* (n=15): 0.82 (0.73, 0.89) *≥19* (n=7): 0.78 (0.65, 0.87)	**0.72 (0.66, 0.78)** [A] N=34,085 By cut-off: *≤15*: 0.66 (0.55, 0.75) *16*: 0.85 (0.70, 0.93) *17*:0.86 (0.50, 0.97) *18*: 0.70 (0.62, 0.77) *≥19*: 0.54 (0.44, 0.63)	**PLR 2.80 (2.30, 3.50)** [A] **NLR 0.30 (0.26, 0.35)** [A]	**9.00 (7.00, 13.00)** [A] By cut-off: *≤15*: 7.00 (4.00, 12.00) *16*: 17.00 (8.00, 36.00) *17*: 14.00 (2.00, 103.00) *18:* 11.00 (6.00, 20.00) *≥19*: 4.00 (2.00, 7.00)	**0.82 (0.79, 0.85)** [A] By cut-off: *≤15*: 0.80 (0.76, 0.83) *16*: 0.84 (0.80, 0.87) *17*: 0.73 (0.69, 0.77) *18*: 0.83 (0.79, 0.86) *≥19*: 0.67 (0.63, 0.71)
Zhang[40] (2021)	n = 18; N = 11,167	**0.78 (0.68, 0.85)** [B]	**0.61 (0.40, 0.79)** [B]	**PLR 2.00 (1.24, 3.24)** **NLR 0.36 (0.25, 0.52)**	**5.52 (2.61, 11.67)**	**0.78**
Park[28] (2016b)	n = 25; N = 10,547	**0.72 (0.69, 0.74)** [A]	**0.63 (0.62, 0.64)** [A]	**PLR 2.31 (1.98, 2.69)** [A] **NLR 0.43 (0.36, 0.51)** [A]	**6.50 (4.64, 9.11)** [A]	**0.79** [A] (SE = 0.02)
TOOL: Modified Braden scales: Braden – modified by Song & Choi[51] (1991)						
Park[53] (2016a)	n = 4; N = 688	**0.97 (0.92, 0.99)** [A] N=125	**0.70 (0.66, 0.73)** [A] N=563	**PLR 3.47 (1.33, 9.06)** [A] **NLR 0.08 (0.04, 0.19)** [A]	**56.56 (21.88, 146.21)** [A]	**0.95** [A] (SE 0.02)
TOOL: Braden – modified by Pang & Wong[52] (1998)						
Park[53] (2016a)	n = 2; N = 626	**0.89 (0.71, 0.98)** [A] N=27	**0.71 (0.67, 0.75)** [A] N=599	**PLR 2.87 (1.88, 4.38)** [A] **NLR 0.17 (0.06, 0.49)** [A]	**16.06 (4.75, 54.35)** [A]	Not calculated
TOOL: Cubbin & Jackson[54] (1991)						
Zhang[40] (2021)	n = 6; N = 800	**0.84 (0.59, 0.95)** [B]	**0.84 (0.66, 0.93)** [B]	**PLR 5.12 (2.70, 9.70)** **NLR 0.19 (0.08, 0.49)**	**26.45 (13.51, 51.78)**	**0.90**
Park[53] (2016a)	n = 4; N = 662	**0.67 (0.60, 0.74)** [A] N=194	**0.75 (0.71, 0.79)** [A] N=468	**PLR 2.80 (1.66, 4.72)** [A] **NLR 0.34 (0.15, 0.76)** [A]	**9.46 (2.41, 37.22)** [A]	**0.82** [A] (SE 0.06)
TOOL: EVARUCI[55] (2001)						
Zhang[40] (2021)	n = 3; N = 3,063	**0.84 (0.79, 0.89)** [A]	**0.68 (0.66, 0.70)** [A]	**PLR 2.32 (2.14, 2.51)** [A] **NLR 0.25 (0.19, 0.35)** [A]	**9.79 (6.81, 14.07)** [A]	**0.82** [A]
TOOL: Norton[7] (1962)						
Park[53] (2016a)	n = 7; N = 2,899	**0.75 (0.70, 0.79)** [A] N=383	**0.57 (0.55, 0.59)** [A] N=2,516	**PLR 1.77 (1.26, 2.50)** [A] **NLR 0.49 (0.32-0.76)** [A]	**7.57 (2.53, 22.64)** [A]	**0.82** [A] (SE 0.05)
Park[28] (2016b)	n = 5; N = 2,408	**0.76 (0.71, 0.80)** [A]	**0.55 (0.53, 0.57)** [A]	**PLR 1.58 (1.07, 2.34)** [A] **NLR 0.47 (0.29, 0.76)** [A]	**6.41 (1.72, 23.88)** [A]	**0.84** [A] (SE 0.07)

Review author (publication year)	**n studies; N participants**	**Sensitivity (95% CI)** (N = no. participants with PI)	**Specificity (95% CI)** (N = no. participants without PI)	**Likelihood ratios (95% CI)**	**DOR (95% CI)**	**AUROC (95% CI)**
TOOL: Waterlow[8] (1985)						
Zhang[40] (2021)	n = 4; N = 1,000	**0.63 (0.48, 0.76)** [B]	**0.46 (0.22, 0.71)** [B]	**PLR 1.16 (0.66, 2.01)** **NLR 0.82 (0.40, 1.67)**	**1.42 (0.40, 5.07)**	**0.56**
Park[53] (2016a)	n = 6; N = 1,268	**0.55 (0.49, 0.62)** [A] N=246	**0.82 (0.80, 0.85)** [A] N=1,222	**PLR 2.89 (1.74, 4.79)** [A] **NLR 0.46 (0.31, 0.70)** [A]	**9.22 (6.43, 13.23)** [A]	**0.82** [A] (SE 0.03)
Park[28] (2016b)	n = 5; N = 1,406	**0.53 (0.47, 0.60)** [A]	**0.84 (0.81, 0.86)** [A]	**PLR 3.09 (1.63, 5.83)** [A] **NLR 0.49 (0.34, 0.72)** [A]	**9.06 (6.30, 13.04)** [A]	**0.81** [A] (SE 0.03)
ML models: [C]						
Qu[30] (2022) **DT models**	n = 14; N = 118,292	**0.66 (0.42, 0.84)** N=7,557	**0.90 (0.78, 0.96)** N=110,735	**PLR 6.9 (3.2, 14.7)** **NLR 0.37 (0.20, 0.69)**	**18 (7, 49)**	**0.88 (0.85, 0.91)**
Qu[30] (2022) **LR models**	n = 14; N = 195,927	**0.71 (0.60, 0.80)** N=9046	**0.83 (0.75, 0.89)** N=186,881	**PLR 4.3 (3.1, 5.9)** **NLR 0.35 (0.26, 0.46)**	**12 (9, 17)**	**0.84 (0.81, 0.87)**
Qu[30] (2022) **NN models**	n = 9; N = 97,815	**0.73 (0.55, 0.86)** N=9488	**0.78 (0.65, 0.87)** N=88,327	**PLR 3.3 (2.1, 5.0)** **NLR 0.35 (0.21, 0.59)**	**9 (5, 19)**	**0.82 (0.79, 0.85)**
Qu[30] (2022) **RF models**	n = 7; N = 161,334	**0.72 (0.26, 0.95)** N=5486	**0.96 (0.80, 0.99)** N=155,848	**PLR 16.3 (2.4, 108.9)** **NLR 0.29 (0.07, 1.29)**	**56 (3, 1258)**	**0.95 (0.93, 0.97)**
Qu[30] (2022) **SVM models**	n = 9; N = 152,068	**0.81 (0.69, 0.90)** N=6562	**0.81 (0.59, 0.93)** N=145,506	**PLR 4.3 (1.8, 9.9)** **NLR 0.23 (0.13, 0.39)**	**19 (6, 54)**	**0.88 (0.85, 0.90)**

[A] summary statistic pooled across multiple thresholds; [B] estimate derived from HSROC, but method for choosing summary point unclear; [C] it is not reported, at review level, whether the results presented below are pooled across development, internal validation or external validation/evaluation studies.

AUROC – area under the receiver operating characteristic curve; CI – confidence interval; DT – decision tree; DOR – diagnostic odds ratio; HSROC – hierarchical summary receiver operating characteristic curve; LR – logistic regression; ML – machine learning; NLR – negative likelihood ratio; NN – neural network; NS – not stated; PLR – positive likelihood ratio; RF – random forest; SE – standard error; SVM – support vector machine.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Norton scale

Park and colleagues[53] pooled data from seven studies (2,899 participants) evaluating the Norton scale, across thresholds ranging from <14 to <16. They reported summary sensitivity of 0.75 (95% CI 0.70, 0.79) and specificity 0.57 (95% CI 0.55, 0.59). A further four reviews presented statistically synthesised results for the Norton scale (Appendix 5), including one review by Chou and colleagues[29] which included nine studies (5,444 participants) but only reported median values for accuracy parameters.

Waterlow scale

Although Zhang and colleagues[40] included the fewest participants (4 studies; 1,000 participants) of all six reviews that conducted a statistical synthesis of the accuracy of the Waterlow scale[8], they provided the most recent review. It was rated highest on AMSTAR-2 criteria and appropriately used the HSROC model for meta-analysis across thresholds ranging from 12 to 25. Summary sensitivity was 0.63 (95% CI 0.48, 0.76) and summary specificity 0.46 (95% CI 0.22, 0.71) (Table 3). A second review[53] reported summary sensitivity of 0.55 (95% CI 0.49, 0.62) and specificity 0.82 (95% CI 0.80, 0.85) (6 studies; 1268 participants).

Machine learning algorithms

Qu and colleagues[30] conducted separate meta-analyses of 25 studies according to ML algorithm type (Table 2). The review rated critically low on AMSTAR-2 items, with only 6/15 (40%) criteria fulfilled, and reported using Bayesian DTA meta-analysis. The review did not restrict inclusion to external evaluations of the models, and the authors did not report which estimates were sourced from development data or external data. The summary AUC for the five algorithms ranged from 0.82 (95% CI 0.79, 0.85; 9 studies with 97,815 participants) for neural network-based models to 0.95 (95% CI 0.93, 0.97; 7 studies with 161,334 participants) for random forest models (Table 3). The latter approach also had the highest summary specificity 0.96 (95% CI 0.80, 0.99), with sensitivity 0.72 (95% CI 0.26, 0.95). The highest summary sensitivity was observed for support vector machine models (0.81, 95% CI 0.69, 0.90) with summary specificity 0.81 (95% CI 0.59, 0.93) (9 studies, 152,068 participants). The remaining algorithms had summary sensitivities ranging from 0.66 (decision tree models) to 0.73 (neural network models) (Table 3). Two additional ML algorithms evaluated in the included studies (Bayesian networks and LOS (abbreviation not explained)) had too few studies to allow meta-analysis (Appendix 5).

Other scales

In addition to the risk prediction tools reported above, Zhang and colleagues[40] reported on the EVARUCI scale[55], presenting summary sensitivity and specificity of 0.84 (95% CI 0.79, 0.89) and 0.68 (95% CI 0.66, 0.70), respectively (3 studies; 3,063 participants). These results were pooled across thresholds, 11 and 11.5 (one not reported).

Beyond the results covered by our five prioritised reviews, three further modifications of the Braden scale were evaluated in statistical syntheses: Braden modified by Kwong[60], the 4-factor model[61] and 'extended Braden'[61], revealing variable performance with high uncertainty.[32 29 42] Another two modified versions of the Norton scale (by Ek[62], and by Bienstein[63]) were also included in one review's meta-analyses[32], but only risk ratios were reported. Three additional scales (revised "Jackson & Cubbin"[64], EMINA[65] and PSPS[66]) were evaluated in one statistical synthesis each.[32 29] Full details can be found in Appendix 5 Table S4.

Appendix 5 Table S5 reports data for another 17 risk prediction tools, each associated with a single primary study (therefore not covered in detail in the text above), and another two tools, Sunderland[67] and RAPS[68], which are assessed in two primary studies each.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Results from reviews evaluating the clinical effectiveness of risk prediction models

Table 4 provides an overview of results from four[29 45 47 50 69] of the 10 reviews reporting clinical effectiveness, including one Cochrane review[47] which identified two randomised controlled trials (RCTs) of risk prediction tools and assessed risk of bias using the Cochrane RoB tool[48] . The remaining reviews used broader eligibility criteria for study inclusion and a range of different quality assessment tools, with some reviews reaching varying conclusions about the methodological quality of the same studies. Given the overlap in study inclusion between reviews, a summary of the included comparative studies is provided below.

One individually randomised trial (Webster and colleagues[70]) and one cluster randomised trial (Saleh and colleagues[71]) were considered to be at high risk of bias by the Cochrane review authors. The individually randomised trial[70] was included in three additional reviews[29 44 47 50], each of which considered the trial to be 'good quality'[29], 'valid'[44], or 'high quality'[50]. The trial was conducted in 1,231 hospital inpatients and found no evidence of a difference in PI incidence between patients assessed with either the Waterlow scale or Ramstadius tool compared with clinical judgment alone (RR 1.10, 95% CI 0.68, 1.81 for Waterlow and RR 0.79, 95% CI 0.46, 1.35 for Ramstadius). The trial further showed no evidence of a difference in patient management or in PI severity when using a risk assessment tool compared to clinical judgement.

The cluster randomised trial[71] was considered to be of poor methodological quality in two reviews.[29 47] The trial included 521 patients at a military hospital and compared nurse training with mandatory use of the Braden scale, to nurse training and optional use of the Braden scale, to no training. No evidence of a difference in PI incidence was observed between groups: incidence rates were 22%, 22% and 15% (p=0.38), for the three groups respectively.

In both reviews by Lovegrove and colleagues,[45 50] an uncontrolled comparison study[72] was included. The study assessed the clinical effectiveness of the Maelor scale,[73] and was rated as high quality within the most recent review.[50] Preventive strategies and PI prevalence were compared across two sites, an Irish hospital that used the Maelor scale (121 patients) and a Norwegian hospital that used nurses' clinical judgement (59 patients). A higher rate of preventive strategies, as well as a lower PI prevalence (12% vs. 54%), was reported for the Irish hospital. However, these results are likely to be highly confounded by inherent differences in population and setting.

A non-randomised study by Gunningberg and colleagues[74] was included in two reviews, one of which is reported in Table 4[33 69] and was considered to be of relatively high quality. The study was conducted in 124 patients in emergency and orthopaedic units and compared the use of a pressure ulcer risk alarm sticker for patients with a modified Norton Score of <21 (indicating high-risk patients) to standard care. No significant difference in the incidence of pressure ulcers between the Norton scale and standard care groups was observed.

A non-randomised study[75] conducted in 233 hospice inpatients was included in three reviews,[29 33 69] one of which is reported in Table 4.[69] The study met six of eight quality criteria used by Health Quality Ontario.[69] Use of a modified version of the Norton scale (Norton modified by Bale), in conjunction with standardised use of preventive interventions based on risk score, was found to be associated with lower risk of pressure ulcers when compared with nurses' clinical judgment alone (RR 0.11, 95% CI 0.03, 0.46). The lack of randomisation limits the reliability of this result, and review authors report that the modified Norton scale had not been validated.

Finally, a 'before-and-after' study[76] of 181 patients in various hospital settings was included in two reviews. [33 69] The Health Quality Ontario considered the study to meet all quality criteria.[69] Use of the

375 *Table 4. Systematic reviews evaluating clinical effectiveness*

Review author (publication year)	**Models included**	**Setting of included studies; study design; sample size**	**Included outcomes**	**Brief description of study quality**	**Relevant results from included studies**
Lovegrove[50] (2021)	Braden; Maelor score; Norton; Ramstadius; Waterlow	Acute care hospital n=1, inpatient units n=1, ICU n=1, internal medicine and oncology wards n=1; Design: cross-sectional survey n=2, RCT n=1, observational inter-rater reliability n=1; Sample size 45 to 1231	PI risk scores; PI incidence; PI preventative interventions; interrater reliability (reliability results covered in Appendix 5)	RoB assessed using JBI tools or analytical cross-sectional study appraisal checklist. The RCT was judged as high quality. Of the remaining studies, two were judged as high quality and one as moderate quality; inclusion criteria not clearly stated and no strategies to deal with confounding.	• There were no differences in patient management ('pressure care plan' and use of a special mattress) based on PI risk assessment method (clinical judgement, Ramstadius tool or Waterlow score). PI incidence difference between groups not significant (p = 0.44) (Webster 2011[70]). • A hospital that used the Maelor scale reported a higher rate of PI preventive strategies, and a lower PI prevalence (12% vs. 54%), than a site that used nurses' clinical judgement (Moore 2015[72]).
Moore[47] (2019)	Braden; Waterlow; Ramstadius	Military hospital n=1, internal medicine and oncology wards n=1; Design: RCT n=1, cluster randomised trial n=1; Sample sizes 286 and 1231	PI incidence; severity of PIs	RoB assessed using Cochrane tool (Higgins 2011[48]). Both studies at high RoB due to blinding issues. One study at RoB also due to baseline imbalance and incorrect analyses.	• No differences in PI incidence when using Braden scale or clinical judgement (Braden vs. clinical judgement+training, RR 0.97, 95% CI 0.53-1.77; Braden vs clinical judgement RR 1.43, 95% CI 0.77-2.68) (Saleh 2009[71]). • No difference in PI incidence when using a risk assessment tool compared to clinical judgement (RR 1.10, 95% CI 0.68-1.81 and RR 0.79, 95% CI 0.46-1.35, for Waterlow and Ramstadius respectively) (Webster 2011[70]). • No difference in PI severity based on risk assessment tools vs. clinical judgement (Webster 2011[70]).
Chou[29] (2013)	Norton modified by Bale; Braden; Waterlow; Ramstadius	Hospital n=2, hospice n=1; Design: non-randomised n=1, cluster randomised trial n=1, RCT n=1; Sample size 240 to 1231	PI incidence, severity of PIs; PI preventative interventions	RoB assessed with criteria consistent with AHRQ Methods Guide for Effectiveness and Comparative Effectiveness Reviews. One RCT was rated as good quality and the other as poor due to randomisation and blinding issues. The cohort study was rated as poor; there were blinding issues and confounding was not investigated.	• No difference in PI incidence when using a risk assessment tool compared to clinical judgement (RR 1.10, 95% CI 0.68-1.81 and RR 0.79, 95% CI 0.46-1.35, for Waterlow and Ramstadius respectively) (Webster 2011[70]). • The modified version of the Norton scale with use of preventive interventions is associated with lower risk of PIs compared with clinical judgment (RR 0.11, 95% CI 0.03-0.46) (Bale 1995[75]). • No difference in risk of PIs when one of three interventions was used (22% vs. 22% vs. 15%, p=0.38 for nurse training+mandatory Braden scale, nurse training+optional Braden scale and no training respectively) (Saleh 2009[71]).

Review author (publication year)	**Models included**	**Setting of included studies; study design; sample size**	**Included outcomes**	**Brief description of study quality**	**Relevant results from included studies**
Health Quality Ontario[69] (2009)	Norton; Norton modified by Bale; Norton modified by Ek 97	Hip fracture inpatients n=1, palliative care/hospice n=1, neurosurgery, general medicine, orthopaedic, and oncology units n=1; Design: prospective controlled (contemporaneous controls) n=1, before-and-after n=1; Sample size 124 to 223	PI incidence; PI preventative interventions	RoB assessment criteria name not given. Two studies met 6/8 and one study met all quality assessment requirements. In the studies that didn't meet all requirements, there were blinding and loss to follow-up issues. One study used a version of the Norton scale that was not validated.	• Compared a strategy that gave high-risk patients (based on modified Norton score) a risk alarm sticker to standard care. No significant difference between the groups in the incidence of PIs (Gunningberg 1999[74]). • Compared a strategy where patients received a pressure support system allocated according to the modified Norton scale to one where the nurse chose whether to give a special mattress. Using the scale significantly reduced the incidence of PIs (22.4% vs. 2.5%, p<0.0001) (Bale 1995[75]). • Compared the Norton scale with training to standard care. There was a significant difference in the number of preventative interventions (18.96 vs. 10.75, for Norton and usual care respectively). Interventions were used earlier for Norton vs. usual care (on day 1, 61% vs. 50%, p<0.002). No significant difference in the incidence of PIs between the groups (Hodge 1990[76]).

AHRQ – Agency for Healthcare Research; CASP – Critical Appraisal Skills Checklist; CI – confidence interval; ICU – intensive care unit; JBI – Joanna Briggs Institute; PI – pressure injury; RCT – randomised controlled trial; RoB – risk of bias; RR – Risk Ratio; S.S. – Suriadi Sanada Scale.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Norton scale with additional training for staff was associated with significant differences in the number of preventative interventions prescribed compared to standard care (18.96 vs. 10.75, respectively). Preventative interventions were also introduced earlier in the intervention group (on day 1, 61% vs. 50%, P<0.002 for Norton and usual care, respectively). However, no significant difference in the incidence of PIs was detected between the groups.

DISCUSSION

This umbrella review summarises data from a total of 23 systematic reviews of studies evaluating the clinical utility of a total of 63 PI risk prediction tools. Despite the large number of available reviews, quality assessment using an adaptation of AMSTAR-2 suggested that the majority were conducted to a relatively poor standard or did not meet reporting standards for systematic reviews.[14 26] Of the 15 items included in AMSTAR-2, only two (for accuracy reviews) and four (for effectiveness reviews) criteria were more consistently met (more than 60% of reviews scoring 'Yes'). All other criteria were fully met by less than half of reviews. The primary studies included in the reviews were particularly poorly described in the accuracy reviews, making it difficult to determine exactly what was evaluated and in whom. In particular, the source of the data was poorly reported. Only one review[33] explicitly restricted to accuracy estimates from external validations, and only one review[29] described whether estimates were sourced from training or external data. The extent to which we could reliably describe and comment on the content of the reviews is limited and high-quality evidence for the accuracy and clinical effectiveness of PI risk prediction models may be lacking.

Prognostic accuracy of risk prediction models

Of the 16 reviews focused on the predictive accuracy of included models, only two used appropriate methods for both quality assessment and statistical synthesis of accuracy data[36 40], one of which[36] evaluated only the Braden scale. Only one review[33] pre-specified the exclusion of studies reporting tool development only, one review restricted to "validated risk assessment instruments" only[29], and none of the reviews discussed the importance of appropriate validation of prediction models. Only two reviews conducted meta-analyses at different cut-offs for determination of high risk[29 36]; the remaining reviews combined data regardless of the threshold used. Combining data across different thresholds to estimate summary sensitivity and specificity is discouraged as it yields clinically uninterpretable and non-generalisable estimates, because the estimates do not relate to a particular threshold.[25]

Results of meta-analyses suggested that risk prediction scales have moderate sensitivities and somewhat lower specificities, typically in the range of around 70% to 85% for sensitivity and as low as 30% to 40% for specificity for some tools. Without a detailed review of the primary study publications for these models, it is not possible to assess which, if any, of these risk assessment scales might outperform the others. It seems that limited comparative studies comparing the accuracy of different tools are available.

For the ML-based models, one review[30] meta-analysed accuracy data by algorithm type. The results of the meta-analyses are not informative for clinical practice but may be a useful way of identifying which ML algorithms may be more suited to the data. Results suggested that specificities for random forest or decision tree models could reach 90% or above with associated sensitivities in the range of 66% to 72%, however relatively wide confidence intervals around these summary estimates reflect considerable variation in model performance. Moreover, some of these estimates came from internal validations within model development studies, and may not be transferable to other settings.[77] Authors should make it clear where accuracy estimates are derived from to avoid overinterpretation of results.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Clinical effectiveness of risk prediction scales

Prediction models, like any test used for diagnostic or prognostic purposes, require evaluation in the care pathway to identify the extent to which their use can impact on health outcomes.[78] Of the 10 reviews assessing clinical effectiveness of PI risk prediction tools, the only primary studies suggesting potential patient benefits from the use of risk prediction tools,[72 75 76] were non-randomised and are likely to be at high risk of bias. In contrast, two randomised trials,[70 71] (both considered at high risk of bias by the Cochrane review[47]) suggest that use of structured risk assessment tools does not reduce the incidence of PIs. We should recognise that effectiveness outcomes largely depend on the availability and efficacy of preventative measures, and conclusions regarding the clinical effectiveness of these risk assessment tools cannot be confidently drawn from the limited evidence available. All reviews included studies that assessed the use of risk assessment scales developed by experts, and no evidence is available evaluating the clinical effectiveness of empirically derived prediction models or ML algorithms.

Other existing evidence

Moore and colleagues[47] recently updated their review (published after our search was conducted[79]) and reported no new randomised trials that assessed the effect of risk assessment tools on PI incidence.

We have separately reviewed[9] available evidence for the development and validation of risk prediction tools for PI occurrence. Almost half (52/116, 45%) of available tools were developed using ML methods (as defined by review authors), 40% (46/116) were based on clinical expertise or unclear methods, and only 18 (16%) were identified as having used statistical modelling methods. The reviews varied in methodological quality and reporting; however, the reporting of prediction model development in the original primary studies appears to be poor. For example, across all prediction tools identified, the internal validation approach was unclear and unidentifiable for 70% (81/116) of tools, and only one review identified and included external validation studies (n=7 studies).

ML-based models may have potential for identifying those at risk of PI, as suggested by one review[30] included in this umbrella review. However, it is important to consider the lack of transparency in reporting of model development methods and model performance, and the concerning lack of model validation in populations outside of the original model development sample.[9]

Strengths and limitations

We have conducted the first umbrella review that summarise the prognostic accuracy and clinical effectiveness of prediction models for risk of PI. We followed Cochrane guidance[13], with a highly sensitive search strategy designed by an experienced information specialist. Although we excluded non-English publications due to time and resource constraints, where possible these publications were used to identify additional eligible risk prediction models. To some extent, our review is limited by the use of AMSTAR-2 for quality assessment of included reviews. AMSTAR-2 was not designed for assessing systematic reviews of diagnostic or prognostic studies. Although we made some adaptations, many of the existing and amended criteria relate to the quality of reporting of the reviews as opposed to methodological quality. There is scope for further work to establish criteria for assessing systematic reviews of prediction models.

The primary limitation of our study lies in the limited detail available on risk prediction tools and their performance within the included systematic reviews. To ensure comprehensive model identification, we adopted a broad definition of 'systematic,' potentially influencing the depth of information provided in the reviews, and the reporting quality in many primary studies contributing

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

to these reviews may be suboptimal. Notably, excluding ML-based models, over half of the existing risk prediction tools were published prior to 2000, before the publication of original versions of reporting guidelines for diagnostic accuracy studies[80] and risk prediction models.[81]

CONCLUSIONS

In conclusion, this umbrella review comprehensively summarises the prognostic accuracy and clinical effectiveness of risk prediction tools for developing PIs. The included systematic reviews used poor methodology and reporting, limiting our ability to reliably describe and evaluate their content. ML-based models demonstrated potential, with high specificity reported for some models. Wide confidence intervals highlight the variability in current evaluations, and external validation of ML tools may be lacking. The prognostic accuracy of clinical scales and statistically derived prediction models has a substantial range of specificities and sensitivities, motivating further model development with high quality data and appropriate statistical methods.

Regarding clinical effectiveness, a reduction of PI incidence is unclear due the overall uncertainty and potential biases in available studies. This underscores the need for further research in this critical area, once promising prediction tools have been developed and appropriately validated. In particular, the clinical impact of newer ML-based models currently remains largely unexplored. Despite these limitations, our umbrella review provides valuable insights into the current state of PI risk prediction tools, emphasising the need for robust research methods to be used in future evaluations.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Acknowledgements

We would like to thank Mrs. Rosie Boodell (University of Birmingham, UK) for her help in acquiring the publications necessary to complete this piece of work.

Author Contributions

Conceptualisation: Bethany Hillier, Katie Scandrett, April Coombe, Tina Hernandez-Boussard, Ewout Steyerberg, Yemisi Takwoingi, Vladica Velickovic, Jacqueline Dinnes

Data curation: Bethany Hillier, Katie Scandrett, April Coombe, Jacqueline Dinnes

Formal analysis: Bethany Hillier, Katie Scandrett, Jacqueline Dinnes

Funding acquisition: Yemisi Takwoingi, Vladica Velickovic, Jacqueline Dinnes

Investigation: Bethany Hillier, Katie Scandrett, April Coombe, Yemisi Takwoingi, Jacqueline Dinnes

Methodology: Bethany Hillier, Katie Scandrett, April Coombe, Tina Hernandez-Boussard, Ewout Steyerberg, Yemisi Takwoingi, Vladica Velickovic, Jacqueline Dinnes

Project administration: Bethany Hillier, Yemisi Takwoingi, Jacqueline Dinnes

Resources: Bethany Hillier, Katie Scandrett

Supervision: Yemisi Takwoingi, Jacqueline Dinnes

Writing – original draft: Bethany Hillier, Katie Scandrett, April Coombe, Jacqueline Dinnes

Writing – review & editing: Bethany Hillier, Katie Scandrett, April Coombe, Tina Hernandez-Boussard, Ewout Steyerberg, Yemisi Takwoingi, Vladica Velickovic, Jacqueline Dinnes

Funding

This work was commissioned and supported by Paul Hartmann AG (Heidenheim, Germany). The contract with the University of Birmingham was agreed on the legal understanding that the authors had the freedom to publish results regardless of the findings.

YT, JD, BH, KS and AC are funded by the National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre (BRC). This paper presents independent research supported by the NIHR Birmingham BRC at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflicting Interests

I have read the journal's policy and the authors of this manuscript have the following competing interests: VV is an employee of Paul Hartmann AG; ES and THB received consultancy fees from Paul Hartmann AG. All other authors received no personal funding or personal compensation from Paul Hartmann AG and have declared that no competing interests exist.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

References

1. Li Z, Lin F, Thalib L, et al. Global prevalence and incidence of pressure injuries in hospitalised adult patients: A systematic review and meta-analysis. *International Journal of Nursing Studies* 2020;105:103-546. doi: 10.1016/j.ijnurstu.2020.103546
2. Padula WV, Delarmente BA. The national cost of hospital-acquired pressure injuries in the United States. *Int Wound J* 2019;16(3):634-40. doi: 10.1111/iwj.13071 [published Online First: 2019/01/28]
3. Sullivan N, Schoelles K. Preventing In-Facility Pressure Ulcers as a Patient Safety Strategy. *Annals of Internal Medicine* 2013;158(5.2):410-16. doi: 10.7326/0003-4819-158-5-201303051-00008
4. Institute for Quality and Efficiency in Health Care (IQWiG). Preventing pressure ulcers. Cologne, Germany 2006 [updated 2018 Nov 15. Available from: https://www.ncbi.nlm.nih.gov/books/NBK326430/?report=classic accessed Feb 2023].
5. Braden B, Bergstrom N. A Conceptual Schema for the Study of the Etiology of Pressure Sores. *Rehabilitation Nursing* 1987;12(1):8-16. doi: 10.1002/j.2048-7940.1987.tb00541.x
6. Bergstrom N, Braden BJ, Laguzza A, et al. The Braden Scale for Predicting Pressure Sore Risk. *Nurs Res* 1987;36(4):205-10.
7. Norton D. Geriatric nursing problems. *Int Nurs Rev* 1962;9:39-41.
8. Waterlow J. Pressure sores: a risk assessment card. *Nursing Times* 1985;81:49-55.
9. Hillier B, Scandrett K, Coombe A, et al. Development and validation of risk prediction tools for pressure injury occurrence: An umbrella review (pre-print). *MedRxiv* 2024 doi: TBC (UPDATE ME)
10. Šimundić AM. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC* 2009;19(4):203-11. [published Online First: 2009/01/20]
11. Leeflang MM, Rutjes AW, Reitsma JB, et al. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ* 2013;185(11):E537-44. doi: 10.1503/cmaj.121286 [published Online First: 2013/06/24]
12. Maiga A, Farjah F, Blume J, et al. Risk Prediction in Clinical Practice: A Practical Guide for Cardiothoracic Surgeons. *Ann Thorac Surg* 2019;108(5):1573-82. doi: 10.1016/j.athoracsur.2019.04.126 [published Online First: 2019/06/27]
13. Pollock M, Fernandes RM BL, Pieper D, Hartling L,. Chapter V: Overviews of Reviews. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA ed. Cochrane Handbook for Systematic Reviews of Interventions version 63 (updated February 2022). Available from www.training.cochrane.org/handbook: Cochrane 2022.
14. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA* 2018;319(4):388-96. doi: 10.1001/jama.2017.19163
15. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc* 2001;8(4):391-7. doi: 10.1136/jamia.2001.0080391 [published Online First: 2001/06/22]
16. Wilczynski NL, Haynes RB. Optimal Search Strategies for Detecting Clinically Sound Prognostic Studies in EMBASE: An Analytic Survey. *Journal of the American Medical Informatics Association* 2005;12(4):481-85. doi: 10.1197/jamia.M1752
17. Geersing G-J, Bouwmeester W, Zuithoff P, et al. Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. *PLOS ONE* 2012;7(2):e32844. doi: 10.1371/journal.pone.0032844
18. NHS. Pressure ulcers: revised definition and measurement. Summary and recommendations 2018 [Available from: https://www.england.nhs.uk/wp-content/uploads/2021/09/NSTPP-summary-recommendations.pdf accessed Feb 2023].
19. Haesler E. European Pressure Ulcer Advisory Panel, National Pressure Injury Advisory Panel and Pan Pacific Pressure Injury Alliance. Prevention and Treatment of Pressure Ulcers/Injuries: Clinical Practice Guideline. 2019 [Available from: https://internationalguideline.com/2019 accessed Feb 2023].

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

20. AHCPR. Pressure ulcer treatment. : Agency for Health Care Policy and Research 1994:1-25.
21. Harker J. Pressure ulcer classification: the Torrance system. *Journal of Wound Care* 2000;9(6):275-77. doi: 10.12968/jowc.2000.9.6.26233
22. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLOS Medicine* 2014;11(10):e1001744. doi: 10.1371/journal.pmed.1001744
23. Cochrane. DE form example prognostic models - scoping review: The Cochrane Collaboration: The Prognosis Methods Group; [Available from: https://methods.cochrane.org/prognosis/tools accessed Feb 2023].
24. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* 2017;358:j4008. doi: 10.1136/bmj.j4008
25. P M, Y T, JJ D, et al. Understanding meta-analysis. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors) Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 ed: Cochrane, 2023 (updated July 2023).
26. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine* 2009;6(7):e1000097. doi: 10.1371/journal.pmed.1000097
27. Munoz N, Posthauer ME. Nutrition strategies for pressure injury management: Implementing the 2019 International Clinical Practice Guideline. *Nutrition in Clinical Practice* 2022;37(3):567-82.
28. Park SH, Lee YS, Kwon YM. Predictive Validity of Pressure Ulcer Risk Assessment Tools for Elderly: A Meta-Analysis. *Western journal of nursing research* 2016;38:459-83. doi: 10.1177/0193945915602259
29. Chou R, Dana T, Bougatsos C, et al. Pressure ulcer risk assessment and prevention: a systematic comparative effectiveness review. *Annals of internal medicine* 2013;159(1):28-38.
30. Qu C, Luo W, Zeng Z, et al. The predictive effect of different machine learning algorithms for pressure injuries in hospitalized patients: A network meta-analyses. *Heliyon* 2022;8(11):e11361. doi: 10.1016/j.heliyon.2022.e11361
31. Wang N, Lv L, Yan F, et al. Biomarkers for the early detection of pressure injury: A systematic review and meta-analysis. *Journal of Tissue Viability* 2022;31:259-67. doi: 10.1016/j.jtv.2022.02.005
32. García-Fernández FP, Pancorbo-Hidalgo PL, Agreda JJS. Predictive Capacity of Risk Assessment Scales and Clinical Judgment for Pressure Ulcers: A Meta-analysis. *Journal of Wound Ostomy & Continence Nursing* 2014;41(1):24-34. doi: 10.1097/01.WON.0000438014.90734.a2
33. Pancorbo-Hidalgo PL, Garcia-Fernandez FP, Lopez-Medina IM, et al. Risk assessment scales for pressure ulcer prevention: a systematic review. *J Adv Nurs* 2006;54(1):94-110. doi: 10.1111/j.1365-2648.2006.03794.x
34. Chen HL, Shen WQ, Liu P. A Meta-analysis to Evaluate the Predictive Validity of the Braden Scale for Pressure Ulcer Risk Assessment in Long-term Care. *Ostomy/wound management* 2016;62(9):20-8.
35. He W, Liu P, Chen HL. The Braden Scale cannot be used alone for assessing pressure ulcer risk in surgical patients: a meta-analysis. *Ostomy/wound management* 2012;58:34-40.
36. Huang C, Ma Y, Wang C, et al. Predictive validity of the braden scale for pressure injury risk assessment in adults: A systematic review and meta-analysis. *Nursing open* 2021;8:2194-207. doi: 10.1002/nop2.792
37. Park SH, Choi YK, Kang CB. Predictive validity of the Braden Scale for pressure ulcer risk in hospitalized patients. *Journal of Tissue Viability* 2015;24:102-13. doi: 10.1016/j.jtv.2015.05.001

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

38. Wei M, Wu L, Chen Y, et al. Predictive Validity of the Braden Scale for Pressure Ulcer Risk in Critical Care: A Meta-Analysis. *Nursing in critical care* 2020;25:165-70. doi: 10.1111/nicc.12500
39. Wilchesky M, Lungu O. Predictive and concurrent validity of the Braden scale in long-term care: A meta-analysis. *Wound Repair and Regeneration* 2015;23:44-56. doi: 10.1111/wrr.12261
40. Zhang Y, Zhuang Y, Shen J, et al. Value of pressure injury assessment scales for patients in the intensive care unit: Systematic review and diagnostic test accuracy meta-analysis. *Intensive & critical care nursing* 2021;64:103009. doi: 10.1016/j.iccn.2020.103009
41. Kottner J, Dassen T, Tannen A. Inter- and intrarater reliability of the Waterlow pressure sore risk scale: A systematic review. *International Journal of Nursing Studies* 2009;46:369-79. doi: 10.1016/j.ijnurstu.2008.09.010
42. Tayyib NAH, Coyer F, Lewis P. Pressure ulcers in the adult intensive care unit: a literature review of patient risk factors and risk assessment scales. *Journal of Nursing Education and Practice* 2013;3(11):28-42.
43. Baris N, Karabacak BG, Alpar SE. The Use of the Braden Scale in Assessing Pressure Ulcers in Turkey: A Systematic Review. *Advances in skin & wound care* 2015;28:349-57. doi: 10.1097/01.ASW.0000465299.99194.e6
44. Gaspar S, Peralta M, Marques A, et al. Effectiveness on hospital-acquired pressure ulcers prevention: a systematic review. *International Wound Journal* 2019;16(5):1087-102. doi: 10.1111/iwj.13147
45. Lovegrove J, Miles S, Fulbrook P. The relationship between pressure ulcer risk assessment and preventative interventions: a systematic review. *Journal of wound care* 2018;27(12):862-75.
46. Kelly J. Inter-rater reliability and Waterlow's pressure ulcer risk assessment tool. *Nurs Stand* 2005;19(32):86-7, 90-2. doi: 10.7748/ns2005.04.19.32.86.c3851
47. Moore ZEH, Patton D. Risk assessment tools for the prevention of pressure ulcers. *Cochrane Database of Systematic Reviews* 2019 doi: 10.1002/14651858.CD006471.pub4
48. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928. doi: 10.1136/bmj.d5928
49. AHRQ Methods for Effective Health Care. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US) 2008.
50. Lovegrove J, Ven S, Miles SJ, et al. Comparison of pressure injury risk assessment outcomes using a structured assessment tool versus clinical judgement: A systematic review. *Journal of Clinical Nursing* 2021 doi: 10.1111/jocn.16154 [published Online First: 2021/12/01]
51. Song M, Choi KS. Factors predicting development of decubitus ulcers among patients admitted for neurological problems. *The Journal of Nurses Academic Society* 1991;21(1):16-26.
52. Pang SM, Wong TK. Predicting pressure sore risk with the Norton, Braden, and Waterlow scales in a Hong Kong rehabilitation hospital. *Nursing Research* 1998;47(3):147-53.
53. Park SH, Lee HS. Assessing Predictive Validity of Pressure Ulcer Risk Scales- A Systematic Review and Meta-Analysis. *Iranian journal of public health* 2016;45(2):122-33.
54. Cubbin B, Jackson C. Trial of a pressure area risk calculator for intensive therapy patients. *Intensive Care Nursing* 1991;7(1):40-44.
55. González-Ruiz J, Carrero AG, Blázquez MH, et al. Factores de riesgo de las úlceras por presión en pacientes críticos. *Enfermería Clinica* 2001;11(5):184-90.
56. Macaskill P, Gatsonis C, Deeks J, et al. Cochrane handbook for systematic reviews of diagnostic test accuracy: Version, 2010.
57. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;21(9):1237-56. doi: 10.1002/sim.1099
58. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;12(14):1293-316. doi: 10.1002/sim.4780121403

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

59. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993;13(4):313-21. doi: 10.1177/0272989x9301300408
60. Kwong E, Pang S, Wong T, et al. Predicting pressure ulcer risk with the modified Braden, Braden, and Norton scales in acute care hospitals in Mainland China. *Appl Nurs Res* 2005;18(2):122-8. doi: 10.1016/j.apnr.2005.01.001
61. Halfens R, Van Achterberg T, Bal R. Validity and reliability of the Braden scale and the influence of other risk factors: a multi-centre prospective study. *International Journal of Nursing Studies* 2000;37(4):313-19.
62. Ek AC. Prediction of pressure sore development. *Scand J Caring Sci* 1987;1(2):77-84. doi: 10.1111/j.1471-6712.1987.tb00603.x
63. Bienstein C. Risikopatienten erkennen mit der erweiterten Nortonskala [Risk patients detected with the extended Norton scale]. Dekubitus - Prophylaxe undTherapie. Frankfurt/Main: Verlag Krankenpflege 1991.
64. Jackson C. The revised Jackson/Cubbin Pressure Area Risk Calculator. *Intensive Crit Care Nurs* 1999;15(3):169-75. doi: 10.1016/s0964-3397(99)80048-2
65. Fuentelsaz C. Validation of the EMINA scale: tool for the evaluation of risk of developing pressure ulcers in hospitalized patients. *Enferm Clin [Internet]* 2001;11(3):97-103.
66. Lowthian P. The practical assessment of pressure sore risk. *Care–Science and Practice* 1987;5(4):3-7.
67. Lowery MT. A pressure sore risk calculator for intensive care patients: 'the Sunderland experience'. *Intensive Crit Care Nurs* 1995;11(6):344-53. doi: 10.1016/s0964-3397(95)80452-8
68. Lindgren M, Unosson M, Krantz AM, et al. A risk assessment scale for the prediction of pressure sore development: reliability and validity. *Journal of advanced nursing* 2002;38(2):190-99.
69. Ontario HQ. Pressure ulcer prevention: an evidence-based analysis. *Ontario health technology assessment series* 2009;9(2):1-104.
70. Webster J, Coleman K, Mudge A, et al. Pressure ulcers: effectiveness of risk-assessment tools. A randomised controlled trial (the ULCER trial). *BMJ Quality & Safety* 2011;20(4):297. doi: 10.1136/bmjqs.2010.043109
71. Saleh M, Anthony D, Parboteeah S. The impact of pressure ulcer risk assessment on patient outcomes among hospitalised patients. *J Clin Nurs* 2009;18(13):1923-9. doi: 10.1111/j.1365-2702.2008.02717.x [published Online First: 2009/04/03]
72. Moore Z, Johansen E, Etten Mv, et al. Pressure ulcer prevalence and prevention practices: a cross-sectional comparative survey in Norway and Ireland. *Journal of Wound Care* 2015;24(8):333-39. doi: 10.12968/jowc.2015.24.8.333
73. Moore Z, Pitman S. Towards establishing a pressure sore prevention and management policy in an acute hospital setting. *The All Ireland Journal of Nursing and Midwifery* 2000;1(1):7-11.
74. Gunningberg L, Lindholm C, Carlsson M, et al. Implementation of risk assessment and classification of pressure ulcers as quality indicators for patients with hip fractures. *J Clin Nurs* 1999;8(4):396-406. doi: 10.1046/j.1365-2702.1999.00287.x
75. Bale S, Finlay I, Harding KG. Pressure sore prevention in a hospice. *J Wound Care* 1995;4(10):465-8. doi: 10.12968/jowc.1995.4.10.465
76. Hodge J, Mounter J, Gardner G, et al. Clinical trial of the Norton Scale in acute care settings. *Aust J Adv Nurs* 1990;8(1):39-46.
77. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245-7. doi: 10.1016/j.jclinepi.2015.04.005 [published Online First: 2015/04/18]
78. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98(9):691-8. doi: 10.1136/heartjnl-2011-301247 [published Online First: 2012/03/07]

medRxiv preprint doi: https://doi.org/10.1101/2024.05.07.24307001; this version posted May 8, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

79. Moore Z, Avsar P, O'Connor T, et al. A systematic review of movement monitoring devices to aid the prediction of pressure ulcers in at-risk adults. *International Wound Journal* 2023;20(2):579-608. doi: 10.1111/iwj.13902

80. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;138(1):W1-12. doi: 10.7326/0003-4819-138-1-200301070-00012-w1

81. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162(1):W1-73. doi: 10.7326/m14-0698