

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

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

Bethany Hillier^{1,2}

Katie Scandrett¹

April Coombe¹

Tina Hernandez-Boussard³

Ewout Steyerberg⁴

Yemisi Takwoingi^{1,2}

Vladica Velickovic^{5,6}

Jacqueline Dinnes^{1,2*}

Affiliations

¹ Department of Applied Health Sciences, College of Medicine and Health, 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NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

34 ABSTRACT

35 Background

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

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

43 Methods and Findings

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

48 We identified 19 reviews that assessed prognostic accuracy and 11 that assessed clinical
49 effectiveness of risk prediction tools for PI. The 19 reviews of prognostic accuracy evaluated 70 tools
50 (39 scales and 31 machine learning models), with the Braden, Norton, Waterlow, Cubbin-Jackson
51 scales (and modifications thereof) the most evaluated tools. Meta-analyses from a focused set of
52 included reviews showed that the scales had sensitivities and specificities ranging from 53%-97% and
53 46%-84%, respectively. Only 2/19 reviews performed appropriate statistical synthesis and quality
54 assessment. Two reviews assessing machine learning based algorithms reported high prognostic
55 accuracy estimates, but some of which were sourced from the same data within which the models
56 were developed, leading to potentially overoptimistic results.

57 Two randomised trials assessing the effect of PI risk assessment tools (within the full test-
58 intervention-outcome pathway) on the incidence of PIs were identified from the 11 systematic
59 reviews of clinical effectiveness; both were included in a Cochrane review and assessed as high risk
60 of bias. Both trials found no evidence of an effect on PI incidence.

61 Conclusions

62 Available systematic reviews suggest a lack of high-quality evidence for the accuracy of risk
63 prediction tools for PI and limited reliable evidence for their use leading to a reduction in incidence
64 of PI. Further research is needed to establish the clinical effectiveness of appropriately developed
65 and validated risk prediction tools for PI.

66

67 Author Summary

68 Why was this study done?

- 69 • Pressure injuries (PIs) are injuries to and below the skin caused by prolonged pressure,
70 especially on bony areas, and people who spend extensive periods in a bed or chair are
71 particularly vulnerable.
- 72 • The majority of pressure injuries are preventable if appropriate preventive measures are put
73 into place, but it is crucial to conduct risk stratification of individuals in order to
74 appropriately allocate preventive measures.
- 75 • Numerous tools that give patients a score (or probability) to signify their risk of developing a
76 PI exist. However, there is a lack of clarity on how accurate the risk scores are, and how
77 effective the scores are at improving patient outcomes (the clinical effectiveness) when
78 patient management is subsequently changed for patients classified as high-risk.

79 What did the researchers do and find?

- 80 • We conducted an umbrella review (an overview of existing systematic reviews), identifying
81 26 systematic reviews which included 70 risk prediction tools.
- 82 • Of these 70 risk prediction tools, 31 were developed using machine learning methods, while
83 the remainder were derived from statistical modelling and/or clinical expertise.
- 84 • Risk prediction tools demonstrated moderate to high accuracy, as measured by a variety of
85 metrics. However, there were concerns regarding the quality of both the systematic reviews,
86 and the primary studies included in these reviews, as reported by the systematic review
87 authors.
- 88 • There were only two randomised controlled trials that investigated the clinical effectiveness
89 of risk prediction tools and subsequent changes in PI management, and neither trial found
90 that use of the tools had an impact on the incidence of PIs.

91 What do these findings mean?

- 92 • Whilst an abundance of risk prediction tools exists, it is unclear how accurate they are due to
93 poor quality evidence and poor reporting, so it is difficult to recommend a particular
94 tool/tools.
- 95 • Even if the tools are shown to be accurate, they are not useful unless they lead to
96 improvement in patient outcomes. There is very limited evidence to determine whether the
97 tools are clinically effective and the evidence that does exist suggests that the tools did not
98 lead to improved patient outcomes.
- 99 • More research into the clinical effectiveness of appropriately developed and evaluated tools,
100 when they are adopted within the clinical pathway, is needed.

101 INTRODUCTION

102 Pressure injuries (PI), also known as pressure ulcers or decubitus ulcers, have an estimated global
103 prevalence of 12.8% among hospitalised adults,¹ and place a significant burden on healthcare
104 systems (estimated at \$26.8 billion per year in the US alone²). PIs are most common in individuals
105 with reduced mobility, limited sensation, poor circulation, or compromised skin integrity, and can
106 affect those in community settings and long-term care as well as hospital settings. Effective
107 prevention of PI requires multicomponent preventive strategies such as mattresses, overlays, and
108 other support systems, nutritional supplementation, repositioning, dressings, creams, lotions, and
109 cleansers.^{3,4} Health economic models have suggested that providing baseline preventive
110 interventions for all with daily risk assessments is more cost-effective than either a less standardised
111 prevention protocol or a targeted risk-stratified prevention strategy.⁵ Nevertheless, the stratification
112 of patients by risk could further improve outcomes by allowing timely and targeted implementation
113 of additional or greater intensity preventive measures in those most at risk, to reduce harm and
114 consequently burden to healthcare systems.⁶

115 Numerous clinical assessment scales and statistical risk prediction models for assessing the risk of PI
116 are available. However, the methodology underlying their development is not always explicit, with
117 scales in routine clinical usage apparently based on epidemiological evidence and clinical judgment
118 about predictors that may not meet accepted principles for the development and reporting of risk
119 prediction models.⁷ The Braden^{8,9}, Norton¹⁰ and Waterlow¹¹ scales are recommended by NICE
120 guidelines¹² in the UK and referenced in international guidelines for PI prevention.¹³ In some
121 hospitals and long-term care settings in the US, healthcare professionals must conduct mandatory
122 risk assessments for PI for all patients for the purposes of risk stratification and clinical triage. The
123 Braden scale, developed in 1987 using a sample of 102 elderly hospital patients in the US includes
124 sensory perception, moisture, activity, mobility, nutrition, friction and shear as predictors.^{8,9} The
125 Norton scale, based on a sample of 250 elderly hospital patients in the UK and published in 1962,
126 includes physical condition, mental status, activity, mobility and continence domains.¹⁰ The Waterlow
127 scale was published in 1985 for use by Waterlow's nursing students in the UK¹⁴, and assesses BMI,
128 assessment of the skin, gender, age, malnutrition, incontinence, mobility, tissue malnutrition,
129 neurological deficits, major surgery or trauma and medication.¹¹

130 Despite the apparent lack of reporting of now standard methods for development and validation
131 (including external validation) of available risk prediction tools, there is a considerable body of
132 evidence evaluating their clinical utility, much of which has been synthesised in systematic reviews
133 and meta-analyses.⁷ Clinical utility includes both prognostic accuracy and clinical effectiveness.
134 Prognostic accuracy is estimated by applying a numeric threshold above (or below) which there is a
135 greater risk of PI, with study results presented using accuracy metrics such as sensitivity, specificity or
136 the area under the receiver operating characteristic (ROC) curve.¹⁵ Resulting accuracy is driven not
137 only by the nominated threshold for defining participants as at low or high risk for PI but by other
138 study factors including population and setting.¹⁶ Clinical effectiveness, or the ability of a tool to
139 ultimately impact on health outcomes such as the incidence or severity of PI, is related both to the
140 accuracy of the tool (or its ability to correctly identify those most likely to develop PI), to the uptake
141 and implementation of the tool in practice and to the consequential changes in PI management
142 based on tool predictions. Demonstrating a change in health outcomes as a result of use of a risk
143 prediction tool is vital to encourage implementation.¹⁷

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

147 METHODS

148 Protocol registration and reporting of findings

149 We followed Cochrane guidance for conducting umbrella reviews¹⁸, and ‘Preferred Reporting Items
150 for Systematic Reviews and Meta-Analyses of Diagnostic Test Accuracy Studies’ (PRISMA-DTA)
151 reporting guidelines¹⁹ (see Appendix 1 in S1 File). The protocol was registered on Open Science
152 Framework (<https://osf.io/tepyk>).

153 Literature search

154 Electronic searches of MEDLINE, Embase via Ovid and CINAHL Plus EBSCO from inception to June
155 2024 were developed and conducted by an experienced information specialist (AC), employing well-
156 established systematic review and prognostic search filters,²⁰⁻²² combined with appropriate keywords
157 related to PIs. Simplified supplementary searches in EPISTEMONIKOS and Google Scholar were also
158 undertaken, with the latter covering the years 2013 to June 2024 (see Appendix 2 in S1 File for
159 further details). Screening of search results and full texts were conducted independently and in
160 duplicate by any two from a group of four reviewers (BH, JD, YT, KS), with arbitration by a third
161 reviewer where necessary (any one of the four reviewers not involved in the independent screening).

162 Eligibility criteria for this umbrella review

163 Published English-language systematic reviews of risk prediction tools developed for adult patients at
164 risk of PI in any setting were included. Clinical risk assessment scales and models developed using
165 statistical or machine learning (ML) methods were eligible (models exclusively using pressure sensor
166 data were not considered). Risk prediction tools could be applied by any healthcare professional
167 using any threshold for classifying patients as high or low risk and using any PI classification system¹³
168 ²³⁻²⁵ as a reference standard. For prognostic accuracy, we required accuracy metrics, such as
169 sensitivity and specificity, to be presented but did not require full 2x2 classification tables to be
170 reported. Reviews on diagnosing or staging suspected or existing PIs were excluded.

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

176 Data extraction and quality assessment

177 Data extraction forms (Appendix 3) were informed by the CHARMS checklist (CHecklist for critical
178 Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) and Cochrane
179 Prognosis group template.^{26,27} Data extraction items included review characteristics, number of
180 studies and participants, study quality and results.

181 The methodological quality of included systematic reviews was assessed using an adapted version of
182 AMSTAR-2 (A Measurement Tool to Assess Systematic Reviews).²⁸ For example, for reviews evaluating
183 the prognostic accuracy of risk prediction tools we assessed eligibility criteria using the PIRT
184 framework (Population, Index test, Reference standard, Target condition)²⁹ and POII framework
185 (Population, Outcome to be predicted, Intended use of model, Intended moment in time)³⁰ and
186 required methodological quality assessment to be conducted using validated and appropriate tools
187 such as QUADAS³¹, QUADAS-2³² or PROBAST³³. We omitted the AMSTAR-2 item relating to
188 publication bias (Item 15) because of the lack of empirical evidence for the effect of publication bias
189 on test accuracy estimates, and limitations in statistical methods for identifying publication bias.^{19, 34}
190 Our adapted AMSTAR-2 contains six critical items, and limitations in any of these items reduces the

191 overall validity of a review.²⁸ Full details can be found in Appendix 4 in S1 File. Quality assessment
192 and data extraction were conducted by one reviewer and checked by a second (BH, JD, KS), with
193 disagreements resolved by consensus.

194 [Synthesis methods](#)

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

199 Considerable overlap in risk prediction tools and included primary studies was noted between
200 reviews. For risk prediction tools that were included in multiple meta-analyses, we focused our
201 synthesis on the review(s) with the most recent search date or most comprehensive (based on
202 number of included studies) and most robust estimate of prognostic accuracy (judged according to
203 the appropriateness of the meta-analytic method used, e.g. use of recommended hierarchical
204 approaches for test accuracy data³⁵). The prognostic accuracy of risk prediction tools that were
205 included in three or fewer reviews, was reported only if an appropriate method of statistical
206 synthesis¹⁸ was used.

207 For clinical effectiveness results, reviews with the most recent search date or most comprehensive
208 overview of available studies, that assessed PI incidence outcomes and that at least partially met
209 more of the AMSTAR-2 criteria²⁸ were prioritised for narrative synthesis.

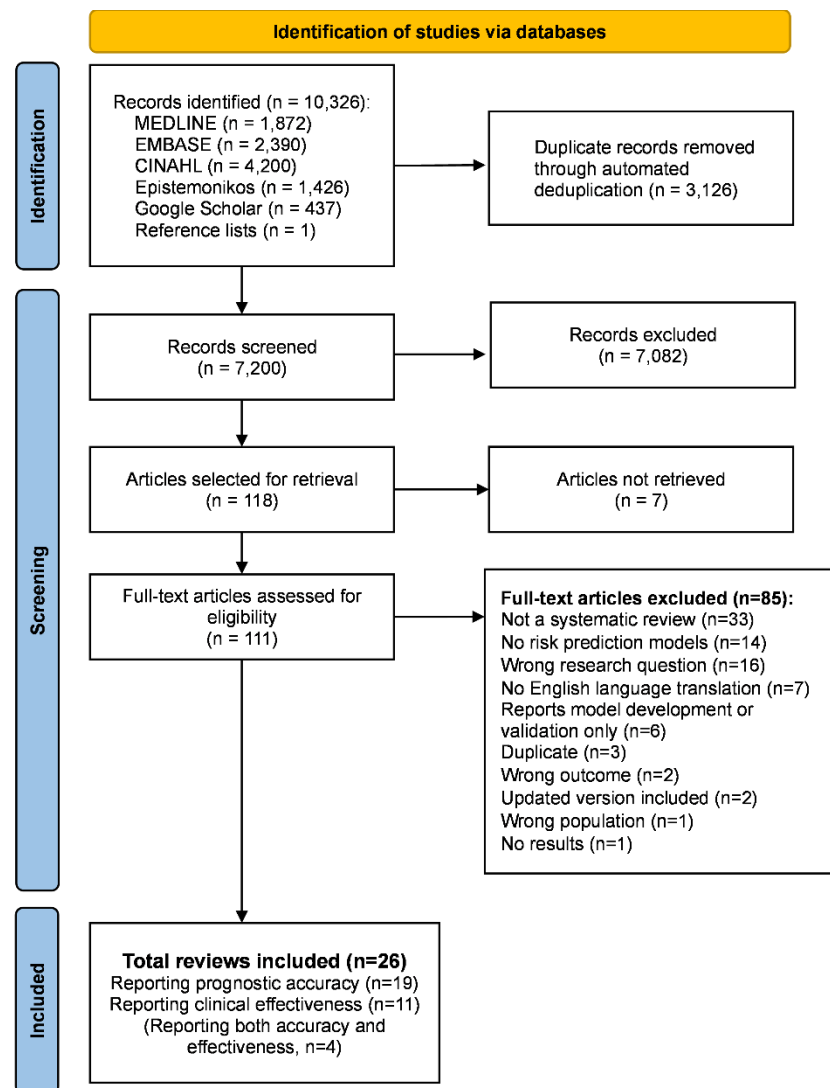
210 [RESULTS](#)

211 [Characteristics of included reviews](#)

212 A total of 118 records were selected for full-text assessment from 7200 unique records. We could
213 obtain the full text of 111 publications, of which 26 reviews met all eligibility criteria (Figure 1), 19
214 reported accuracy data³⁶⁻⁵⁴ and 11 reported clinical effectiveness data^{38 42 43 49 55-61} (four reported both
215 accuracy and effectiveness data^{38 42 43 49}). Tables 1-2 provide an overview of the characteristics,
216 methods and methodological quality of all 26 reviews (see Appendix 5 in S1 File for full details).

217

218 *Figure 1. PRISMA flowchart: identification, screening and selection process*



219

220 *List of full-text articles excluded, with reasons, is given in Appendix 5 in S1 File.*

221 Reviews were published between 2006 and 2024. Over half (15/26, 58%) restricted inclusion to adult
 222 populations (Table 1), two (8%) included any age group, and nine (35%) did not report any age
 223 restrictions. Six reviews (6/26, 23%) only included study populations with no PI at baseline. Acute
 224 care was the most frequent setting across both review questions (7/19 (37%) accuracy reviews and
 225 3/11 (27%) effectiveness reviews). Quality assessment tools included QUADAS-2 (n=8) or QUADAS
 226 (n=2) in more than half of reviews of accuracy (10/19, 53%). One review⁴⁷ utilised and reported
 227 PROBAST assessments for risk of bias. Another review⁴⁸ reported using QUADAS-2 and PROBAST
 228 tools in their methods, but only reported QUADAS-2 results.

229 Reviews of accuracy either included studies evaluating any tool (5/19, 26%) or pre-specified tools
 230 (10/19, 53%); two^{47 48} included only ML-based prediction models, while the remaining two^{49 50} did
 231 not specify the tools to be included. A total of 70 risk prediction tools were reported across the
 232 reviews (from one^{37 40 41 46 51 52} to 28³⁹ tools included per review), including 31 ML models. Only two
 233 reviews reported eligibility criteria related to the development or validation of the risk prediction
 234 tools. One⁴³ (6%) excluded evaluation studies that used the same data that was used to develop the

235 tool and the other³⁸ included only “validated risk assessment instruments” with no further definition
236 (yet included studies reporting original tool development).

237 The majority (15/19, 79%) of accuracy reviews conducted a meta-analysis, but only two utilised
238 currently recommended hierarchical approaches for the meta-analysis of test accuracy data.^{41 53} Eight
239 reviews conducted univariate meta-analysis of individual accuracy measures (e.g. sensitivity and
240 specificity separately, or area under the curve (AUC)⁵⁰, risk ratios (RR)³⁹ or odds ratio⁴³) and five did
241 not clearly report the type of analysis approach used.

242 Of the 11 systematic reviews evaluating clinical effectiveness, two only considered the reliability of
243 risk assessment scales^{49 58}, one considered reliability and other ‘psychometric’ properties⁴², and eight
244 considered effects on patient outcomes (one of which also considered tool reliability⁵⁵). More than
245 half of reviews (6, 55%) compared use of PI risk assessment scales to clinical judgement alone or
246 ‘standard care’. The number of included studies ranged from one⁵⁶ to 20⁶⁰, with sample sizes ranging
247 from one (one subject and 110 raters, in an inter-rater reliability study⁶²) to 4,137 patients. Reported
248 outcomes included the incidence of PIs (7/11), preventive interventions prescribed (5/11) and
249 interrater reliability (4/11), internal consistency, measurement error and convergent validity (1/11)
250 (latter four properties reported in Appendix 5 in S1 File). One review⁶¹ used the Cochrane risk of bias
251 (RoB) tool for quality assessment of included studies, and three used JBI (n=2) or CASP (n=1) tools.
252 Due to heterogeneity in study design, risk prediction tools and outcomes evaluated, none of the
253 included reviews provided any form of statistical synthesis of study results.

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

Review characteristic	Reviews on prognostic accuracy of risk prediction tools (N=19)	Reviews on clinical effectiveness of risk prediction tools (N=11)	All included reviews (N=26)
Median (range) year of publication	2016 (2006 – 2024)	2015 (2006 – 2024)	2017 (2006 – 2024)
Eligibility criteria			
Participants			
Adults only	11 (58) ^A	6 (55)	15 (58) ^A
Review states 'Any age'	1 (5)	1 (9)	2 (8)
No age restriction reported	7 (37)	4 (36)	9 (35)
Presence of PI at baseline			
Excluded those with PI at baseline ^B	5 (26)	2 (18)	6 (23)
NS	14 (74)	8 (73)	20 (77)
Setting			
Any healthcare setting	1 (5)	1 (9)	2 (8)
Hospital	3 (16)	0 (0)	3 (12)
Acute care (incl. surgical and ICU)	7 (37)	3 (27)	8 (31)
Hospital or acute care	0 (0)	2 (18)	2 (8)
Long-term care	2 (11)	0 (0)	2 (8)
Long-term, acute or community settings	1 (5)	1 (9)	1 (4)
NS	5 (26)	4 (36)	8 (31)
Risk assessment tools			
Any prediction tool or scale	5 (26)	6 (55)	9 (35)
Specified clinical scale(s)	10 (53)	3 (27)	12 (46)
ML-based prediction models	2 (11)	0 (0)	2 (8)
PI prevention strategies	0 (0)	1 (9)	1 (4)
NS	2 (11)	1 (9)	2 (8)
PI classification system			
Any PI classification system	1 (5)	0 (0)	1 (4)
Accepted standard classifications	1 (5)	1 (9)	2 (8)
Several specified classification systems (NPUAP, EPUAP, AHCP or TDCPS)	3 (16)	1 (9)	3 (12)
PI stage predefined ⁶³ /defined by study authors	1 (5)	0 (0)	1 (4)
NS	13 (68)	9 (82)	19 (73)
Source of data			
Prospective only	4 (21)	2 (18)	4.5 (17) ^C
Prospective or retrospective	1 (5)	2 (18)	2.5 (10) ^C
NS	14 (74)	7 (64)	19 (73)
Study design restrictions			
Yes	9 (47)	7 (64)	14 (54)
No	2 (11)	2 (18)	3 (12)
NS	8 (42)	2 (18)	9 (35)
Phase of development/evaluation of tools			
Development studies (with internal/external validation NS)	1 (5)	N/A	N/A
External evaluations only	2 (11)	N/A	N/A
Validation studies (internal or external NS)	1 (5)	N/A	N/A
NS	15 (79)	N/A	N/A
Review methods			
Median (range) no. sources^D searched	6 (2 – 14)	5 (3 – 14)	6 (2 – 14)
Publication restrictions:			

End date (year)			
2000-2009	1 (5)	3 (27)	3 (12)
2010-2019	12 (63)	6 (55)	16 (62)
2020-2023	6 (32)	2 (18)	7 (27)
Language			
English only	7 (37)	7 (64)	10 (38)
2 languages	1 (5)	2 (18)	3 (12)
>2 languages	2 (11)	2 (18)	3 (12)
No restrictions	3 (16)	1 (9)	4 (15)
NS	6 (32)	0 (0)	6 (23)
Quality assessment tool ^E			
PROBAST	1 (5) ^F	N/A	1 (4) ^F
QUADAS	2 (11)	N/A	2 (8)
QUADAS-2	8 (42)	N/A	8 (31)
JBI tools	1 (5)	2 (18)	3 (12)
CASP	2 (11)	1 (9)	2 (8)
Cochrane RoB tool	0 (0)	1 (9)	1 (4)
Other	2 (5)	5 (45)	6 (23)
None	3 (16)	2 (18)	4 (15)
Meta-analysis included	15 (79)	0 (0)	15 (58)
Method of meta-analysis (% of reviews incl. meta-analysis)			
Univariate RE/FE model (depending on heterogeneity assessment)	2 (13) ^G	N/A	N/A
Univariate RE model	6 (40) ^G	N/A	N/A
Hierarchical model (for DTA studies)	2 (13)	N/A	N/A
Unclear/NS	5 (33) ^G	N/A	N/A
Volume of evidence			
Median (range) no. studies	18 (2 – 70)	5 (1 – 20)	15 (1 – 70)
Median (range) no. participants	13,464 (609 – 408,504)	1,951 (528 – 5,052)	7,684 (528 – 408,504)
Median (range) no. tools	3 (1 – 28)	3 (1 – 9)	3 (1 – 28)

255 Figures are number (%) of reviews, unless otherwise specified. ^A one review⁴⁵ restricted to aged >60 years; ^B 'baseline'
256 refers to beginning of study or hospital admission depending on included reviews; ^C one review³⁸ states either prospective
257 or retrospective data eligible for Research Question 1, but prospective only for Research Question 2, hence 0.5 added to
258 each category; ^D including databases, bibliographies or registries; ^E reviews may fall into multiple categories, therefore total
259 number within domain not necessarily equal to N (100%); ^F another review⁴⁸ reported use of PROBAST in methods, but did
260 not present any PROBAST results, therefore not included; ^G one review conducts univariate meta-analysis for single
261 estimate, e.g. AUC⁵⁰, RR³⁹ or OR⁴³; AHCPR – Agency for Health Care Policy and Research; AUC – area under the curve; CASP
262 – Critical Appraisal Skills Programme; DTA – diagnostic test accuracy; EPUAP – European Pressure Ulcer Advisory Panel; FE –
263 fixed effects; ICU – intensive care unit; JBI – Joanna Briggs Institute; ML – machine learning; NPUAP – National Pressure
264 Ulcer Advisory Panel; NS – not stated; OR – odds ratio; PI – pressure injury; PROBAST – Prediction model Risk of Bias
265 Assessment; QUADAS (2) – Quality Assessment of Diagnostic Accuracy Studies (Version 2); RE – random effects; RoB – Risk
266 of Bias; RR – risk ratio; TDCPS – Torrance Developmental Classification of Pressure Sore.
267

268 **Methodological quality of included reviews**

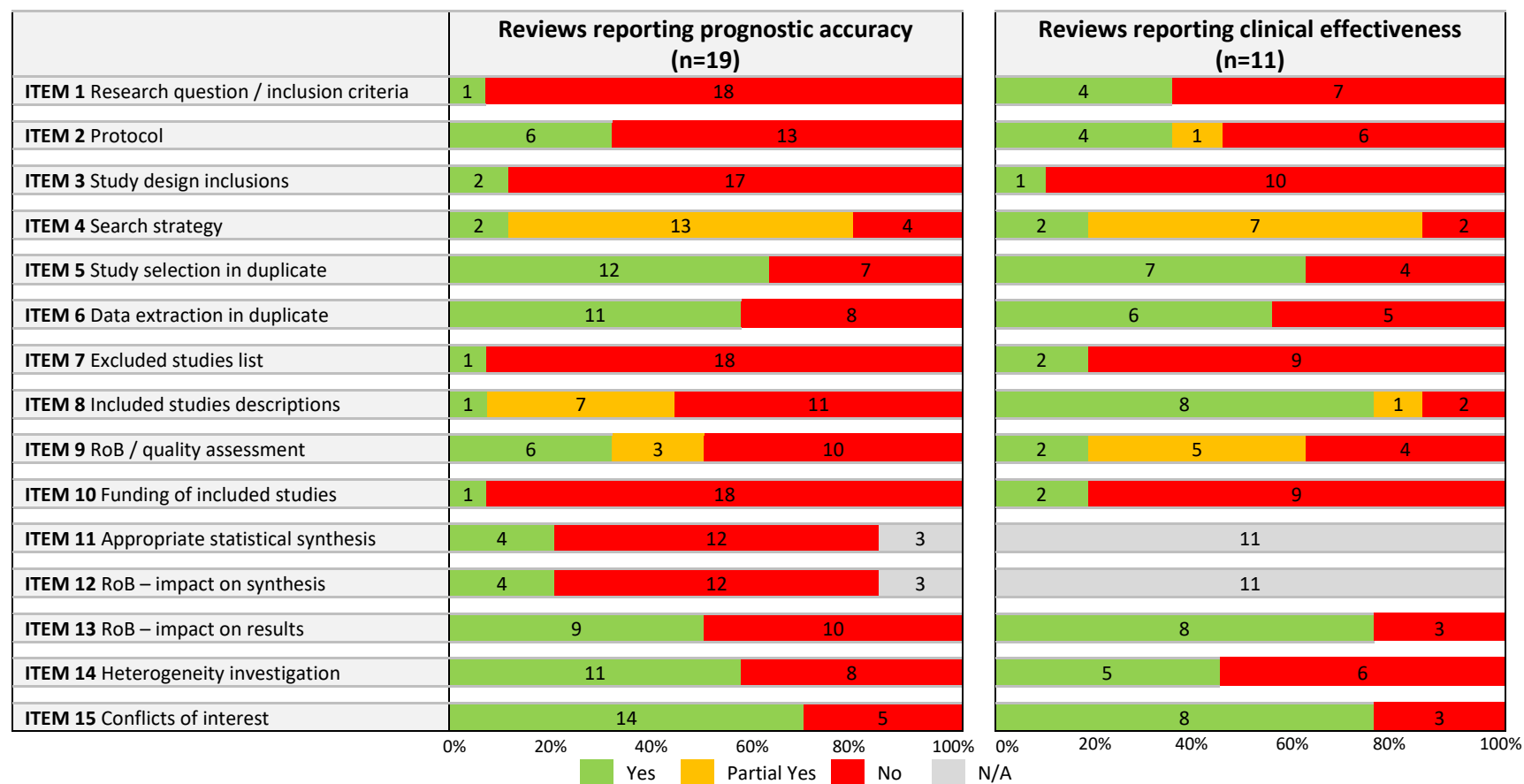
269 The quality of included reviews was generally poor (Table 2; Appendix 5 in S1 File). The AMSTAR-2
270 items that were most consistently met (yes or partial yes) were: comprehensiveness of the search
271 (21/26, 81%), study selection independently in duplicate (17/26, 65%), data extraction in
272 independently in duplicate (15/26, 58%), and conflicts of interest reported (20/26, 77%).

273 Six (32%) accuracy reviews^{36 40 41 47 48 53} and two (18%) effectiveness reviews used an appropriate
274 method of quality assessment of included studies (i.e. QUADAS or QUADAS-2 dependent on
275 publication year, or PROBAST for accuracy and the Cochrane tool for assessing risk of bias⁶⁴ and
276 criteria consistent with AHRQ Methods Guide for Effectiveness and Comparative Effectiveness
277 Reviews⁶⁵ for effectiveness reviews) and also presented judgements per study. Five reviews either
278 reported quality assessment results per study (n=4^{42 58-60}) or were considered to use an appropriate
279 quality assessment tool (n=1⁴³) (AMSTAR-2 criterion partially met).

280 Of the accuracy reviews that included a statistical synthesis, 25% (4/16)^{39 41 50 53} used an appropriate
281 meta-analytic method and investigated sources of heterogeneity. Two reviews^{41 53} used
282 recommended hierarchical approaches to meta-analysis of test accuracy data (the bivariate model⁴¹
283 and hierarchical summary ROC (HSROC) model⁵³) and two reviews calculated summary estimates of
284 individual measures, using random effects meta-analyses (AUC⁵⁰ or RR⁶⁶).

285 Compared to the reviews of accuracy, reviews of effectiveness more commonly provided adequate
286 descriptions of primary studies (8/11, 73% vs 1/19, 5%) and adequately defined their inclusion
287 criteria (4/11, 36% vs 1/19, 5%) (Table 2). No other major differences across review questions were
288 noted.

289 Table 2. Summary of AMSTAR-2 assessment results.



290

291 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
 292 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?;
 293 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
 294 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
 295 Applicable; RoB – Risk of Bias. Further details on AMSTAR items are given in Appendix 4 in S1 File, and results per review are given in Appendix 5 in S1 File.

296 Results from reviews evaluating the prognostic accuracy of risk prediction tools

297 Seven of 19 accuracy reviews were prioritised for narrative synthesis (Tables 3-4) and are reported
298 below according to risk prediction tool. Five of the seven reviews did not include development study
299 estimates within their meta-analyses, one review of ML models did not report this information⁴⁸ and
300 one⁴⁷ restricted inclusion to studies reporting model development studies. The latter review was the
301 only one to consider the effect of study quality in their statistical syntheses.

302 Braden, and modified Braden scales

303 The most recent and largest review⁴¹ of the Braden scale (60 studies, including 49,326 patients),
304 which used hierarchical bivariate meta-analysis, reported an overall summary sensitivity of 0.78 (95%
305 CI 0.74, 0.82; 15,241 patients) and specificity of 0.72 (95% CI 0.66, 0.78; 34,085 patients) across all
306 reported thresholds (range ≤ 10 to ≤ 20). Summary sensitivities and specificities ranged from 0.79
307 (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
308 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
309 18 (15 studies), respectively. Heterogeneity investigations suggested higher accuracy for predicting PI
310 risk in patients with a mean age of 60 years or less, in hospitalised patients (compared to long-term
311 care facility residents) and in Caucasian populations (compared to Asian populations).⁴¹ The review
312 noted a high risk of bias for the 'index test' section of the QUADAS-2 assessment in approximately a
313 third of included studies, but failed to provide further details.

314 Two modified versions of the Braden scale^{67 68} were included in another review.⁴⁴ Summary
315 sensitivities were 0.97 (95% CI 0.92, 0.99; 125 patients from four studies)⁶⁷ and 0.89 (95% CI 0.71,
316 0.98; 27 patients from two studies)⁶⁸, and summary specificities were 0.70 (95% CI 0.66, 0.73; 563
317 patients)⁶⁷ and 0.71 (95% CI 0.67, 0.75; 599 patients).⁶⁸ The review was rated critically low on the
318 AMSTAR-2 assessment, with only 1/15 (13%) criteria fulfilled. QUADAS-2 was reportedly used but
319 results not reported in any detail, other than to indicate that none of the included studies were
320 considered at high risk of bias.

321 Cubbin & Jackson scale

322 The most recent and comprehensive review³⁶ of the Cubbin & Jackson scale (9 studies, including
323 7,684 patients) reported summary sensitivity of 0.81 (95% CI 0.51, 0.95; 1,558 patients) and
324 specificity of 0.76 (95% CI 0.58, 0.88; 6,126 patients). However, this review scored critically low on
325 AMSTAR-2 (3/15, 20%, criteria fulfilled), with authors utilising inappropriate methods for statistical
326 synthesis, not investigating causes of heterogeneity and poor reporting of results throughout. Their
327 meta-analysis approach was also not clearly reported, but it appears that univariate meta-analyses
328 were conducted separately for sensitivity and specificity, across studies with different Cubbin &
329 Jackson thresholds.

330 Zhang and colleagues⁵³ included six studies evaluating the original Cubbin & Jackson scale⁶⁹ (800
331 patients). Summary sensitivity and specificity were both reported as 0.84 (95% CIs 0.59, 0.95 and
332 0.66, 0.93, respectively)⁵³ suggesting that this represents the point on the HSROC curve where
333 sensitivity equals specificity, particularly as reported thresholds ranged from 24 to 34. The review
334 authors concluded that although the accuracy of the Cubbin & Jackson scale was higher than the
335 EVARUCI scale and the Braden scale, low quality of evidence and significant heterogeneity limit the
336 strength of conclusions that can be drawn.

337 Norton scale

338 Park and colleagues⁴⁴ synthesised data from seven studies (2,899 participants) evaluating the Norton
339 scale, across thresholds ranging from <14 to <16 . They reported summary sensitivity of 0.75 (95% CI
340 0.70, 0.79) and specificity 0.57 (95% CI 0.55, 0.59). A further four reviews presented statistically

341 synthesised results for the Norton scale (Appendix 5 in S1 File), including one review by Chou and
342 colleagues³⁸ which included nine studies (5,444 participants) but only reported median values for
343 accuracy parameters.

344 Waterlow scale

345 Although Zhang and colleagues⁵³ included the fewest participants (4 studies; 1,000 participants) of all
346 six reviews that conducted a statistical synthesis of the accuracy of the Waterlow scale¹¹, they
347 provided the most recent review. It was rated highest on AMSTAR-2 criteria and appropriately used
348 the HSROC model for meta-analysis across thresholds ranging from 12 to 25. Summary sensitivity
349 was 0.63 (95% CI 0.48, 0.76) and summary specificity 0.46 (95% CI 0.22, 0.71) (Table 4). A second
350 review⁴⁴ reported contrasting results with summary sensitivity of 0.55 (95% CI 0.49, 0.62) and
351 specificity 0.82 (95% CI 0.80, 0.85) (6 studies; 1268 participants), however authors synthesised data
352 across multiple thresholds without utilising hierarchical methods.

353 Machine learning algorithms

354 Pei and colleagues⁴⁷ included 18 ML models, seven of which were not covered by any other included
355 review. Accuracy measures were combined across all models that provided 2x2 data (n=14 models).
356 The summary AUC across the 14 models was 0.94, summary sensitivity was 0.79 (95% CI 0.78, 0.80)
357 and summary specificity was 0.87 (95% CI 0.88, 0.87) (Table 4). Meta-regression found no significant
358 effect by ML algorithm or data type. Clinical heterogeneity was not investigated. The majority of
359 studies (89%, 16/18) were considered at high risk of bias based on PROBAST. Our confidence in the
360 review was critically low, with only 6/15 (40%) AMSTAR-2 criteria fulfilled. One critical flaw was the
361 use of inappropriate meta-analysis methods (failing to use a hierarchical model for synthesising
362 sensitivity and specificity).

363 Qu and colleagues⁴⁸ conducted separate meta-analyses of 25 studies by ML algorithm type using
364 Bayesian hierarchical methods (Table 3). The review rated critically low on AMSTAR-2 items, with
365 only 6/15 (40%) criteria fulfilled. The review did not restrict inclusion to external evaluations of the
366 models, and the authors did not report which estimates were sourced from development data or
367 external data. The summary AUC for the five algorithms ranged from 0.82 (95% CI 0.79, 0.85; 9
368 studies with 97,815 participants) for neural network-based models to 0.95 (95% CI 0.93, 0.97; 7
369 studies with 161,334 participants) for random forest models (Table 4).

370 The latter approach also had the highest summary specificity 0.96 (95% CI 0.80, 0.99), with sensitivity
371 0.72 (95% CI 0.26, 0.95). The highest summary sensitivity was observed for support vector machine
372 models (0.81, 95% CI 0.69, 0.90) with summary specificity 0.81 (95% CI 0.59, 0.93) (9 studies,
373 152,068 participants). The remaining algorithms had summary sensitivities ranging from 0.66
374 (decision tree models) to 0.73 (neural network models) (Table 4). Two additional ML algorithms
375 evaluated in the included studies (Bayesian networks and LOS (abbreviation not explained)) had too
376 few studies to allow meta-analysis (Appendix 5 in S1 File).

377 Other scales

378 In addition to the risk prediction tools reported above, Zhang and colleagues⁵³ reported on the
379 EVARUCI scale⁷⁰, presenting summary sensitivity and specificity of 0.84 (95% CI 0.79, 0.89) and 0.68
380 (95% CI 0.66, 0.70), respectively (3 studies; 3,063 participants). These results were synthesised across
381 thresholds, 11 and 11.5 (one not reported).

382 Additional statistical syntheses covering three further modifications of the Braden scale (Braden
383 modified by Kwong⁷¹, the 4-factor model⁷² and 'extended Braden'⁷²), two modified versions of the
384 Norton scale (by Ek⁷³, and by Bienstein⁷⁴), a revised "Jackson & Cubbin"⁷⁵, and the EMINA⁷⁶ and

385 PPS⁷⁷ tools) were also identified.^{39 38 49} These analyses showed variable performance, often with
386 high uncertainty. Full details can be found in Table A4 in S1 File.

387 Table A5 in S1 File reports data for another 17 risk prediction tools, each associated with a single
388 primary study (therefore not covered in detail in the text above), and another two tools,
389 Sunderland⁷⁸ and RAPS⁷⁹, which are assessed in two primary studies each.

Table 3. 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 ⁴¹ (2021)	Braden ^{8,9} 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: Studies performed worst overall in the patient selection domain, with low RoB in only 18% (11/60) and high concern for applicability in 27% (16/60). The index test domain also revealed some issues, with about one-third of studies having high RoB. Studies performed well in the remaining domains.	Bivariate meta-analysis; SROC constructed; and subgroup/stratified analyses to explore heterogeneity
Chen ³⁶ (2023)	Cubbin & Jackson ⁶⁹ n = 9; N = 7,684	Study designs included prospective studies (n=2), a prospective and cross-sectional study (n=1), retrospective studies (n=3), an observational study (n=1), a predictive correlational study (n=1) and a longitudinal study (n=1). Studies were conducted in South Korea (n=3), the US (n=2), Turkey (n=2), China (n=1) and Portugal (n=1). Studies targeted research participants >18 years old (n=6), >16 years old (n=1), ≥21 years old (n=1) or did not set an age limit (n=1). All studies restricted to patients without PIs on ICU admission.	Authors state that " <i>none of [the 9 included studies] has a high risk of bias in any field</i> ", despite indicating some as 'high risk' in the table of QUADAS-2 results. From table of QUADAS-2 results: 5/9 (56%) studies are at high RoB, 3/9 (33%) are at unclear RoB and 1/9 (11%) is at low RoB.	Method unclear, but synthesised across different thresholds of Cubbin & Jackson; SROC constructed
Zhang ⁵³ (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 ^{8,9} n = 18; N = 11,167	Cut-offs used range from 10.5 to 20. Development study not included.		
	Cubbin & Jackson ⁶⁹ n = 6; N = 800	Cut-offs used range from 24 to 34. Development study not included.		
	EVARUCI ⁷⁰ 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 ¹¹ n = 4; N = 1,000	Cut-offs: 12, 16, <25, 20.5 Development study not included.		

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
Park ⁴⁵ (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 ^{8,9} n = 25; N = 10,547	Cut-offs: 13 (n=2); 16 (n=8); 17 (n=2); 18 (n=9); 19 (n=3); 20 (n=1) Development study not included.		
	Norton ¹⁰ n = 5; N = 2,408	Cut-offs: 14 (n=2); 16 (n=3) Development study not included.		
	Waterlow ¹¹ n = 5; N = 1,406	Cut-offs: 15 (n=1); 16 (n=2); 17 (n=1); NS (n=1) Development study not included.		
Park ⁴⁴ (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) ³⁴ and Walter 2002 ⁸⁰ cited
	Braden – modified by Song & Choi ⁶⁷ 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 ⁶⁸ 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 ⁶⁹ 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.		

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
	Norton ¹⁰ 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 ¹¹ 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.		
Pei ⁴⁷ (2023)	Various ML models n = 14; N = 408,504	Studies published from 2012-2022; studies conducted in China (n=4), Taiwan (n=3), USA (n=6), Germany (n=1), Japan (n=1), Australia (n=1), Spain (n=1) and Czech Republic (n=1); 15 studies utilised retrospective data, while 3 used prospective data; studies focused on adult ICU inpatients (n=5), hospitalised patients (n=8), adult hospitalised patients awaiting surgery (n=3), cancer patients (n=1) and end-of-life adult inpatients (n=1); sample size range 168-149,006; number of patients with PI range 8-4663 Validation methods: both sample splitting and k-fold cross-validation (n= 10), sample splitting only (n=4), k-fold cross-validation only (n=1). External verification conducted (n=1). Validation methods not reported (n=2). Handling of missing data not disclosed (n=7).	PROBAST: Overall, 16/18 (88.9%) papers were at high RoB, 1 (5.6%) was at unclear RoB and only 1 (5.6%) was at low RoB. 14 (77.8%) studies were at high RoB in the analysis domain. The most common factors contributing to the high risk of bias in the analysis domain included an inadequate number of events per candidate predictor, poor handling of missing data and failure to deal with overfitting.	Meta-analysis conducted for the 14 studies that presented 2x2 information; DTA meta-analysis (random-effects) using MetaDiSc (no further details); SROC constructed; meta-regression heterogeneity investigation (only p-values reported)
Qu ⁴⁸ (2022)	Models by ML algorithm type Decision Tree n = 14; N = 118,292	Characteristics only reported overall, not by algorithm type.	2/14 high RoB; 10/14 low RoB; 2/14 unclear RoB	RevMan (Moses & Littenberg method ^{81,82}) for analysis of quantitative data (SROC plot)

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
	Logistic Regression n = 14; N = 195,927	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.	4/14 high RoB; 9/14 low RoB; 1/14 unclear RoB	presented) and Bayesian DTA-NMA
	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.	

391 * 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%.

392 AHCPR – Agency for Health Care Policy and Research; CI – confidence interval; CVD – cardiovascular disease; DTA – diagnostic test accuracy; EPUAP – European Pressure Ulcer Advisory Panel;

393 (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 –

394 number of studies; NMA – network meta-analysis; NS – not stated; NPUAP – National Pressure Ulcer Advisory Panel; PI – pressure injury; PPU – Panel for the Prediction and Prevention of

395 Pressure Ulcers; QUADAS – Quality Assessment of Diagnostic Accuracy Studies; RoB – risk of bias; TDCPS – Torrance Developmental Classification of Pressure Sore.

Table 4. 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^{8,9} (1987)						
Huang ⁴¹ (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 ⁵³ (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 ⁴⁵ (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 & Choi⁶⁷ (1991)						
Park ⁴⁴ (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 & Wong⁶⁸ (1998)						
Park ⁴⁴ (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 & Jackson⁶⁹ (1991)						
Chen ³⁶ (2023)	n = 9; N = 7,684	0.81 (0.51, 0.95) N=1,558	0.76 (0.58, 0.88) N=6,126	PLR 3.34 (2.14, 5.21) NLR 0.25 (0.09, 0.68)	13.24 (5.41, 32.40)	0.84 (0.81, 0.87)
Zhang ⁵³ (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 ⁴⁴ (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: EVARUCI⁷⁰ (2001)						
Zhang ⁵³ (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: Norton¹⁰ (1962)						
Park ⁴⁴ (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 ⁴⁵ (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: Waterlow¹¹ (1985)						
Zhang ⁵³ (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 ⁴⁴ (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 ⁴⁵ (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						
Pei ⁴⁷ (2023) Various ML models	n = 14; N = 328,789	0.79 (0.78, 0.80) N=18,807	0.87 (0.88, 0.87) N=309,982	PLR 10.71 (5.98, 19.19) NLR 0.21 (0.08, 0.50)	52.39 (24.83, 110.55)	0.94
Qu ⁴⁸ (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 ⁴⁸ (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 ⁴⁸ (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 ⁴⁸ (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 ⁴⁸ (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)

397 ^A summary statistic 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 from
398 the Qu review⁴⁸ include data from development, internal validation or external validation/evaluation studies.
399 AUROC – area under the receiver operating characteristic curve; CI – confidence interval; DT – decision tree; DOR – diagnostic odds ratio; HSROC – hierarchical summary receiver operating
400 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;
401 SE – standard error; SVM – support vector machine.

402 Results from reviews evaluating the clinical effectiveness of risk prediction tools

403 The 11 reviews reporting clinical effectiveness, used a range of eligibility criteria and a number of
404 different quality assessment tools, leading to varying conclusions about the methodological quality
405 of the same studies across reviews. Given the overlap in study inclusion between reviews Table 5
406 provides an overview of results from four^{38 57 59 61} of the 11 reviews, and a summary of the included
407 comparative studies is provided below.

408 Two randomised controlled trials (RCTs) of risk prediction tools^{83 84} were identified, both of which
409 were considered at high risk of bias in the Cochrane review (assessed using the Cochrane RoB tool⁶⁴).
410 One of the trials (an individually randomised study⁸³) was included in a further three reviews which
411 considered it to be 'good quality'³⁸, 'valid'⁵⁶, or 'high quality'⁵⁹. The trial was conducted in 1,231
412 hospital inpatients and the only intervention was that the staff must use the tool that was allocated
413 to them, with no other protocol prescribed changes made to routine care. However, no evidence of a
414 difference in PI incidence was found between patients assessed with either the Waterlow scale or
415 Ramstadius tool compared with clinical judgment alone (RR 1.10 (95% CI 0.68, 1.81) and RR 0.79
416 (95% CI 0.46, 1.35), respectively). The trial further showed no evidence of a difference in patient
417 management or in PI severity when using a risk assessment tool compared to clinical judgement.

418 A further cluster randomised trial⁸⁴ was considered to be of poor methodological quality both in the
419 Cochrane review³⁸ and one other review⁶¹. The trial included 521 patients at a military hospital and
420 compared nurse training with mandatory use of the Braden scale, to nurse training and optional use
421 of the Braden scale, to no training. No evidence of a difference in PI incidence was observed between
422 the three groups: incidence rates were 22%, 22% and 15% (p=0.38), respectively.

423 Two reviews by Lovegrove and colleagues^{59 60} included an uncontrolled comparison study⁸⁵ rated as
424 high quality⁵⁹. The study compared the clinical effectiveness of the Maelor scale⁸⁶ used in an Irish
425 hospital (121 patients) with nurses' clinical judgement at a Norwegian hospital (59 patients). A higher
426 rate of preventive strategies, as well as a lower PI prevalence (12% vs. 54%), was reported for the
427 Irish hospital. However, these results are likely to be highly confounded by inherent differences
428 in population and setting.

429 A non-randomised study by Gunningberg and colleagues⁸⁷ included in two reviews^{43 57} was
430 considered by review authors to be of relatively high quality. The study was conducted in 124
431 patients in emergency and orthopaedic units and compared the use of a PI risk alarm sticker for
432 patients with a modified Norton Score of <21 (indicating high-risk patients) to standard care. No
433 significant difference in the incidence of PIs between the Norton scale and standard care groups was
434 observed.

435 A non-randomised study⁸⁸ conducted in 233 hospice inpatients was included in three reviews,^{38 43 57}
436 one of which is reported in Table 5.⁵⁷ The study met six of eight quality criteria used by Health
437 Quality Ontario.⁵⁷ Use of a modified version of the Norton scale (Norton modified by Bale), in
438 conjunction with standardised use of preventive interventions based on risk score, was found to be
439 associated with lower risk of PIs when compared with nurses' clinical judgment alone (RR 0.11, 95%
440 CI 0.03, 0.46). The lack of randomisation limits the reliability of this result, and review authors report
441 that the modified Norton scale had not been validated.

442 Finally, a 'before-and-after' study⁸⁹ of 181 patients in various hospital settings was included in two
443 reviews,^{43 57} one of which considered the study to meet all quality criteria.⁵⁷ Use of the Norton scale
444 with additional training for staff was associated with significant differences in the number of
445 preventive interventions prescribed compared to standard care (18.96 vs. 10.75, respectively).

446 Preventive interventions were also introduced earlier in the intervention group (on day 1, 61% vs.
447 50%, $p < 0.002$ for Norton and usual care, respectively). However, no significant difference in the
448 incidence of PIs was detected between the groups.

Table 5. Systematic reviews evaluating clinical effectiveness

Review author (publication year)	Tools included	Setting of included studies; study design; sample size	Included outcomes	Brief description of study quality	Relevant results from included studies
Lovegrove ⁵⁹ (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 preventive interventions; interrater reliability (reliability results covered in Appendix 5 in S1 File)	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.	<ul style="list-style-type: none"> • 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 2011⁸³). • 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 2015⁸⁵).
Moore ⁶¹ (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 ⁶⁴). Both studies at high RoB due to blinding issues. One study at RoB also due to baseline imbalance and incorrect analyses.	<ul style="list-style-type: none"> • 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 2009⁸⁴). • 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 2011⁸³). • No difference in PI severity based on risk assessment tools vs. clinical judgement (Webster 2011⁸³).
Chou ³⁸ (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 preventive 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.	<ul style="list-style-type: none"> • 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 2011⁸³). • 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 1995⁸⁸). • 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 2009⁸⁴).

Review author (publication year)	Tools included	Setting of included studies; study design; sample size	Included outcomes	Brief description of study quality	Relevant results from included studies
Health Quality Ontario ⁵⁷ (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 preventive 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.	<ul style="list-style-type: none"> • 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 1999⁸⁷). • 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<0.001) (Bale 1995⁸⁸). • Compared the Norton scale with training to standard care. There was a significant difference in the number of preventive 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 1990⁸⁹).

450

451

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.

452 DISCUSSION

453 This umbrella review summarises data from a total of 26 systematic reviews of studies evaluating the
454 prognostic accuracy and clinical effectiveness of a total of 70 PI risk prediction tools. Despite the
455 large number of available reviews, quality assessment using an adaptation of AMSTAR-2 suggested
456 that the majority were conducted to a relatively poor standard or did not meet reporting standards
457 for systematic reviews.^{19 90} Of the 15 AMSTAR-2 items assessed, only two (for accuracy reviews) and
458 four (for effectiveness reviews) criteria were more consistently met (more than 60% of reviews
459 scoring 'Yes'). Whilst AMSTAR-2 Item 6 (data extraction independent in duplicate) was fulfilled by
460 over half of all reviews (15/26, 58%), and Item 14 (adequate heterogeneity investigation) was fulfilled
461 by around half of the accuracy reviews (10/19, 53%), all other criteria were fully met by less than half
462 of the reviews. The primary studies included in the reviews were particularly poorly described in the
463 accuracy reviews, making it difficult to determine exactly what was evaluated and in whom. The
464 extent to which we could reliably describe and comment on the content of the reviews is limited and
465 high-quality evidence for the accuracy and clinical effectiveness of PI risk prediction tools may be
466 lacking.

467 Prognostic accuracy of risk prediction tools

468 Of the 19 reviews reporting the accuracy of included tools, only two used appropriate methods for
469 both quality assessment and statistical synthesis of accuracy data^{41 53}, one of which⁴¹ evaluated only
470 the Braden scale. Only two reviews^{42 43} pre-specified the exclusion of studies reporting accuracy data
471 from tool development studies, one review restricted to "validated risk assessment instruments"
472 only³⁸ and one review⁴⁷ was limited to development studies only. This was the only review⁴⁷, that
473 discussed the importance of appropriate validation of prediction tools. Only two reviews conducted
474 meta-analyses at different cut-offs for determination of high risk^{38 41}; the remaining reviews
475 combined data regardless of the threshold used. Combining data across different thresholds to
476 estimate summary sensitivity and specificity yields clinically uninterpretable and non-generalisable
477 estimates that do not relate to a particular threshold.³⁵ Only one review³⁸ considered timing in their
478 inclusion criteria or in the description of primary studies. It is important to interpret the findings
479 below with these limitations in mind.

480 The included meta-analyses consistently suggested that risk prediction scales have moderate
481 sensitivities and somewhat lower specificities, typically in the range of around 70% to 85% for
482 sensitivity and as low as 30% to 40% for specificity for some tools. Although these ranges in
483 sensitivities and specificities would be considered on the lower end of acceptable within a diagnostic
484 accuracy paradigm, they may have greater utility in a prognostic context. Without a detailed review
485 of the primary study publications for these tools, it is not possible to assess which, if any, of these
486 risk assessment scales might outperform the others. It seems that limited comparative studies
487 comparing the accuracy of different tools are available.

488 For the ML-based models, one review⁴⁷ combined multiple ML models into one meta-analysis and
489 another⁴⁸ meta-analysed accuracy data by algorithm type. The results of the latter meta-analyses are
490 not informative for clinical practice but may be a useful way of identifying which ML algorithms may
491 be more suited to the data. Results suggested that specificities for random forest or decision tree
492 models could reach 90% or above with associated sensitivities in the range of 66% to 72%, however
493 relatively wide confidence intervals around these summary estimates reflect considerable variation
494 in model performance. Moreover, some of these estimates came from internal validations within
495 model development studies, and may not be transferable to other settings.⁹¹ Authors should make it
496 clear where accuracy estimates are derived from to avoid overinterpretation of results.

497 Diagnostic accuracy studies are typically cross-sectional in the sense that there should be no, or only
498 minimal delay between application of the test and the reference standard.^{92 93} For prognostic
499 accuracy however, there is a time delay between the application of the test and the outcome that
500 the tool aims to predict. If the use of an accurate PI risk prediction tool is combined with effective
501 and appropriate preventive measures in those identified as most at risk, the incidence of PI would
502 decline, reducing the positive predictive value of the original risk assessment and potentially the
503 sensitivity of the tool.⁹⁴ Sensitivity and specificity can be optimised by methods which directly
504 consider the cost of misclassification, including both the harms associated with applying more
505 intensive prevention in those with a false positive result and the benefits of preventive measures in
506 those with a true positive result. One solution to determine the preventive treatment threshold risk
507 is through net benefit calculations,^{95 96} which can be visualised in decision curves and are common in
508 prognostic research. These calculations can assist in providing a balanced use of resources while
509 maximising positive health outcomes, such as lowering incidence of PI.

510 It is important to also consider that not all predictors have a causal relationship with the outcome,
511 therefore, not every predictor will be a clinical risk modifier. Risk assessment tools that allow a more
512 personalised-risk approach, i.e. that identify and flag predictors that are risk modifiers to the end-
513 users of the tool, would make predictions more interpretable and actionable. Some such
514 developments exist,^{97 98} but future validation of these methods is needed. Where risk assessment
515 tools are developed for enriching study design (for example, as a means of recruiting only high-risk
516 patients to studies evaluating preventive measures), a different approach and optimisation of
517 performance metrics would be needed. Risk prediction models should therefore pre-specify their
518 intended application before development to allow their clinical utility for a given context to be
519 addressed.⁹⁹

520 Clinical effectiveness of risk prediction scales

521 Prediction models, like any test used for diagnostic or prognostic purposes, require evaluation in the
522 care pathway to identify the extent to which their use can impact on health outcomes.¹⁰⁰ Of the 11
523 reviews assessing clinical effectiveness of PI risk prediction tools, the only primary studies suggesting
524 potential patient benefits from the use of risk prediction tools^{85 88 89} were non-randomised and are
525 likely to be at high risk of bias. In contrast, two randomised trials^{83 84} (both considered at high risk of
526 bias by the Cochrane review⁶¹) suggest that use of structured risk assessment tools does not
527 ultimately lead to the reduction in incidence of PIs. We should recognise that effectiveness outcomes
528 from using a risk prediction tool depend on the timely implementation of effective preventive
529 measures, a step that is frequently poorly described in studies evaluating the effectiveness of risk
530 assessment tools, restricting the conclusions that can be drawn from the limited evidence available.
531 One possible explanation for the lack of differences in PI incidence is the implementation of
532 preventive measures that have not been proven effective in preventing PIs, such as alternating air-
533 mattresses.⁴ All reviews included studies that assessed the use of risk assessment scales developed
534 by clinical experts, and no evidence is available evaluating the clinical effectiveness of empirically
535 derived prediction models or ML algorithms.

536 Other existing evidence

537 We have separately reviewed⁷ available evidence for the development and validation of risk
538 prediction tools for PI occurrence. Almost half (60/124, 48%) of available tools were developed using
539 ML methods (as defined by review authors), 37% (46/124) were based on clinical expertise or
540 unclear methods, and only 18 (15%) were identified as having used statistical modelling methods.
541 The reviews varied in methodological quality and reporting; however, the reporting of prediction
542 model development in the original primary studies appears to be poor. For example, across all

543 prediction tools identified, the internal validation approach was unclear and unidentifiable for 72%
544 (89/124) of tools, and only one review¹⁰¹ identified and included external validation studies (n=7
545 studies).

546 ML-based models may have potential for identifying those at risk of PI, as suggested by two reviews⁴⁷
547 ⁴⁸ included in this umbrella review. However, it is important to consider the lack of transparency in
548 reporting of model development methods and model performance, and the concerning lack of
549 model validation in populations outside of the original model development sample.⁷

550 Strengths and limitations

551 We have conducted the first umbrella review that summarises the prognostic accuracy and clinical
552 effectiveness of prediction tools for risk of PI. We followed Cochrane guidance¹⁸, with a highly
553 sensitive search strategy designed by an experienced information specialist. Although we excluded
554 non-English publications due to time and resource constraints, where possible these publications
555 were used to identify additional eligible risk prediction tools.

556 To some extent, our review is limited by the use of AMSTAR-2 for quality assessment of included
557 reviews. AMSTAR-2 was not designed for assessing systematic reviews of diagnostic or prognostic
558 studies. Although we made some adaptations, many of the existing and amended criteria relate to
559 the quality of reporting of the reviews as opposed to methodological quality. There is scope for
560 further work to establish criteria for assessing systematic reviews of prediction tools. Additionally, we
561 chose not to exclude reviews based on low AMSTAR-2 ratings to provide a comprehensive overview
562 of all available evidence. However, by doing so, we acknowledge that many included reviews are of
563 poor quality (with critically low confidence in 81%, 21/26, reviews), reducing the reliability of the
564 evidence presented and the ability to make conclusions or recommendations based on this evidence.

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

570 Although standards for reporting of test accuracy studies have been available since the year 2000,⁹²
571 standards for reporting risk prediction models were not published until 2015.¹⁰² Similarly, quality
572 assessment tools highlighting important areas for consideration in primary studies have been
573 available for DTA studies since 2003, with an adaption to prognostic accuracy published in 2022,¹⁰³
574 and PROBAST for prediction model studies in 2019.³³ This lag in methodological developments for
575 studies and systematic reviews of risk prediction tools has likely contributed to the observed
576 emphasis on the application of DTA principles in our set of reviews, without sufficient consideration
577 of the prognostic context and effect on accuracy of intervening and effective preventive
578 interventions.

579 While 18/19 (95%) accuracy reviews aimed to evaluate the 'predictive' validity of PI risk assessment
580 tools, the majority (16/19, 84%) relied on DTA principles without any consideration of the time
581 interval between the test and the outcome, i.e. occurrence of PI. This approach does not account for
582 the prognostic nature of these tools or address longitudinal questions, such as censoring and
583 competing events.¹⁰³ Another fundamental flaw in these accuracy assessments is that risk scales may
584 actually appear to perform worse in settings where risk prediction and preventive care are most
585 effective, as accurate risk prediction combined with effective preventive measures may prevent
586 patients classified as 'high-risk' from developing PIs.⁹⁴

587 CONCLUSIONS

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

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

602

603 Supporting Information

604 **S1 File. Appendices.**

605 Acknowledgements

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

608 Author Contributions

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

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

612 **Formal analysis:** Bethany Hillier, Katie Scandrett, Jacqueline Dinnes

613 **Funding acquisition:** Yemisi Takwoingi, Vladica Velickovic, Jacqueline Dinnes

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

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

617 **Project administration:** Bethany Hillier, Yemisi Takwoingi, Jacqueline Dinnes

618 **Resources:** Bethany Hillier, Katie Scandrett

619 **Supervision:** Yemisi Takwoingi, Jacqueline Dinnes

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

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

623 Funding

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

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

632 Conflicting Interests

633 I have read the journal's policy, and the authors of this manuscript have the following competing
634 interests: VV is an employee of Paul Hartmann AG; ES and THB received consultancy fees from Paul
635 Hartmann AG. VV, ES and THB were not involved in data curation, screening, data extraction, analysis
636 of results or writing of the original draft. These roles were conducted independently by authors at
637 the University of Birmingham. All other authors received no personal funding or personal
638 compensation from Paul Hartmann AG and have declared that no competing interests exist.

639 References

- 640 1. Li Z, Lin F, Thalib L, et al. Global prevalence and incidence of pressure injuries in hospitalised adult
641 patients: A systematic review and meta-analysis. *International Journal of Nursing Studies*
642 2020;105:103-546. doi: 10.1016/j.ijnurstu.2020.103546
- 643 2. Padula WV, Delarmente BA. The national cost of hospital-acquired pressure injuries in the United
644 States. *Int Wound J* 2019;16(3):634-40. doi: 10.1111/iwj.13071 [published Online First:
645 2019/01/28]
- 646 3. Sullivan N, Schoelles K. Preventing In-Facility Pressure Ulcers as a Patient Safety Strategy. *Annals of*
647 *Internal Medicine* 2013;158(5.2):410-16. doi: 10.7326/0003-4819-158-5-201303051-00008
- 648 4. Qaseem A, Mir TP, Starkey M, et al. Risk Assessment and Prevention of Pressure Ulcers: A Clinical
649 Practice Guideline From the American College of Physicians. *Annals of Internal Medicine*
650 2015;162(5):359-69. doi: 10.7326/m14-1567
- 651 5. Padula WV, Pronovost PJ, Makic MBF, et al. Value of hospital resources for effective pressure injury
652 prevention: a cost-effectiveness analysis. *BMJ Quality & Safety* 2019;28(2):132.
653 doi: 10.1136/bmjqs-2017-007505
- 654 6. Institute for Quality and Efficiency in Health Care (IQWiG). Preventing pressure ulcers. Cologne,
655 Germany 2006 [updated 2018 Nov 15. Available from:
656 <https://www.ncbi.nlm.nih.gov/books/NBK326430/?report=classic> accessed Feb 2023].
- 657 7. Hillier B, Scandrett K, Coombe A, et al. Development and validation of risk prediction tools for
658 pressure injury occurrence: An umbrella review (pre-print). *MedRxiv* 2024 doi:
659 10.1101/2024.05.07.24306999
- 660 8. Braden B, Bergstrom N. A Conceptual Schema for the Study of the Etiology of Pressure Sores.
661 *Rehabilitation Nursing* 1987;12(1):8-16. doi: 10.1002/j.2048-7940.1987.tb00541.x
- 662 9. Bergstrom N, Braden BJ, Laguzza A, et al. The Braden Scale for Predicting Pressure Sore Risk. *Nurs*
663 *Res* 1987;36(4):205-10.
- 664 10. Norton D. Geriatric nursing problems. *Int Nurs Rev* 1962;9:39-41.
- 665 11. Waterlow J. Pressure sores: a risk assessment card. *Nursing Times* 1985;81:49-55.
- 666 12. NICE. Pressure ulcers: prevention and management. Clinical guideline [CG179]. 2014 [Available
667 from: <https://www.nice.org.uk/guidance/cg179> accessed Aug 2024].
- 668 13. Haesler E. European Pressure Ulcer Advisory Panel, National Pressure Injury Advisory Panel and
669 Pan Pacific Pressure Injury Alliance. Prevention and Treatment of Pressure Ulcers/Injuries:
670 Clinical Practice Guideline. 2019 [Available from: <https://internationalguideline.com/2019>
671 accessed Feb 2023].
- 672 14. Scott K, Longstaffe S. Judy Waterlow. 2020 [Available from: <https://litfl.com/judy-waterlow/>
673 accessed Aug 2024].
- 674 15. Šimundić AM. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC* 2009;19(4):203-11.
675 [published Online First: 2009/01/20]
- 676 16. Leeftang MM, Rutjes AW, Reitsma JB, et al. Variation of a test's sensitivity and specificity with
677 disease prevalence. *CMAJ* 2013;185(11):E537-44. doi: 10.1503/cmaj.121286 [published
678 Online First: 2013/06/24]
- 679 17. Maiga A, Farjah F, Blume J, et al. Risk Prediction in Clinical Practice: A Practical Guide for
680 Cardiothoracic Surgeons. *Ann Thorac Surg* 2019;108(5):1573-82. doi:
681 10.1016/j.athoracsur.2019.04.126 [published Online First: 2019/06/27]
- 682 18. Pollock M, Fernandes RM BL, Pieper D, Hartling L. Chapter V: Overviews of Reviews. In: Higgins
683 JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA ed. *Cochrane Handbook for*
684 *Systematic Reviews of Interventions* version 63 (updated February 2022). Available from
685 www.training.cochrane.org/handbook: Cochrane 2022.
- 686 19. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and
687 Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA*
688 2018;319(4):388-96. doi: 10.1001/jama.2017.19163

- 689 20. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc*
690 2001;8(4):391-7. doi: 10.1136/jamia.2001.0080391 [published Online First: 2001/06/22]
- 691 21. Wilczynski NL, Haynes RB. Optimal Search Strategies for Detecting Clinically Sound Prognostic
692 Studies in EMBASE: An Analytic Survey. *Journal of the American Medical Informatics*
693 *Association* 2005;12(4):481-85. doi: 10.1197/jamia.M1752
- 694 22. Geersing G-J, Bouwmeester W, Zuithoff P, et al. Search Filters for Finding Prognostic and
695 Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. *PLOS ONE*
696 2012;7(2):e32844. doi: 10.1371/journal.pone.0032844
- 697 23. NHS. Pressure ulcers: revised definition and measurement. Summary and recommendations 2018
698 [Available from: [https://www.england.nhs.uk/wp-content/uploads/2021/09/NSTPP-](https://www.england.nhs.uk/wp-content/uploads/2021/09/NSTPP-summary-recommendations.pdf)
699 [summary-recommendations.pdf](https://www.england.nhs.uk/wp-content/uploads/2021/09/NSTPP-summary-recommendations.pdf) accessed Feb 2023].
- 700 24. AHCPR. Pressure ulcer treatment. : Agency for Health Care Policy and Research 1994:1-25.
- 701 25. Harker J. Pressure ulcer classification: the Torrance system. *Journal of Wound Care* 2000;9(6):275-
702 77. doi: 10.12968/jowc.2000.9.6.26233
- 703 26. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical Appraisal and Data Extraction for
704 Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLOS Medicine*
705 2014;11(10):e1001744. doi: 10.1371/journal.pmed.1001744
- 706 27. Cochrane. DE form example prognostic models - scoping review: The Cochrane Collaboration: The
707 Prognosis Methods Group; [Available from: <https://methods.cochrane.org/prognosis/tools>
708 accessed Feb 2023].
- 709 28. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that
710 include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*
711 2017;358:j4008. doi: 10.1136/bmj.j4008
- 712 29. World Health O. WHO handbook for guideline development: Chapter 17: developing guideline
713 recommendations for tests or diagnostic tools. 2nd ed. Geneva: World Health Organization
714 2014:218.
- 715 30. Whiting P, Savović J, Higgins JP, et al. ROBIS: A new tool to assess risk of bias in systematic reviews
716 was developed. *J Clin Epidemiol* 2016;69:225-34. doi: 10.1016/j.jclinepi.2015.06.005
717 [published Online First: 20150616]
- 718 31. Whiting P, Rutjes AWS, Reitsma JB, et al. The development of QUADAS: a tool for the quality
719 assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res*
720 *Methodol* 2003;3(25) doi: 10.1186/1471-2288-3-25
- 721 32. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment
722 of diagnostic accuracy studies. *Ann Intern Med* 2011;155(8):529-36. doi: 10.7326/0003-4819-
723 155-8-201110180-00009
- 724 33. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability
725 of Prediction Model Studies. *Annals of Internal Medicine* 2019;170(1):51-58. doi:
726 10.7326/M18-1376
- 727 34. Macaskill P, Gatsonis C, Deeks J, et al. Cochrane handbook for systematic reviews of diagnostic
728 test accuracy: Version, 2010.
- 729 35. Macaskill P, Takwoingi Y, Deeks J, et al. Chapter 9: Understanding meta-analysis. In: Deeks J,
730 Bossuyt P, Leeflang M, et al., eds. *Cochrane Handbook for Systematic Reviews of Diagnostic*
731 *Test Accuracy*. Version 2.0 ed: Cochrane, 2023 (updated July 2023).
- 732 36. Chen X, Diao D, Ye L. Predictive validity of the Jackson–Cubbin scale for pressure ulcers in
733 intensive care unit patients: A meta-analysis. *Nursing in Critical Care* 2023;28(3):370-78. doi:
734 10.1111/nicc.12818
- 735 37. Chen HL, Shen WQ, Liu P. A Meta-analysis to Evaluate the Predictive Validity of the Braden Scale
736 for Pressure Ulcer Risk Assessment in Long-term Care. *Ostomy/wound management*
737 2016;62(9):20-8.
- 738 38. Chou R, Dana T, Bougatsos C, et al. Pressure ulcer risk assessment and prevention: a systematic
739 comparative effectiveness review. *Annals of internal medicine* 2013;159(1):28-38.

- 740 39. García-Fernández FP, Pancorbo-Hidalgo PL, Agreda JJS. Predictive Capacity of Risk Assessment
741 Scales and Clinical Judgment for Pressure Ulcers: A Meta-analysis. *Journal of Wound Ostomy*
742 *& Continence Nursing* 2014;41(1):24-34. doi: 10.1097/01.WON.0000438014.90734.a2
- 743 40. He W, Liu P, Chen HL. The Braden Scale cannot be used alone for assessing pressure ulcer risk in
744 surgical patients: a meta-analysis. *Ostomy/wound management* 2012;58:34-40.
- 745 41. Huang C, Ma Y, Wang C, et al. Predictive validity of the braden scale for pressure injury risk
746 assessment in adults: A systematic review and meta-analysis. *Nursing open* 2021;8:2194-207.
747 doi: 10.1002/nop2.792
- 748 42. Mehicic A, Burston A, Fulbrook P. Psychometric properties of the Braden scale to assess pressure
749 injury risk in intensive care: A systematic review. *Intensive & critical care nursing*
750 2024;83:103686. doi: 10.1016/j.iccn.2024.103686
- 751 43. Pancorbo-Hidalgo PL, Garcia-Fernandez FP, Lopez-Medina IM, et al. Risk assessment scales for
752 pressure ulcer prevention: a systematic review. *J Adv Nurs* 2006;54(1):94-110. doi:
753 10.1111/j.1365-2648.2006.03794.x
- 754 44. Park SH, Lee HS. Assessing Predictive Validity of Pressure Ulcer Risk Scales- A Systematic Review
755 and Meta-Analysis. *Iranian journal of public health* 2016;45(2):122-33.
- 756 45. Park SH, Lee YS, Kwon YM. Predictive Validity of Pressure Ulcer Risk Assessment Tools for Elderly:
757 A Meta-Analysis. *Western journal of nursing research* 2016;38:459-83. doi:
758 10.1177/0193945915602259
- 759 46. Park SH, Choi YK, Kang CB. Predictive validity of the Braden Scale for pressure ulcer risk in
760 hospitalized patients. *Journal of Tissue Viability* 2015;24:102-13. doi:
761 10.1016/j.jtv.2015.05.001
- 762 47. Pei J, Guo X, Tao H, et al. Machine learning-based prediction models for pressure injury: A
763 systematic review and meta-analysis. *Int Wound J* 2023 doi: 10.1111/iwj.14280 [published
764 Online First: 2023/06/20]
- 765 48. Qu C, Luo W, Zeng Z, et al. The predictive effect of different machine learning algorithms for
766 pressure injuries in hospitalized patients: A network meta-analyses. *Heliyon*
767 2022;8(11):e11361. doi: 10.1016/j.heliyon.2022.e11361
- 768 49. Tayyib NAH, Coyer F, Lewis P. Pressure ulcers in the adult intensive care unit: a literature review of
769 patient risk factors and risk assessment scales. *Journal of Nursing Education and Practice*
770 2013;3(11):28-42.
- 771 50. Wang N, Lv L, Yan F, et al. Biomarkers for the early detection of pressure injury: A systematic
772 review and meta-analysis. *Journal of Tissue Viability* 2022;31:259-67. doi:
773 10.1016/j.jtv.2022.02.005
- 774 51. Wei M, Wu L, Chen Y, et al. Predictive Validity of the Braden Scale for Pressure Ulcer Risk in
775 Critical Care: A Meta-Analysis. *Nursing in critical care* 2020;25:165-70. doi:
776 10.1111/nicc.12500
- 777 52. Wilchesky M, Lungu O. Predictive and concurrent validity of the Braden scale in long-term care: A
778 meta-analysis. *Wound Repair and Regeneration* 2015;23:44-56. doi: 10.1111/wrr.12261
- 779 53. Zhang Y, Zhuang Y, Shen J, et al. Value of pressure injury assessment scales for patients in the
780 intensive care unit: Systematic review and diagnostic test accuracy meta-analysis. *Intensive &*
781 *critical care nursing* 2021;64:103009. doi: 10.1016/j.iccn.2020.103009
- 782 54. Zimmermann GS, Cremasco MF, Zanei SSV, et al. Pressure injury risk prediction in critical care
783 patients: an integrative review. *Texto & Contexto-Enfermagem* 2018;27(3)
- 784 55. Baris N, Karabacak BG, Alpar SE. The Use of the Braden Scale in Assessing Pressure Ulcers in
785 Turkey: A Systematic Review. *Advances in skin & wound care* 2015;28:349-57. doi:
786 10.1097/01.ASW.0000465299.99194.e6
- 787 56. Gaspar S, Peralta M, Marques A, et al. Effectiveness on hospital-acquired pressure ulcers
788 prevention: a systematic review. *International Wound Journal* 2019;16(5):1087-102. doi:
789 10.1111/iwj.13147

- 790 57. Ontario HQ. Pressure ulcer prevention: an evidence-based analysis. *Ontario health technology*
791 *assessment series* 2009;9(2):1-104.
- 792 58. Kottner J, Dassen T, Tannen A. Inter- and intrarater reliability of the Waterlow pressure sore risk
793 scale: A systematic review. *International Journal of Nursing Studies* 2009;46:369-79. doi:
794 10.1016/j.ijnurstu.2008.09.010
- 795 59. Lovegrove J, Ven S, Miles SJ, et al. Comparison of pressure injury risk assessment outcomes using
796 a structured assessment tool versus clinical judgement: A systematic review. *Journal of*
797 *Clinical Nursing* 2021 doi: 10.1111/jocn.16154 [published Online First: 2021/12/01]
- 798 60. Lovegrove J, Miles S, Fulbrook P. The relationship between pressure ulcer risk assessment and
799 preventative interventions: a systematic review. *Journal of wound care* 2018;27(12):862-75.
- 800 61. Moore ZEH, Patton D. Risk assessment tools for the prevention of pressure ulcers. *Cochrane*
801 *Database of Systematic Reviews* 2019 doi: 10.1002/14651858.CD006471.pub4
- 802 62. Kelly J. Inter-rater reliability and Waterlow's pressure ulcer risk assessment tool. *Nurs Stand*
803 2005;19(32):86-7, 90-2. doi: 10.7748/ns2005.04.19.32.86.c3851
- 804 63. Munoz N, Posthauer ME. Nutrition strategies for pressure injury management: Implementing the
805 2019 International Clinical Practice Guideline. *Nutrition in Clinical Practice* 2022;37(3):567-
806 82.
- 807 64. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of
808 bias in randomised trials. *BMJ* 2011;343:d5928. doi: 10.1136/bmj.d5928
- 809 65. AHRQ Methods for Effective Health Care. Methods Guide for Effectiveness and Comparative
810 Effectiveness Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US)
811 2008.
- 812 66. Zahia S, Garcia Zapirain MB, Sevillano X, et al. Pressure injury image analysis with machine
813 learning techniques: A systematic review on previous and possible future methods. *Artificial*
814 *Intelligence in Medicine* 2020;102:101742. doi: 10.1016/j.artmed.2019.101742
- 815 67. Song M, Choi KS. Factors predicting development of decubitus ulcers among patients admitted
816 for neurological problems. *The Journal of Nurses Academic Society* 1991;21(1):16-26.
- 817 68. Pang SM, Wong TK. Predicting pressure sore risk with the Norton, Braden, and Waterlow scales in
818 a Hong Kong rehabilitation hospital. *Nursing Research* 1998;47(3):147-53.
- 819 69. Cubbin B, Jackson C. Trial of a pressure area risk calculator for intensive therapy patients.
820 *Intensive Care Nursing* 1991;7(1):40-44.
- 821 70. González-Ruiz J, Carrero AG, Blázquez MH, et al. Factores de riesgo de las úlceras por presión en
822 pacientes críticos. *Enfermería Clínica* 2001;11(5):184-90.
- 823 71. Kwong E, Pang S, Wong T, et al. Predicting pressure ulcer risk with the modified Braden, Braden,
824 and Norton scales in acute care hospitals in Mainland China. *Appl Nurs Res* 2005;18(2):122-8.
825 doi: 10.1016/j.apnr.2005.01.001
- 826 72. Halfens R, Van Achterberg T, Bal R. Validity and reliability of the Braden scale and the influence of
827 other risk factors: a multi-centre prospective study. *International Journal of Nursing Studies*
828 2000;37(4):313-19.
- 829 73. Ek AC. Prediction of pressure sore development. *Scand J Caring Sci* 1987;1(2):77-84. doi:
830 10.1111/j.1471-6712.1987.tb00603.x
- 831 74. Bienstein C. Risikopatienten erkennen mit der erweiterten Nortonskala [Risk patients detected
832 with the extended Norton scale]. *Dekubitus - Prophylaxe und Therapie*. Frankfurt/Main:
833 Verlag Krankenpflege 1991.
- 834 75. Jackson C. The revised Jackson/Cubbin Pressure Area Risk Calculator. *Intensive Crit Care Nurs*
835 1999;15(3):169-75. doi: 10.1016/s0964-3397(99)80048-2
- 836 76. Fuentelsaz C. Validation of the EMINA scale: tool for the evaluation of risk of developing pressure
837 ulcers in hospitalized patients. *Enferm Clin [Internet]* 2001;11(3):97-103.
- 838 77. Lowthian P. The practical assessment of pressure sore risk. *Care-Science and Practice*
839 1987;5(4):3-7.

- 840 78. Lowery MT. A pressure sore risk calculator for intensive care patients: 'the Sunderland
841 experience'. *Intensive Crit Care Nurs* 1995;11(6):344-53. doi: 10.1016/s0964-3397(95)80452-
842 8
- 843 79. Lindgren M, Unosson M, Krantz AM, et al. A risk assessment scale for the prediction of pressure
844 sore development: reliability and validity. *Journal of advanced nursing* 2002;38(2):190-99.
- 845 80. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for
846 diagnostic test data. *Stat Med* 2002;21(9):1237-56. doi: 10.1002/sim.1099
- 847 81. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a
848 summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med*
849 1993;12(14):1293-316. doi: 10.1002/sim.4780121403
- 850 82. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new
851 meta-analytic method. *Med Decis Making* 1993;13(4):313-21. doi:
852 10.1177/0272989x9301300408
- 853 83. Webster J, Coleman K, Mudge A, et al. Pressure ulcers: effectiveness of risk-assessment tools. A
854 randomised controlled trial (the ULCER trial). *BMJ Quality & Safety* 2011;20(4):297. doi:
855 10.1136/bmjqs.2010.043109
- 856 84. Saleh M, Anthony D, Parboteeah S. The impact of pressure ulcer risk assessment on patient
857 outcomes among hospitalised patients. *J Clin Nurs* 2009;18(13):1923-9. doi: 10.1111/j.1365-
858 2702.2008.02717.x [published Online First: 2009/04/03]
- 859 85. Moore Z, Johansen E, Etten Mv, et al. Pressure ulcer prevalence and prevention practices: a cross-
860 sectional comparative survey in Norway and Ireland. *Journal of Wound Care* 2015;24(8):333-
861 39. doi: 10.12968/jowc.2015.24.8.333
- 862 86. Moore Z, Pitman S. Towards establishing a pressure sore prevention and management policy in
863 an acute hospital setting. *The All Ireland Journal of Nursing and Midwifery* 2000;1(1):7-11.
- 864 87. Gunningberg L, Lindholm C, Carlsson M, et al. Implementation of risk assessment and
865 classification of pressure ulcers as quality indicators for patients with hip fractures. *J Clin*
866 *Nurs* 1999;8(4):396-406. doi: 10.1046/j.1365-2702.1999.00287.x
- 867 88. Bale S, Finlay I, Harding KG. Pressure sore prevention in a hospice. *J Wound Care* 1995;4(10):465-
868 8. doi: 10.12968/jowc.1995.4.10.465
- 869 89. Hodge J, Mounter J, Gardner G, et al. Clinical trial of the Norton Scale in acute care settings. *Aust*
870 *J Adv Nurs* 1990;8(1):39-46.
- 871 90. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-
872 Analyses: The PRISMA Statement. *PLOS Medicine* 2009;6(7):e1000097. doi:
873 10.1371/journal.pmed.1000097
- 874 91. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and
875 external validation. *J Clin Epidemiol* 2016;69:245-7. doi: 10.1016/j.jclinepi.2015.04.005
876 [published Online First: 2015/04/18]
- 877 92. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic
878 accuracy: explanation and elaboration. *Ann Intern Med* 2003;138(1):W1-12. doi:
879 10.7326/0003-4819-138-1-200301070-00012-w1
- 880 93. Reitsma J, Rutjes A, Whiting P, et al. Chapter 8: Assessing risk of bias and applicability. In: Deeks J,
881 Bossuyt P, Leeflang M, et al., eds. *Cochrane Handbook for Systematic Reviews of Diagnostic*
882 *Test Accuracy*. Version 2.0 (updated July 2023) ed. Cochrane, 2023.
- 883 94. Deeks JJ, Dealey C. Pressure sore prevention: using and evaluating risk assessment tools. *British*
884 *Journal of Nursing* 1996;5(5):313-20. doi: 10.12968/bjon.1996.5.5.313
- 885 95. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction
886 models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6. doi: 10.1136/bmj.i6
- 887 96. Trikalinos TA, Siebert U, Lau J. Decision-Analytic Modeling to Evaluate Benefits and Harms of
888 Medical Tests: Uses and Limitations. *Medical Decision Making* 2009;29(5):E22-E29. doi:
889 10.1177/0272989X09345022

- 890 97. Dweekat OY, Lam SS, McGrath L. Machine Learning Techniques, Applications, and Potential Future
891 Opportunities in Pressure Injuries (Bedsore) Management: A Systematic Review.
892 *International journal of environmental research and public health* 2023;20(1) doi:
893 10.3390/ijerph20010796
- 894 98. Berlowitz DR, VanDeusen Lukas C, Parker V, et al. 3F: Care Plan. Preventing pressure ulcers in
895 hospitals: A toolkit for improving quality of care: Agency for Healthcare Research and Quality,
896 2014:140-42.
- 897 99. Hingorani AD, Windt DAvd, Riley RD, et al. Prognosis research strategy (PROGRESS) 4: Stratified
898 medicine research. *BMJ : British Medical Journal* 2013;346:e5793. doi: 10.1136/bmj.e5793
- 899 100. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model
900 updating, and impact assessment. *Heart* 2012;98(9):691-8. doi: 10.1136/heartjnl-2011-
901 301247 [published Online First: 2012/03/07]
- 902 101. Shi C, Dumville JC, Cullum N. Evaluating the development and validation of empirically-derived
903 prognostic models for pressure ulcer risk assessment: A systematic review. *International*
904 *journal of nursing studies* 2019;89:88-103. doi: 10.1016/j.ijnurstu.2018.08.005
- 905 102. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction
906 model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann*
907 *Intern Med* 2015;162(1):W1-73. doi: 10.7326/m14-0698
- 908 103. Lee J, Mulder F, Leeftang M, et al. QUAPAS: An Adaptation of the QUADAS-2 Tool to Assess
909 Prognostic Accuracy Studies. *Annals of Internal Medicine* 2022;175(7):1010-18. doi:
910 10.7326/m22-0276