

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
31
32
33
34

Development and validation of risk prediction tools for pressure injury occurrence: An umbrella review

Bethany Hillier^{1,2}

Katie Scandrett^{1,2}

April Coombe^{1,2}

Tina Hernandez-Boussard³

Ewout Steyerberg⁴

Yemisi Takwoingi^{1,2}

Vladica Velickovic^{5,6}

Jacqueline Dinnes^{1,2*}

Affiliations

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

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

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

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

⁵ Evidence Generation Department, HARTMANN GROUP, Heidenheim, Germany

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

Email addresses

b.hillier@bham.ac.uk (BH); k.e.scandrett@bham.ac.uk (KS); a.r.coombe@bham.ac.uk (AC); boussard@stanford.edu (THB); e.w.steyerberg@lumc.nl (ES); y.takwoingi@bham.ac.uk (YT); vladica.velickovic@hartmann.info (VV)

* Corresponding author: j.dinnes@bham.ac.uk (JD)

Keywords

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Development, internal, external validation, prediction, prognostic, pressure injury, ulcer, overview

35 ABSTRACT

36 Background

37 Pressure injuries (PIs) place a substantial burden on healthcare systems worldwide. Risk stratification
38 of those who are at risk of developing PIs allows preventive interventions to be focused on patients
39 who are at the highest risk. The considerable number of risk assessment scales and prediction
40 models available underscore the need for a thorough evaluation of their development, validation and
41 clinical utility.

42 Our objectives were to identify and describe available risk prediction tools for PI occurrence, their
43 content and development and validation methods used.

44 Methods

45 The umbrella review was conducted according to Cochrane guidance. MEDLINE, Embase, CINAHL,
46 EPISTEMONIKOS, Google Scholar and reference lists were searched to identify relevant systematic
47 reviews. Risk of bias was assessed using adapted AMSTAR-2 criteria. Results were described
48 narratively. All included reviews contributed to build a comprehensive list of risk prediction tools.

49 Results

50 We identified five systematic reviews describing the development and validation of risk prediction
51 tools for pressure injuries, 16 that assessed the prognostic accuracy of the tools and 10 that assessed
52 the clinical effectiveness. Of the five reviews of model development and validation, four included
53 only machine learning models. One review included detail about external validation, and this was the
54 only review to include model performance metrics. Where quality assessment was completed (3 out
55 of 5 reviews), most prediction tools were assessed by review authors as being at high risk of bias and
56 no tools were assessed as being at low risk of bias.

57 Conclusions

58 Available tools do not meet current standards for the development or reporting of risk prediction
59 models. The majority of tools have not been externally validated. Standardised and rigorous
60 approaches to risk prediction model development and validation are needed.

61 Registration

62 The protocol was registered on the Open Science Framework (<https://osf.io/tepyk>).

63

64

65 INTRODUCTION

66 Pressure injuries (PI) carry a significant healthcare burden. A recent meta-analysis estimated the
67 global burden of PIs to be 13%, two-thirds of which are hospital-acquired PIs (HAPI).¹ The average
68 cost of a HAPI has been estimated as \$11k per patient, totalling at least \$27 billion a year in the
69 United States based on 2.5 million reported cases.² Length of hospital stay is a large contributing
70 cost, with patients over the age of 75 who develop HAPI having on average a 10-day longer hospital
71 stay compared to those without PI.³

72 PIs result from prolonged pressure, typically on bony areas like heels, ankles, and the coccyx, and are
73 more common in those with limited mobility, including those who are bedridden or wheelchair
74 users. PIs can develop rapidly, and pose a threat in community, hospital and long-term care settings.
75 Multicomponent preventive strategies are needed to reduce PI incidence⁴ with timely
76 implementation to both reduce harm and burden to healthcare systems.⁵ Where preventive
77 measures fail or are not introduced in adequate time, PI treatment involves cleansing, debridement,
78 topical and biophysical agents, biofilms, growth factors and dressings^{6-7,8}, and in severe cases, surgery
79 may be necessary.^{5,9}

80 A number of clinical assessment scales for assessing the risk of PI are available (e.g. Braden^{10,11},
81 Norton¹², Waterlow¹³) but are limited by reliance on subjective clinical judgment. Statistical risk
82 prediction models may offer improved accuracy over clinical assessment scales, however appropriate
83 methods of development and validation are required.^{14,15,16} Although methods for developing risk
84 prediction models have developed considerably,^{14,15,17,18} methodological standards of available
85 models have been shown to remain relatively low.^{17,19-22} Machine learning (ML) algorithms to
86 develop prediction models are increasingly commonplace, but these models are at similarly high risk
87 of bias²³ and do not necessarily offer any model performance benefit over the use of statistical
88 methods such as logistic regression.²⁴ Methods for systematic reviews of risk prediction model
89 studies have also improved,²⁵⁻²⁷ with tools such as PROBAST (Prediction model Risk of Bias
90 Assessment Tool)²⁸ now available to allow critical evaluation of study methods.

91 Although several systematic reviews of PI risk assessment scales and risk prediction models for PI
92 (subsequently referred to as risk prediction tools) are available²⁹⁻³⁸, these have been demonstrated to
93 frequently focus on single or small numbers of scales or models, use variable review methods and
94 show a lack of consensus about the accuracy and clinical effectiveness of available tools.³⁹ We
95 conducted an umbrella review of systematic reviews of risk prediction tools for PI to gain further
96 insight into the methods used for tool development and validation, and to summarise the content of
97 available tools.

98 METHODS

99 Protocol registration and reporting of findings

100 We followed guidance for conducting umbrella reviews provided in the Cochrane Handbook for
101 Intervention Reviews.⁴⁰ The review was reported in accordance with guidelines for Preferred
102 Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)⁴¹ (see Appendix 1), adapted for
103 risk prediction model reviews as required. The protocol was registered on the Open Science
104 Framework (<https://osf.io/tepyk>).

105 Literature search

106 A single sensitive search strategy, developed and tested by an experienced information specialist
107 (AC), was conducted in January 2023. Electronic searches of MEDLINE and Embase via Ovid and

108 CINAHL Plus EBSCO were conducted, employing well-established systematic review and prognostic
109 search filters⁴²⁻⁴⁴ combined with specific keyword and controlled vocabulary terms relating to PI
110 appropriate to each source, without restrictions on publication year. Additional simplified searches
111 were undertaken in EPISTEMONIKOS and Google Scholar due to the more limited search functionality
112 of these two sources. The reference lists of all publications reporting reviews of prediction tools
113 (systematic or non-systematic) were reviewed to identify additional eligible systematic reviews and
114 to populate a list of PI risk prediction tools. Title and abstract screening and full text screening were
115 conducted independently and in duplicate by two of four reviewers (BH, JD, YT, KS). Any
116 disagreements were resolved by discussion or referral to a third reviewer.

117 Eligibility criteria for this umbrella review

118 Published English-language systematic reviews of risk prediction models developed for adult patients
119 at risk of PI in any setting were included. Reviews of clinical risk assessment tools or models
120 developed using statistical or ML methods were included, both with or without internal or external
121 validation. The use of any PI classification system^{6 45-47} as a reference standard was eligible. Reviews
122 of the diagnosis or staging of those with suspected or existing PIs or chronic wounds, reviews of
123 prognostic factor and predictor finding studies, and models exclusively using pressure sensor data
124 were excluded.

125 Systematic reviews were required to report a comprehensive search of at least two electronic
126 databases, and at least one other indicator of systematic methods (i.e. explicit eligibility criteria,
127 formal quality assessment of included studies, sufficient data presented to allow results to be
128 reproduced, or review stages (e.g. search screening) conducted independently in duplicate.

129 Data extraction and quality assessment

130 Data extraction forms (Appendix 3) were developed using the CHARMS checklist (CHecklist for critical
131 Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) and Cochrane
132 Prognosis group template.^{48 49} One reviewer extracted data concerning: review characteristics, model
133 details, number of studies and participants, study quality and results. Extractions were
134 independently checked by a second reviewer. Where discrepancies in model or primary study details
135 were noted between reviews, we accessed the primary model development publications where
136 possible.

137 The methodological quality of included systematic reviews was assessed using AMSTAR-2 (A
138 Measurement Tool to Assess Systematic Reviews)⁵⁰, adapted for systematic reviews of risk prediction
139 models (Appendix 4). Quality assessment and data extraction were conducted by one reviewer and
140 checked by a second (BH, JD, KS), with disagreements resolved by consensus. Our adapted AMSTAR-2
141 contains six critical items, and limitations in any of these items reduces the overall validity of a
142 review.⁵⁰

143 Synthesis methods

144 Reviews were considered according to whether any information concerning model development and
145 validation were reported. Available data were tabulated and a narrative synthesis provided. All risk
146 prediction models identified are listed in Appendix 5 Table S4, including those for which no
147 information about model development or validation was provided at systematic review level. Where
148 possible, the predictors included in the tools were extracted at review level and categorised into
149 relevant groups in order to describe the candidate predictors associated with risk of PI. No statistical
150 synthesis of systematic review results was conducted. Reviews reporting results as prognostic
151 accuracy (i.e. risk classification according to a binary decision) or clinical effectiveness (i.e. impact on
152 patient management and outcomes) are reported elsewhere.³⁹

153 RESULTS

154 Characteristics of included reviews

155 Following de-duplication of search results, 6301 unique records remained, of which 110 were
156 selected for full text assessment. We obtained the full text of 104 publications of which 28 met all
157 eligibility criteria for inclusion (see Figure 1). Five reviews reported details about model development
158 and internal validation^{36 37 51-53}, one of which also considered external validation⁵²; 16 reported
159 accuracy data^{29 31-35 38 54-62}; and 10 reported clinical effectiveness data.^{30 54 56 59 63-68} Three reviews
160 reported both accuracy and effectiveness data.^{54 56 59}

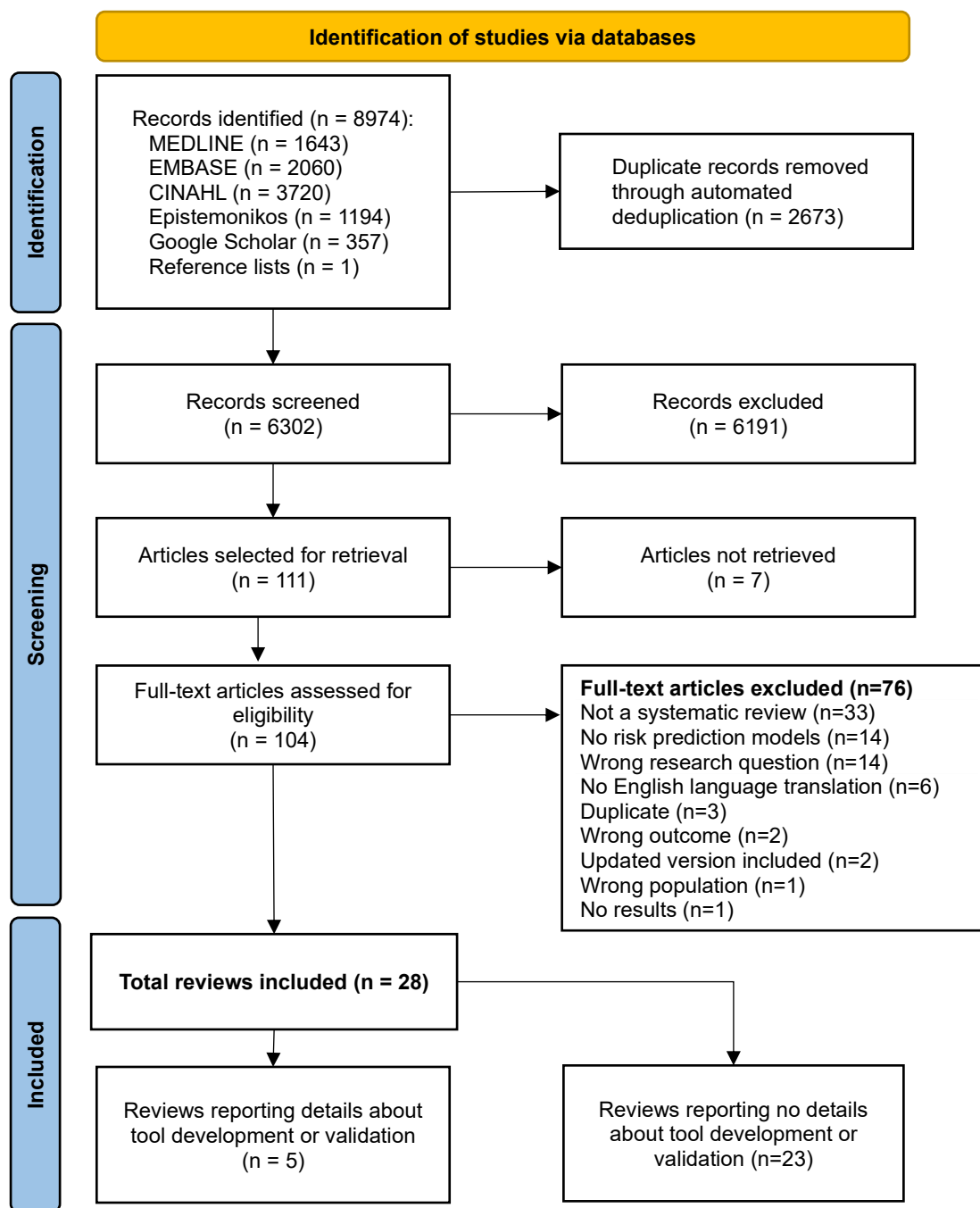
161 Table 1 provides a summary of systematic review methods for all 28 reviews according to whether or
162 not they reported any tool development methods (see Appendix 5 for full details). The five reviews
163 reporting prediction tool development and validation were all published within the last five years
164 (2019 to 2023) compared to reviews focused on the clinical utility of available tools (published from
165 2006 to 2022). Reviews about tool development focused primarily on ML-based models (one⁵⁸ of the
166 five reviews limited inclusion to empirically derived models including ML-based models, and four
167 (80%) considered only ML models) and frequently did not report study eligibility criteria related
168 study participants or setting. Only one review (4%) concerning the clinical utility of models included
169 ML-based models,³⁸ but more often restricted study inclusion by population or setting: hospital
170 inpatients (ward or acute unit) (n = 4),^{33 38 63 67} acute settings (n=5),^{34 59 61 62 66} or surgical patients
171 (n=1)³¹, or studies in long-term care settings (n=2)^{29 35} or the elderly (n=1).⁵⁸

172 On average reviews about tool development included more studies than reviews of clinical utility
173 (median 22 compared to 13), more participants (median 234,105 compared to 6,106) and covered
174 more prediction tools (median 21 compared to 3) (Table 1). Eight reviews (35%) about clinical utility
175 included only one risk assessment scale, whereas reviews of tool development included at least 3
176 different risk prediction models. The PROBAST tool for quality assessment of prediction model
177 studies was used in 60% (n=3) of tool development reviews^{37 52 53} compared to none of the reviews of
178 clinical utility, however the remaining two reviews of tool development did not report any quality
179 assessment of included studies (2 (40%) compared to 4 (17%) of reviews of clinical utility). Meta-
180 analysis was conducted in one of five (20%) reviews of tool development compared to more than
181 half of reviews of clinical utility (13, 57%).

182 Methodological quality of included reviews

183 The quality of included reviews was generally low (Figure 2; Appendix 5 for detailed assessments).
184 The majority of reviews (all five reviews about tool development and 70% (16/23) reviews of clinical
185 utility) partially met the AMSTAR-2 criteria for the literature search (i.e. searched two databases,
186 reported search strategy or key words, and justified language restrictions if used), with only two
187 (both reviews of clinical utility) meeting all criteria for 'Yes' (i.e. searching grey literature and
188 reference lists, with the search conducted within 2 years of publication). Nineteen reviews (68%)
189 conducted study selection in duplicate (4/5 (80%) of review about tool development and 15/23
190 (65%) of clinical utility reviews). Conflicts of interest were reported in all five tool development
191 reviews and 74% of clinical utility reviews (17/23). Reviews scored poorly on the remaining AMSTAR-
192 2 items, with at most half of reviews meeting the stipulated AMSTAR-2 criteria. Seven reviews (25%)
193 used an appropriate method of quality assessment of included studies and provided itemisation of
194 judgements per study. No review scored 'Yes' for all AMSTAR-2 items in either category.

195 Figure 1. PRISMA⁴¹ flowchart: identification, screening and selection process



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

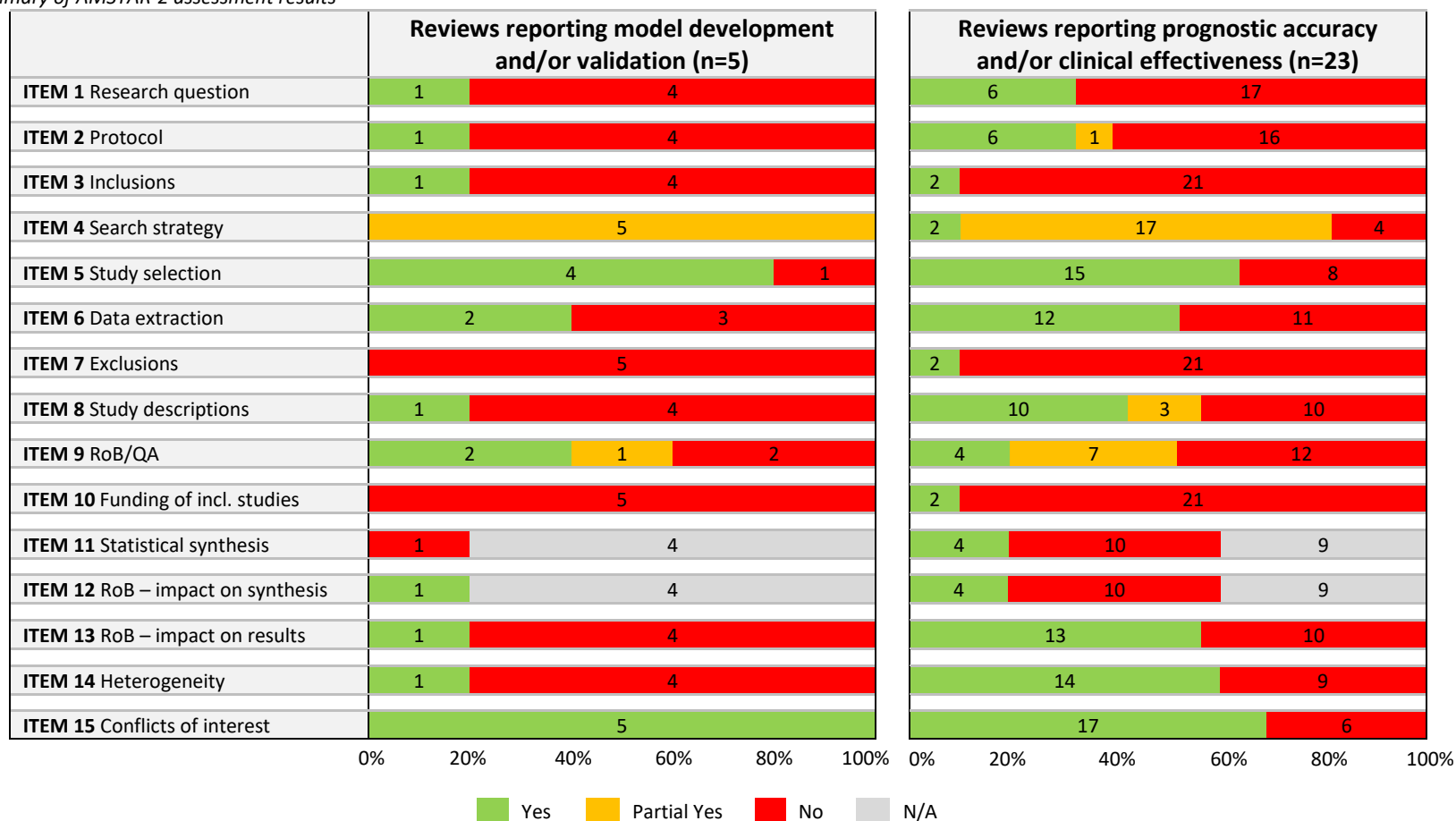
197 Table 1. Summary of included systematic review characteristics

Review characteristics	Reviews on model development and validation (N=5)	Reviews on accuracy or clinical effectiveness (N=23)	All included reviews (N=28)
Median (range) year of publication	2021 (2019 – 2023)	2016 (2006 – 2022)	2018 (2006 – 2023)
Eligibility criteria			
Participants			
Adults only	0 (0)	13 (57) ^A	13 (46) ^A
Any age	0 (0)	3 (13)	3 (11)
No age restriction reported	5 (100)	7 (30)	12 (43)
Presence of PI at baseline			
No PIs at baseline	0 (0)	6 (26)	6 (21)
NS	5 (100)	17 (74)	22 (79)
Setting			
Any healthcare setting	0 (0)	10 (43)	10 (36)
Hospital	1 (20)	4 (22)	5 (18)
Long-term care	0 (0)	2 (9)	2 (7)
Acute care (incl. surgical and ICU)	0 (0)	6 (26)	6 (21)
Long-term, acute or community settings	0 (0)	1 (4)	1 (4)
NS	4 (80)	0 (0)	4 (14)
Risk assessment tools			
ML-based prediction models	4 (80)	1 (4)	5 (18)
ML or statistical models	1 (20)	0 (0)	1 (4)
Any prediction tool or scale	0 (0)	9 (39)	9 (32)
Specified clinical scale(s)	0 (0)	10 (43)	10 (36)
PI prevention strategies	0 (0)	1 (4)	1 (4)
NS	0 (0)	2 (9)	2 (7)
PI classification system			
Any	0 (0)	1 (4)	1 (4)
Accepted standard classifications	0 (0)	2 (9)	2 (7)
Several specified classification systems (NPUAP, EPUAP, AHCPH or TDCPS)	0 (0)	3 (13)	3 (11)
Other	0 (0)	1 (4)	1 (4)
NS	5 (100)	16 (70)	21 (75)
Source of data			
Prospective only	0 (0)	4.5 (20) ^B	4.5 (16) ^B
Prospective or retrospective	1 (20)	2.5 (11) ^B	3.5 (13) ^B
NS	4 (80)	16 (70)	20 (71)
Study design restrictions			
Yes	1 (20)	12 (52)	13 (46)
No	0 (0)	3 (13)	3 (11)
NS	4 (80)	8 (39)	12 (43)
Review methods			
Median (range) no. sources^C searched	5 (2 – 8)	6 (2 – 14)	5 (2 – 14)
Publication restrictions:			
End date (year)			
2000-2009	0 (0)	3 (13)	3 (11)
2010-2019	1 (20)	16 (70)	17 (61)
2020-2023	4 (80)	4 (17)	8 (29)
Language			
English only	4 (80)	9 (39)	13 (46)
2 languages	0 (0)	2 (9)	2 (7)
>2 languages	0 (0)	3 (13)	3 (11)
No restrictions	0 (0)	4 (17)	4 (14)
NS	1 (20)	5 (22)	6 (21)

Quality assessment tool ^D			
PROBAST	3 (60)	0 (0) ^E	3 (11) ^E
QUADAS	0 (0)	2 (9)	2 (7)
QUADAS-2	0 (0)	7 (30)	7 (25)
JBI tools	0 (0)	3 (13)	3 (11)
CASP	0 (0)	2 (9)	2 (7)
Other	0 (0)	6 (26)	6 (21)
None	2 (40)	4 (17)	6 (21)
Meta-analysis included	1 (20)	13 (57)	14 (50)
Method of meta-analysis (% of reviews incl. meta-analysis)			
Univariate RE/FE model (depending on heterogeneity assessment)	1 (100) ^F	2 (15) ^F	3 (21)
Univariate RE model	0 (0)	5 (38) ^F	5 (36) ^F
Hierarchical model (for DTA studies)	0 (0)	2 (15)	2 (14)
Unclear/NS	0 (0)	4 (31) ^F	4 (29) ^F
Volume of evidence			
Median (range) no. studies	22 (3 – 35)	13 (1 – 70)	16 (1 – 70)
Median (range) no. participants	234105 (6674 – 1278148)	6106 (528 – 221541)	10044 (528 – 1278148)
Median (range) no. tools	21 (3 – 35)	3 (1 – 28)	4 (1 – 35)

198 Figures are number (%) of reviews, unless otherwise specified. ^A one review⁵⁸ restricted to aged >60 years; ^B one review ⁵⁴
199 states either prospective or retrospective data eligible for Research Question 1, but prospective only for Research Question
200 2, hence 0.5 added to each category; ^C including databases, bibliographies or registries; ^D reviews may fall into multiple
201 categories, therefore total number within domain not necessarily equal to N (100%); ^E one review³⁸ reported use of
202 PROBAST in methods, but did not present any PROBAST results; ^F one review conducts univariate meta-analysis for a single
203 estimate, e.g. c-statistic⁵², AUC⁶⁰, RR⁵⁵, or OR.⁵⁶
204 AHCPR – Agency for Health Care Policy and Research; CASP – Critical Appraisal Skills Programme; DTA – diagnostic test
205 accuracy; EPUAP – European Pressure Ulcer Advisory Panel; FE – fixed effects; ICU – intensive care unit; JBI – Joanna Briggs
206 Institute; ML – machine learning; NPUAP – National Pressure Ulcer Advisory Panel; NS – not stated; PI – pressure injury;
207 PROBAST – Prediction model Risk of Bias Assessment; QUADAS (2) – Quality Assessment of Diagnostic Accuracy Studies
208 (Version 2); RE – random effects; TDCPS – Torrance Developmental Classification of Pressure Sore.

209 *Figure 2. Summary of AMSTAR-2 assessment results*



210 AMSTAR – A MeaSurement Tool to Assess systematic Reviews; Item 1 – Adequate research question/ inclusion criteria?; Item 2 – Protocol and justifications for deviations?; Item 3 – Reasons
 211 for study design inclusions?; Item 4 – Comprehensive search strategy?; Item 5 – Study selection in duplicate?; Item 6 – Data extraction in duplicate?; Item 7 – Excluded studies list (with
 212 justifications)?; Item 8 – Included studies description adequate?; Item 9 – Assessment of RoB/quality satisfactory?; Item 10 – Studies’ sources of funding reported?; Item 11 – Appropriate
 213 statistical synthesis method?; Item 12 – Assessment of impact of RoB on synthesised results?; Item 13 – Assessment of impact of RoB on review results?; Item 14 – Discussion/investigation of
 214 heterogeneity?; Item 15 – Conflicts of interest reported?; N/A – not applicable; RoB – risk of bias; QA – quality assessment. Further details on AMSTAR items are given in Appendix 4, and
 215 results per review are given in Appendix 5. Note that where AMSTAR-2 assessment was applied to overlapping reviews (n=3) for prognostic accuracy and clinical effectiveness separately, and
 216 resulted in differing judgements for each review question, the judgements for the prognostic accuracy review question are displayed here for simplicity.

Table 2. Results of reviews reporting model development and validation

Review author (publication year)	DEV/ VAL (no. studies)	Setting of included studies; data sources	Model development algorithms	Internal validation method	Brief description of study quality	Summary of model performance results
Dweekat ³⁶ (2023)	DEV (34); unclear (1) ^A	HAPI/CAPI n=32; SRPI n=2; detection of PI (effect on length of stay) n=1; nursing home residents n=2	LR n=20; RF n=18; DT n=12; SVM n=12; MLP n=9; KNN n=4; LDA n=1; other n=19	CV n=10; split sample n=10; split sample and CV n=8; NS=7	No RoB assessment	N/A
Jiang ³⁷ (2021)	DEV (9)	ICU n=3; operating room n=2; acute care hospital n=1; oncology department n=1; end-of-life care n=1; mobility-related disabilities n=1 EHRs used in all models	DT n=5; LR n=3; SVM n=2; NN n=2; RF n=1; MTS n=1; BN n=1; gradient boosting n=1	Split sample n=4; NS n=9	RoB assessed using PROBAST. Overall RoB high for all predictive models. All models at high RoB in analysis domain.	F-score ranged between 0.377 (ML Su MTS) and 0.670 (ML Su LR); g-means ranged between 0.628 (ML Kaewprag BN) and 0.822 (ML Su MTS); sensitivity ranged between 0.478 (ML Kawprag) and 0.848 (ML Yang); specificity ranged between 0.703 (ML Deng) and 0.988 (ML Su LR)
Ribeiro ⁵¹ (2021)	DEV (3)	SRPI cardiovascular n=2; SRPI critical care n=1 EHRs used in n=2 models	ANN n=1; XGBoost n=1; RF n=1	Split sample n=2; NS n=1	No RoB assessment	N/A
Shi ⁵² (2019)	DEV (21); VAL (7)	<i>DEV</i> General acute care hospital n=7; long-term care n=5; specific acute care (e.g. ICU) n=4; cardiovascular surgery n=2; trauma and burn centres n=1; rehabilitation units n=1; unclear n=1 Prospective n=10; retrospective n=11 <i>VAL</i> Long-term care n=3; specific acute care (e.g. ICU) n=2; general (acute care) hospital n=2	LR n=16; cox regression n=5; ANN n=1; DT n=1; discriminant analysis n=1; C4.5 machine learning (DT induction algorithm) n=1; NS n=1	CV n=1; tree-pruning n=1; split sample n=1; re-sampling n=2; NS n=16	RoB assessed using PROBAST. <i>DEV</i> Overall RoB unclear for two models. Overall RoB high for the remaining 19 models. Analysis and outcome domains were mostly at high RoB. <i>VAL</i> Overall RoB unclear for three validation studies. Overall RoB high for the remaining four validation studies. Analysis and outcome domains were mostly at high RoB.	C-statistics ^B ranged between 0.61 (interRAI PURS) and 0.90 (TNH-PUPP); O/E ratios ^B ranged between 0.91 (Berlowitz MDS) and 1.0 (prePURSE study tool) <i>Pooled C-statistics^B</i> TNH-PUPP: 0.86 (95% CI 0.81–0.90), n=2 Fragmment scale: 0.79 (95% CI 0.77–0.82), n=1 ^C Berlowitz 11-item model: 0.75 (95% CI 0.74–0.76), n=2 Berlowitz MDS model: 0.73 (95% CI 0.72–0.74), n=2 interRAI PURS: 0.65 (95% CI 0.60–0.69), n=3 Compton: 0.81 (95% CI 0.78–0.84), n=2 <i>Pooled O/E ratios^B</i> 0.99 (95% CI 0.95–1.04), n=2

Review author (publication year)	DEV/ VAL (no. studies)	Setting of included studies; data sources	Model development algorithms	Internal validation method	Brief description of study quality	Summary of model performance results
		Prospective n=3; retrospective n=4				Berlowitz MDS 0.94 (95% CI 0.88–1.01), n=2
Zhou ⁵³ (2022)	DEV (22)	SRPI n=3; ICU n=11; hospitalised n=6; rehabilitation centre n=1; hospice n=1 EHR n=18; Medical Information Mart for Intensive care III database n=4	LR n=15; RF n=10; DT n=9; SVM; n=9; ANN n=8; BN n=3; XGBoost n=3; GB n=2; AdaBoost n=1; CANTRIP n=1; LSTM n=1; EN n=1; KNN n=1; MTS n=1; NB n=1	CV n=12; NS n=10	RoB assessed using PROBAST. Overall RoB unclear for five studies. Overall RoB high for 15 models. RoB not assessed in two studies due to use of unstructured data.	F1 score ranged between 0.02 (ML Nakagami) and 0.99 (ML Song [2]); AUC ranged between 0.78 (ML Delparte) and 0.99 (ML Song [2]); sensitivity ranged between 0.08 (ML Cai) and 0.99 (ML Song [2]); specificity ranged between 0.63 (ML Delparte) and 1 (ML Cai)

218 ^A Appears to be a model validation study but the review only included model development studies.

219 ^B Values from fixed-effects meta-analyses, pooling development and external validation study estimates together.

220 ^C One data source but included two C-statistic values (one for model development and one for internal validation) that were subsequently pooled.

221 AUC – area under curve; ANN – artificial neural network; BN – Bayesian network; CAPI – community-acquired pressure injury; CANTRIP - reCurrent Additive Network for Temporal Risk

222 Prediction; CV – cross-validation; DEV – development; DT – decision tree; EHRs – electronic health records; EN – elastic net; GB – gradient boosting; HAPI – hospital-acquired pressure injury;

223 ICU – intensive care unit; KNN – k-nearest neighbours; LDA – linear discriminant analysis; LSTM – long short-term memory; LR – logistic regression; ML – machine learning; MLP – multilayer

224 perception; MTS – Mahalanobis-Taguchi system; N/A – not applicable; NB – naïve Bayes; NN – neural network; O/E – observed vs expected; PI – pressure injury; PROBAST – Prediction model

225 Risk of Bias ASsessment Tool; RF – random forest; RoB – risk of bias; SRPI – surgery-related pressure injury; SVM – support vector machine; VAL – validation.

226 Findings

227 Of the 28 reviews, 23 reviews focused on the clinical utility of prediction tools provided no details
228 about the development or validation of included models (see Appendix 5), and gave only limited
229 detail about setting and study design. These reviews included a total of 63 different prediction tools,
230 predominantly derived by clinical experts, as opposed to empirically-derived models. The most
231 commonly included tools were the Braden (included in 20 reviews), Waterlow (n=14 reviews),
232 Norton (n=11 reviews), and Cubbin and Jackson scales (n=8 reviews).

233 The five systematic reviews that reported detailed information about model development and
234 validation included 62 prediction models, 40 of which were unique to these five reviews (between
235 one³⁷ and 20⁵² unique models were included per review) (Table 2). Between three⁵¹ and 35³⁶ model
236 development studies were included; one review⁵² also included 7 external validation studies.
237 Electronic health records (EHRs) were used for model development in all studies in one review³⁷ and
238 in 18 of 22 models (82%) in a second review.⁵³ One review⁵² reported the use of prospectively or
239 retrospectively collected data (n=10 studies and n=11 studies, respectively). No review included
240 information about the thresholds used define whether a patient is at risk of developing PIs. Four
241 reviews included detail about the predictors included in each model.

242 The largest review³⁶ reported that logistic regression was the most commonly reported modelling
243 approach (20/35 models), followed by random forest (n=18), decision tree (n=12) and support vector
244 machine (n=12) approaches. Logistic regression was also the most frequently used in two other
245 reviews (16/21⁵² and 15/22⁵³). Primary studies frequently compared the use of different ML methods
246 using the same datasets, such that 'other' ML methods were reported with little to no further detail
247 (e.g. 19 studies in the review by Dweekat and colleagues³⁶).

248 Approaches to internal validation were not well reported in the primary studies. One review⁵² found
249 no information on internal validation for 76% (16/21) of studies; with re-sampling reported in two
250 and tree-pruning, cross-validation and split sample reported in one study each. Another³⁶ reported
251 no information about internal validation for 20% of studies (7/35) and the use of cross-validation
252 (n=10), split sample (n=10) techniques, or both (n=8) for the remainder. Cross-validation was used in
253 more than half (12/22) of studies in another.⁵³ Only one review reported any detail about methods
254 for selection of model predictors⁵²: 29% (6/21) selected predictors by univariate analysis prior to
255 modelling and 9 used stepwise selection for final model predictors; 11 (52%) clearly reported
256 candidate predictors, and all 21 clearly reported final model predictors. The same review⁵² reported
257 15 models (71%) with no information about missing data, and only two using imputation techniques.

258 Model performance measures were reported by three reviews^{37 52 53}, all of which noted considerable
259 variation in reported metrics and model performance including C-statistics (0.71 to 0.89 in 10
260 studies⁵³), F1 score (0.02 to 0.99 in 9 studies⁵³), G-means (0.628 to 0.822 in four studies³⁷), and
261 observed versus expected ratios (0.97 to 1 in 3 studies⁵²). Two reviews reported accuracy metrics
262 associated with included models: sensitivity ranged between 0.48 and 0.85 and specificity between
263 0.70 and 0.99 for 7 models in one review,³⁷ compared to sensitivity 0.08 to 0.99 (for 19 studies⁵²) and
264 specificity 0.63 to 1.00 (for 18 studies⁵²) in another. AUC ranged between 0.78 and 0.99 for 16
265 studies⁵².

266 Shi and colleagues⁵² included 7 external validations using data from long-term care (n=3) or acute
267 hospital care (n=4) settings. All were judged to be at unclear (n=3) or high (n=4) risk of bias using
268 PROBAST. Model performance metrics for five models (TNH-PUPP⁶⁹, Berlowitz 11-item model⁷⁰,
269 Berlowitz MDS adjustment model⁷¹, interRAI PURS⁷², Compton ICU model⁷³) included C-statistics
270 between 0.61 and 0.9 and reported observed versus expected ratios were between 0.91 and 0.97.

271 The review also reported external validation studies for the 'SS scale' and the prePURSE study tool,
272 but no model performance metrics were given. A meta-analysis of C-statistics and O/E ratios was
273 performed, including values from both development and external validation cohorts (Table 2).
274 Parameters related to model development were not consistently reported: C-statistics ranged
275 between 0.71 and 0.89 (n = 10 studies); observed versus expected ratios ranged between 0.97 and 1
276 (n=3 studies). For more detailed information about individual models, including predictors, specific
277 model performance metrics and sample sizes (Appendix 5).

278 Included tools and predictors

279 A total of 116 risk prediction tools were identified (Table 3); 103 were identified from the 28 included
280 systematic reviews and 13 were identified from screening the reference lists of the 'non-systematic'
281 reviews that were considered during full text assessment. Full details obtained at review-level are
282 reported in Appendix 5 Table S4.

283 Tools were categorised as having been developed with (52/116, 45%) or without (64, 55%) the use of
284 ML methods. Prospectively collected data was used for model development for 19% of tools
285 (22/116), retrospectively collected data for 40% (46/116), including 18 ML-based models using EHRs,
286 or was not reported (48/116). Information about the study populations was poorly reported,
287 however study setting was reported for 102 prediction tools. Thirty-four tools were reported to have
288 been developed in hospital inpatients, and 20 were developed in long-term care settings,
289 rehabilitation units or nursing homes or hospices. Where reported (n=92), sample sizes ranged from
290 15⁷⁴ to 1,252,313.⁷⁵⁻⁸² The approach to internal validation used for the prediction tools (e.g. cross-
291 validation versus split sample) was not reported at review-level for almost three quarters of tools
292 (81/116, 70%).

293 We could extract information about the predictors for only 53 of the 116 tools (Table 4 and Appendix
294 5). The most frequently included predictor was mobility (27/53, 51%), followed by pre-disposing
295 diseases/conditions (26/53, 49%), medical treatment/care received (22/53, 42%) and continence
296 (22/53, 42%). Tools often (23/53, 45%) included multiple pre-existing conditions or comorbidities and
297 multiple types of treatment or medication as predictors. Other common predictors include age,
298 nutrition, mental status, activity, skin conditions and laboratory values (34% to 40% of models).
299 Seven tools incorporated scores from other established risk prediction scales as a predictor, with six
300 including Braden scores and one including the Norton score.

301 Only one review⁵² reported the presentation format of included tools, coded as 'score system'
302 (n=11), 'formula equation' (n=3), 'nomogram scale' (n=2), or 'not reported' (n=6).

303

304

305 *Table 3. Summary of tool characteristics, extracted at review-level*

Tool characteristics	ML-based models (N=52, 45%)	Non-ML tools (N=64, 55%)	Total (N=116)
No. of included reviews^A considered in			
0	0 (0)	13 (20)	13 (11)
1	30 (58)	23 (36)	53 (46)
2	8 (15)	9 (14)	17 (15)
>2	14 (27)	19 (30)	33 (28)
Development study details			
Median (range) year of publication	2020 (2000 – 2022)	1998 (1962 – 2015)	2006 (1962 – 2022)
Source of data			
Prospective	4 (8)	18 (28)	22 (19)
Retrospective	36 (69)	10 (16)	46 (40)
NS	12 (23)	36 (56)	48 (41)
Setting			
Hospital	13 (25)	11 (17)	24 (21)
Long-term care (incl. end-of-life and rehab)	6 (12)	14 (22)	20 (17)
Acute care (incl. surgical and ICU)	28 (54)	24 (38)	52 (45)
Mixed settings	1 (2)	1 (2)	2 (2)
Other	2 (4)	2 (3)	4 (3)
NS	2 (4)	12 (19)	14 (12)
Study population age			
Adults	31 (60)	34 (53)	65 (56)
Any	5 (10)	3 (5)	8 (7)
NS	16 (31)	27 (42)	43 (37)
Baseline condition			
PIs at baseline	1 (2)	0 (0)	1 (1)
No PIs at baseline	11 (21)	19 (30)	30 (26)
NS	40 (77)	45 (70)	85 (73)
Development methods			
Development method/algorithm^B			
ML algorithms	52 (100)	0 (0)	52 (45)
Logistic regression	35 (67)	15 (23) ^C	50 (43)
Cox regression	0 (0)	5 (8)	5 (4)
Clinical expertise	0 (0)	2 (3)	2 (2)
NS	0 (0)	44 (69) ^D	44 (38)
Internal validation method			
Cross-validation	21 (40)	3 (5) ^G	24 (21)
Data splitting	11 (21)	0 (0)	11 (9)
Not done / NS	20 (38) ^F	61 (95)	81 (70)
Median (range) no. of final predictors^E	6 (3 – 23)	8 (2 – 15)	7 (2 – 23)
Study cohort			
Median (range) total sample size	3000 (27 – 1252313)	320 (15 – 31150)	686 (15 – 1252313)
Median (range) number of events	206 (8 – 86410)	51 (9 – 1350)	94 (8 – 86410)
Median (range) proportion of events (% of sample size)	10.65% (0.42% – 80.00%)	14.84% (1.18% – 46.67%)	14.40% (0.42% – 80.00%)

306 ^A the 28 included systematic reviews; ^B tools use multiple methods, therefore total number not equal to N (100%); ^C one
307 study also used discriminant analysis for model development; ^D many seemed to use clinical expertise, but development
308 methods were not clearly reported; ^E counting of final predictors may vary between models: some authors may count
309 individual factors, while others consider domains or subscales; ^F one review³⁶ implies 5 models did not implement internal
310 validation; ^G ‘resampling’ (not described further) was used for the development of 2 models; ML – machine learning; NS –
311 not stated; ICU – intensive care unit; PI – pressure injury.

312

313

Table 4. Predictor categories and frequency (%) of inclusion in N=53 models.

Predictor category	No. of tools predictor appears in
Mobility	27 (51)
Pre-disposing conditions	26 (49)
Continence	22 (42)
Receiving medical treatment/care	22 (42)
Age	21 (40)
Mental Status	20 (38)
Nutrition	20 (38)
Activity	19 (36)
Laboratory values	18 (34)
Skin	18 (34)
General Health	16 (30)
Body	13 (25)
Gender	11 (21)
Surgery duration	7 (13)
Ability to ambulate	6 (11)
Braden score	6 (11)
Pressure injury	6 (11)
Friction, shear, pressure	5 (9)
Medical unit, ward, visit	5 (9)
Length of stay	4 (8)
Hygiene	3 (6)
Ethnicity	2 (4)
Pain	2 (4)
Smoking	2 (4)
'Special' (not explained)	2 (4)
Isolation	1 (2)
Norton score	1 (2)

314 Figures are given as count (% out of 53 tools with information on predictors). Note that multiple predictors may fall within
 315 the same predictor category. For instance, the category 'skin' may encompass both 'skin moisture' and 'skin integrity', with
 316 the frequency count reflecting the entire predictor category rather than individual predictors.

317 DISCUSSION

318 This umbrella review summarises data from 28 eligible systematic reviews of PI risk prediction tools.
 319 Quality assessment using an adaptation of AMSTAR-2 revealed that most reviews were conducted to
 320 a relatively poor standard. Critical flaws were identified, including inadequate or absent reporting of
 321 protocols (23/28, 82%), inappropriate meta-analysis methods (11/15, 73%) and lack of consideration
 322 for risk of bias judgements when discussing review results (14/28, 50%). Despite the large number of
 323 risk prediction models identified, only five reviews focused on the development and validation of
 324 predominantly ML-based prediction models. The remaining reviews aimed to summarise evidence
 325 for the accuracy (sensitivity and specificity) or effectiveness of identified models. For many (44/64,
 326 69%) prediction tools that were developed without the use of ML, we were not able to determine
 327 whether reliable and robust statistical methods were used or whether models were essentially risk
 328 assessment tools developed based on expert knowledge. For over half (63/116, 54%) of the
 329 identified tools, predictors included in the final models were not reported. Details of study
 330 populations and settings were also lacking. It was not always clear from the reviews whether the
 331 poor reporting occurred at review level or in the original primary study publications.

332 A recent systematic review of risk of bias in machine learning developed prediction models found
 333 that most models are of poor methodological quality and are at high risk of bias.²³ In our set of
 334 reviews, of the three reviews that conducted a risk of bias assessment using the PROBAST tool, all
 335 models were found to be at high or unclear risk of bias.⁵²

336 Where the method of internal validation was reported, split-sample and cross-validation were the
337 most commonly used techniques, however, detail was limited, and it was not possible to determine
338 whether appropriate methods had been used. Although split-sample approaches have been favoured
339 for model validation, more recent empirical work suggests that bootstrap-based optimism
340 correction⁸³ or cross-validation⁸⁴ are preferred approaches. None of the included reviews reported
341 the use of optimism correction approaches.

342 Model development algorithms included logistic regression, decision trees and random forests, with
343 a vast number of ML-based models having been developed in the last five years. In contrast to
344 logistic regression approaches, decision trees and random forests while providing valuable risk
345 classification, may not give a proper risk probability. Instead, they commonly categorise patients into
346 binary 'at risk' or 'not at risk' groups. Although the risk probabilities generated in logistic regression
347 prediction models can be useful for clinical decision making, it was not possible to derive any
348 information about thresholds used to define 'at risk' or 'not at risk', and for most reviews, it was
349 unclear what the final model comprised of. This lack of transparency poses potential hurdles in
350 applying these models effectively in clinical settings.

351 Only one review included external validations of previously developed models⁵², however limited
352 details of model performance were presented. External validation is necessary to ensure a model is
353 both reproducible and generalisable^{85 86}, bringing the usefulness of the models included in these
354 reviews into question. The PROGRESS framework suggests that multiple external validation studies
355 should be conducted using independent datasets from different locations.¹⁵ In the only review that
356 included model validation studies⁵², it is unclear whether these studies were conducted in different
357 locations, and all were conducted in the same setting. PROGRESS also suggests that external
358 validations are carried out in a variety of relevant settings. Four out of eight validations were
359 described as using 'temporal' data, which suggests that the validation population is largely the same
360 as the development population but with use of data from different timeframes. This approach has
361 been described as 'between' internal and external validation, further emphasising the need for well-
362 designed external validation studies.⁸⁵ Furthermore, none of the eight external validations reported
363 model recalibration. Recent evidence suggests focus should be placed on large, well-designed
364 external validation studies to validate and improve promising models (using recalibration and
365 updating⁸⁷), rather than developing a multitude of new ones.^{18 15} Model validation and recalibration
366 should be a continuous process, and this is something that future research should address. Following
367 external validation, the PROGRESS framework¹⁵ suggests that effectiveness studies should be
368 conducted following external validation to assess the effect of the model on decision making, patient
369 outcomes and costs.

370 Despite the advances in methods for developing risk prediction models, scales developed using
371 clinical expertise such as the Braden Scale, Norton Scale, Waterlow Score and Cubbin-Jackson Scale
372 are extensively discussed in numerous clinical practice guidelines for patient risk assessment, and are
373 commonly used in clinical practice.^{6 88} Although guidelines recognise their low accuracy, they are still
374 acknowledged, while other risk prediction models are not even considered. This may be due to the
375 availability of at least some clinical trials evaluating the clinical utility of scales.³⁹ Some scales, such as
376 the Braden scale, are so widely used that they have become an integral component of risk
377 assessment for PI in clinical practice, and have even been incorporated into EHRs. Their widespread
378 use may impede the progress towards development, validation and evaluation of more accurate and
379 innovative risk prediction models. Striking a balance between tradition and embracing advancements
380 is crucial for effective implementation in healthcare settings and improving patient outcomes.

381 Other existing evidence

382 We are aware of one additional systematic review of ML prediction models for PIs published after
383 our search was conducted.⁸⁹ Pei and colleagues included 18 models, all of which were already
384 included in our list of identified models. The aim of the review was to assess risk factors related to
385 HAPIs, rather than assess tools to predict PIs and only ML-based models were included. A meta-
386 analysis was conducted by pooling prognostic accuracy measures across all models that provided 2x2
387 data (n=14 models). The pooled AUC across the 14 models was 0.94, pooled sensitivity was 0.79
388 (95% CI 0.78–0.80) and pooled specificity was 0.87 (95% CI 0.88–0.87).⁸⁹ A meta-regression was
389 conducted to investigate whether these values differed by machine learning algorithm; no difference
390 based on algorithm was found. Review authors found 16/18 studies at high risk of bias based on
391 PROBAST. We had low confidence in the review itself, through applying our adapted AMSTAR-2
392 criteria; one critical flaw was their use of inappropriate meta-analysis methods (not using a
393 hierarchical model for pooling sensitivity and specificity).

394 Strengths and limitations

395 Our umbrella reviews were conducted to a high standard, following Cochrane guidance⁴⁰, and with a
396 highly sensitive search strategy designed by an experienced information specialist. Although we
397 excluded non-English publications due to time and resource constraints, where possible these
398 publications were used to identify additional eligible risk prediction models. To some extent our
399 review is limited by the use of AMSTAR-2 for quality assessment of included reviews. AMSTAR-2 was
400 not designed for assessment of diagnostic or prognostic studies and, although we made some
401 adaptations, many of the existing and amended criteria relate to the quality of reporting of the
402 reviews as opposed to methodological quality. There is scope for further work to establish criteria for
403 assessing systematic reviews of prediction models.

404 The main limitation, however, was the lack of detail about risk prediction models and risk prediction
405 model performance that could be determined from the included systematic reviews. To be as
406 comprehensive as possible in model identification, we were relatively generous in our definition of
407 'systematic', and this may have contributed to the often poor level of detail provided by included
408 reviews. It is likely, however, that reporting was poor in many of the primary studies contributing to
409 these reviews. Excluding the ML-based models, more than half of available risk prediction scales or
410 tools were published prior to the year 2000. The fact that the original versions of reporting
411 guidelines for diagnostic accuracy studies⁹⁰ and risk prediction models⁹¹ were not published until
412 2003 and 2015 respectively, is likely to have contributed to poor reporting. In contrast, the ML-based
413 models were published between 2000 and 2022, with a median year of 2020. Reporting guidelines
414 for development and validation of ML-based models are more recent⁹² or still in development⁹³, but
415 aim to improve the reporting standards and understanding of evolving ML technologies in
416 healthcare.

417 CONCLUSIONS

418 There is a very large body of evidence reporting various risk prediction scales, tool and models for PI
419 which has been summarised across multiple systematic reviews of varying methodological quality.
420 Only five systematic reviews reported the development and validation of models to predict risk of
421 PIs. It seems that for the most part, available models do not meet current standards for the
422 development or reporting of risk prediction models. Furthermore, most available models, including
423 ML-based models have not been validated beyond the original population in which they were
424 developed. Identification of the optimal risk prediction model for PI from those currently available
425 would require a high-quality systematic review of the primary literature, ideally limited to studies

426 conducted to a high methodological standard. It is evident from our findings that there is still a lack
427 of consensus on the optimal risk prediction model for PI, highlighting the need for more standardised
428 and rigorous approaches in future research.

429

430 Acknowledgements

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

433 Author Contributions

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

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

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

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

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

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

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

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

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

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

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

448 Funding

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

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

457 Conflicting Interests

458 The authors of this manuscript have the following competing interests: VV is an employee of Paul
459 Hartmann AG; ES and THB received consultancy fees from Paul Hartmann AG. All other authors
460 received no personal funding or personal compensation from Paul Hartmann AG and have declared
461 that no competing interests exist.

References

- 462 1. Li Z, Lin F, Thalib L, et al. Global prevalence and incidence of pressure injuries in hospitalised adult
463 patients: A systematic review and meta-analysis. *International Journal of Nursing Studies*
464 2020;105:103-546. doi: 10.1016/j.ijnurstu.2020.103546
- 465 2. Padula WV, Delarmente BA. The national cost of hospital-acquired pressure injuries in the United
466 States. *Int Wound J* 2019;16(3):634-40. doi: 10.1111/iwj.13071 [published Online First:
467 2019/01/28]
- 468 3. Theisen S, Drabik A, Stock S. Pressure ulcers in older hospitalised patients and its impact on length
469 of stay: a retrospective observational study. *J Clin Nurs* 2012;21(3-4):380-7. doi:
470 10.1111/j.1365-2702.2011.03915.x [published Online First: 2011/12/09]
- 471 4. Sullivan N, Schoelles K. Preventing In-Facility Pressure Ulcers as a Patient Safety Strategy. *Annals of*
472 *Internal Medicine* 2013;158(5.2):410-16. doi: 10.7326/0003-4819-158-5-201303051-00008
- 473 5. Institute for Quality and Efficiency in Health Care (IQWiG). Preventing pressure ulcers. Cologne,
474 Germany 2006 [updated 2018 Nov 15. Available from:
475 <https://www.ncbi.nlm.nih.gov/books/NBK326430/?report=classic> accessed Feb 2023].
- 476 6. Haesler E. European Pressure Ulcer Advisory Panel, National Pressure Injury Advisory Panel and
477 Pan Pacific Pressure Injury Alliance. Prevention and Treatment of Pressure Ulcers/Injuries:
478 Clinical Practice Guideline. 2019 [Available from: <https://internationalguideline.com/2019>
479 accessed Feb 2023].
- 480 7. Walker RM, Gillespie BM, McInnes E, et al. Prevention and treatment of pressure injuries: A meta-
481 synthesis of Cochrane Reviews. *Journal of Tissue Viability* 2020;29(4):227-43. doi:
482 10.1016/j.jtv.2020.05.004
- 483 8. Shi C, Dumville JC, Cullum N, et al. Beds, overlays and mattresses for preventing and treating
484 pressure ulcers: an overview of Cochrane Reviews and network meta-analysis. *Cochrane*
485 *Database Syst Rev* 2021;8(8):Cd013761. doi: 10.1002/14651858.CD013761.pub2 [published
486 Online First: 2021/08/16]
- 487 9. Russo CA, Steiner C, Spector W. Hospitalizations Related to Pressure Ulcers, 2006. HCUP Statistical
488 Brief: Agency for Healthcare Research and Quality, Rockville, MD. 2008.
- 489 10. Braden B, Bergstrom N. A Conceptual Schema for the Study of the Etiology of Pressure Sores.
490 *Rehabilitation Nursing* 1987;12(1):8-16. doi: 10.1002/j.2048-7940.1987.tb00541.x
- 491 11. Bergstrom N, Braden BJ, Laguzza A, et al. The Braden Scale for Predicting Pressure Sore Risk. *Nurs*
492 *Res* 1987;36(4):205-10.
- 493 12. Norton D. Geriatric nursing problems. *Int Nurs Rev* 1962;9:39-41.
- 494 13. Waterlow J. Pressure sores: a risk assessment card. *Nursing Times* 1985;81:49-55.
- 495 14. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and
496 external validation. *J Clin Epidemiol* 2016;69:245-7. doi: 10.1016/j.jclinepi.2015.04.005
497 [published Online First: 2015/04/18]
- 498 15. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3:
499 Prognostic Model Research. *PLOS Medicine* 2013;10(2):e1001381. doi:
500 10.1371/journal.pmed.1001381
- 501 16. Siontis GCM, Tzoulaki I, Castaldi PJ, et al. External validation of new risk prediction models is
502 infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*
503 2015;68(1):25-34. doi: 10.1016/j.jclinepi.2014.09.007
- 504 17. Bouwmeester W, Zuithoff NPA, Mallett S, et al. Reporting and Methods in Clinical Prediction
505 Research: A Systematic Review. *PLOS Medicine* 2012;9(5):e1001221. doi:
506 10.1371/journal.pmed.1001221
- 507 18. Van Calster B, Steyerberg EW, Wynants L, et al. There is no such thing as a validated prediction
508 model. *BMC Medicine* 2023;21(1):70. doi: 10.1186/s12916-023-02779-w
- 509 19. Wynants L, Calster BV, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-
510 19: systematic review and critical appraisal. *BMJ* 2020;369:m1328. doi: 10.1136/bmj.m1328

- 511 20. Ma J, Dhiman P, Qi C, et al. Poor handling of continuous predictors in clinical prediction models
512 using logistic regression: a systematic review. *J Clin Epidemiol* 2023;161:140-51. doi:
513 10.1016/j.jclinepi.2023.07.017 [published Online First: 2023/08/02]
- 514 21. Dhiman P, Ma J, Qi C, et al. Sample size requirements are not being considered in studies
515 developing prediction models for binary outcomes: a systematic review. *BMC Medical*
516 *Research Methodology* 2023;23(1):188. doi: 10.1186/s12874-023-02008-1
- 517 22. Moriarty AS, Meader N, Snell KIE, et al. Predicting relapse or recurrence of depression: systematic
518 review of prognostic models. *Br J Psychiatry* 2022;221(2):448-58. doi: 10.1192/bjp.2021.218
- 519 23. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models
520 developed using supervised machine learning techniques: systematic review. *BMJ*
521 2021;375:n2281. doi: 10.1136/bmj.n2281
- 522 24. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of
523 machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*
524 2019;110:12-22. doi: 10.1016/j.jclinepi.2019.02.004 [published Online First: 20190211]
- 525 25. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of
526 prediction model performance. *BMJ* 2017;356:i6460. doi: 10.1136/bmj.i6460
- 527 26. Riley RD, van der Windt D, Croft P, et al. Prognosis research in healthcare: concepts, methods, and
528 impact: Oxford University Press 2019.
- 529 27. Snell KIE, Levis B, Damen JAA, et al. Transparent reporting of multivariable prediction models for
530 individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses
531 (TRIPOD-SRMA). *BMJ* 2023;381:e073538. doi: 10.1136/bmj-2022-073538
- 532 28. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability
533 of Prediction Model Studies. *Annals of Internal Medicine* 2019;170(1):51-58. doi:
534 10.7326/M18-1376
- 535 29. Chen HL, Shen WQ, Liu P. A Meta-analysis to Evaluate the Predictive Validity of the Braden Scale
536 for Pressure Ulcer Risk Assessment in Long-term Care. *Ostomy/wound management*
537 2016;62(9):20-8.
- 538 30. Baris N, Karabacak BG, Alpar SE. The Use of the Braden Scale in Assessing Pressure Ulcers in
539 Turkey: A Systematic Review. *Advances in skin & wound care* 2015;28:349-57. doi:
540 10.1097/01.ASW.0000465299.99194.e6
- 541 31. He W, Liu P, Chen HL. The Braden Scale cannot be used alone for assessing pressure ulcer risk in
542 surgical patients: a meta-analysis. *Ostomy/wound management* 2012;58:34-40.
- 543 32. Huang C, Ma Y, Wang C, et al. Predictive validity of the braden scale for pressure injury risk
544 assessment in adults: A systematic review and meta-analysis. *Nursing open* 2021;8:2194-207.
545 doi: 10.1002/nop2.792
- 546 33. Park SH, Choi YK, Kang CB. Predictive validity of the Braden Scale for pressure ulcer risk in
547 hospitalized patients. *Journal of Tissue Viability* 2015;24:102-13. doi:
548 10.1016/j.jtv.2015.05.001
- 549 34. Wei M, Wu L, Chen Y, et al. Predictive Validity of the Braden Scale for Pressure Ulcer Risk in
550 Critical Care: A Meta-Analysis. *Nursing in critical care* 2020;25:165-70. doi:
551 10.1111/nicc.12500
- 552 35. Wilchesky M, Lungu O. Predictive and concurrent validity of the Braden scale in long-term care: A
553 meta-analysis. *Wound Repair and Regeneration* 2015;23:44-56. doi: 10.1111/wrr.12261
- 554 36. Dweekat OY, Lam SS, McGrath L. Machine Learning Techniques, Applications, and Potential Future
555 Opportunities in Pressure Injuries (Bedsore) Management: A Systematic Review.
556 *International journal of environmental research and public health* 2023;20(1) doi:
557 10.3390/ijerph20010796
- 558 37. Jiang M, Ma Y, Guo S, et al. Using Machine Learning Technologies in Pressure Injury Management:
559 Systematic Review. *JMIR Medical Informatics* 2021;9(3):e25704. doi: 10.2196/25704

- 560 38. Qu C, Luo W, Zeng Z, et al. The predictive effect of different machine learning algorithms for
561 pressure injuries in hospitalized patients: A network meta-analyses. *Heliyon*
562 2022;8(11):e11361. doi: 10.1016/j.heliyon.2022.e11361
- 563 39. Hillier B, Scandrett K, Coombe A, et al. Accuracy and clinical effectiveness of risk prediction tools
564 for pressure injury occurrence: An umbrella review (pre-print). *MedRxiv* 2024 doi: TBC
565 (UPDATE ME)
- 566 40. Pollock M, Fernandes RM BL, Pieper D, Hartling L,. Chapter V: Overviews of Reviews. In: Higgins
567 JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA ed. *Cochrane Handbook for*
568 *Systematic Reviews of Interventions* version 63 (updated February 2022). Available from
569 www.training.cochrane.org/handbook: Cochrane 2022.
- 570 41. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-
571 Analyses: The PRISMA Statement. *PLOS Medicine* 2009;6(7):e1000097. doi:
572 10.1371/journal.pmed.1000097
- 573 42. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. *J Am Med Inform Assoc*
574 2001;8(4):391-7. doi: 10.1136/jamia.2001.0080391 [published Online First: 2001/06/22]
- 575 43. Wilczynski NL, Haynes RB. Optimal Search Strategies for Detecting Clinically Sound Prognostic
576 Studies in EMBASE: An Analytic Survey. *Journal of the American Medical Informatics*
577 *Association* 2005;12(4):481-85. doi: 10.1197/jamia.M1752
- 578 44. Geersing G-J, Bouwmeester W, Zuithoff P, et al. Search Filters for Finding Prognostic and
579 Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. *PLOS ONE*
580 2012;7(2):e32844. doi: 10.1371/journal.pone.0032844
- 581 45. NHS. Pressure ulcers: revised definition and measurement. Summary and recommendations 2018
582 [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)
583 [summary-recommendations.pdf](https://www.england.nhs.uk/wp-content/uploads/2021/09/NSTPP-summary-recommendations.pdf) accessed Feb 2023].
- 584 46. AHCPR. Pressure ulcer treatment. : Agency for Health Care Policy and Research 1994:1-25.
- 585 47. Harker J. Pressure ulcer classification: the Torrance system. *Journal of Wound Care* 2000;9(6):275-
586 77. doi: 10.12968/jowc.2000.9.6.26233
- 587 48. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical Appraisal and Data Extraction for
588 Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLOS Medicine*
589 2014;11(10):e1001744. doi: 10.1371/journal.pmed.1001744
- 590 49. Cochrane. DE form example prognostic models - scoping review: The Cochrane Collaboration: The
591 Prognosis Methods Group; [Available from: <https://methods.cochrane.org/prognosis/tools>
592 accessed Feb 2023].
- 593 50. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that
594 include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*
595 2017;358:j4008. doi: 10.1136/bmj.j4008
- 596 51. Ribeiro F, Fidalgo F, Silva A, et al. Literature review of machine-learning algorithms for pressure
597 ulcer prevention: Challenges and opportunities: MDPI 2021.
- 598 52. Shi C, Dumville JC, Cullum N. Evaluating the development and validation of empirically-derived
599 prognostic models for pressure ulcer risk assessment: A systematic review. *International*
600 *journal of nursing studies* 2019;89:88-103. doi: 10.1016/j.ijnurstu.2018.08.005
- 601 53. Zhou Y, Yang X, Ma S, et al. A systematic review of predictive models for hospital-acquired
602 pressure injury using machine learning. *Nursing open* 2022;30 doi: 10.1002/nop2.1429
- 603 54. Chou R, Dana T, Bougatsos C, et al. Pressure ulcer risk assessment and prevention: a systematic
604 comparative effectiveness review. *Annals of internal medicine* 2013;159(1):28-38.
- 605 55. García-Fernández FP, Pancorbo-Hidalgo PL, Agreda JJS. Predictive Capacity of Risk Assessment
606 Scales and Clinical Judgment for Pressure Ulcers: A Meta-analysis. *Journal of Wound Ostomy*
607 *& Continence Nursing* 2014;41(1):24-34. doi: 10.1097/01.WON.0000438014.90734.a2
- 608 56. Pancorbo-Hidalgo PL, Garcia-Fernandez FP, Lopez-Medina IM, et al. Risk assessment scales for
609 pressure ulcer prevention: a systematic review. *J Adv Nurs* 2006;54(1):94-110. doi:
610 10.1111/j.1365-2648.2006.03794.x

- 611 57. Park SH, Lee HS. Assessing Predictive Validity of Pressure Ulcer Risk Scales- A Systematic Review
612 and Meta-Analysis. *Iranian journal of public health* 2016;45(2):122-33.
- 613 58. Park SH, Lee YS, Kwon YM. Predictive Validity of Pressure Ulcer Risk Assessment Tools for Elderly:
614 A Meta-Analysis. *Western journal of nursing research* 2016;38:459-83. doi:
615 10.1177/0193945915602259
- 616 59. Tayyib NAH, Coyer F, Lewis P. Pressure ulcers in the adult intensive care unit: a literature review of
617 patient risk factors and risk assessment scales. *Journal of Nursing Education and Practice*
618 2013;3(11):28-42.
- 619 60. Wang N, Lv L, Yan F, et al. Biomarkers for the early detection of pressure injury: A systematic
620 review and meta-analysis. *Journal of Tissue Viability* 2022;31:259-67. doi:
621 10.1016/j.jtv.2022.02.005
- 622 61. Zhang Y, Zhuang Y, Shen J, et al. Value of pressure injury assessment scales for patients in the
623 intensive care unit: Systematic review and diagnostic test accuracy meta-analysis. *Intensive &*
624 *critical care nursing* 2021;64:103009. doi: 10.1016/j.iccn.2020.103009
- 625 62. Zimmermann GS, Cremasco MF, Zanei SSV, et al. Pressure injury risk prediction in critical care
626 patients: an integrative review. *Texto & Contexto-Enfermagem* 2018;27(3)
- 627 63. Gaspar S, Peralta M, Marques A, et al. Effectiveness on hospital-acquired pressure ulcers
628 prevention: a systematic review. *International Wound Journal* 2019;16(5):1087-102. doi:
629 10.1111/iwj.13147
- 630 64. Ontario HQ. Pressure ulcer prevention: an evidence-based analysis. *Ontario health technology*
631 *assessment series* 2009;9(2):1-104.
- 632 65. Kottner J, Dassen T, Tannen A. Inter- and intrarater reliability of the Waterlow pressure sore risk
633 scale: A systematic review. *International Journal of Nursing Studies* 2009;46:369-79. doi:
634 10.1016/j.ijnurstu.2008.09.010
- 635 66. Lovegrove J, Ven S, Miles SJ, et al. Comparison of pressure injury risk assessment outcomes using
636 a structured assessment tool versus clinical judgement: A systematic review. *Journal of*
637 *Clinical Nursing* 2021 doi: 10.1111/jocn.16154 [published Online First: 2021/12/01]
- 638 67. Lovegrove J, Miles S, Fulbrook P. The relationship between pressure ulcer risk assessment and
639 preventative interventions: a systematic review. *Journal of wound care* 2018;27(12):862-75.
- 640 68. Moore ZEH, Patton D. Risk assessment tools for the prevention of pressure ulcers. *Cochrane*
641 *Database of Systematic Reviews* 2019 doi: 10.1002/14651858.CD006471.pub4
- 642 69. Page KN, Barker AL, Kamar J. Development and validation of a pressure ulcer risk assessment tool
643 for acute hospital patients. *Wound Repair and Regeneration* 2011;19(1):31-37. doi:
644 10.1111/j.1524-475X.2010.00647.x
- 645 70. Berlowitz DR, Ash AS, Brandeis GH, et al. Rating long-term care facilities on pressure ulcer
646 development: Importance of case-mix adjustment. *Annals of Internal Medicine*
647 1996;124(6):557-63. doi: 10.7326/0003-4819-124-6-199603150-00003
- 648 71. Berlowitz DR, Brandeis GH, Morris JN, et al. Deriving a risk-adjustment model for pressure ulcer
649 development using the Minimum Data Set. *Journal of the American Geriatrics Society*
650 2001;49(7):866-71. doi: 10.1046/j.1532-5415.2001.49175.x
- 651 72. Poss J, Murphy KM, Woodbury MG, et al. Development of the interRAI Pressure Ulcer Risk Scale
652 (PURS) for use in long-term care and home care settings. *BMC geriatrics* 2010;10:67. doi:
653 10.1186/1471-2318-10-67
- 654 73. Compton F, Hoffmann F, Hortig T, et al. Pressure ulcer predictors in ICU patients: nursing skin
655 assessment versus objective parameters. *J Wound Care* 2008;17(10):417-20, 22-4. doi:
656 10.12968/jowc.2008.17.10.31304
- 657 74. Lowery MT. A pressure sore risk calculator for intensive care patients: 'the Sunderland
658 experience'. *Intensive Crit Care Nurs* 1995;11(6):344-53. doi: 10.1016/s0964-3397(95)80452-
659 8
- 660 75. Baldwin KM, Ziegler SM. Pressure ulcer risk following critical traumatic injury. *Advances in wound*
661 *care : the journal for prevention and healing* 1998;11(4):168-73.

- 662 76. Bergquist S. Subscales, subscores, or summative score: evaluating the contribution of Braden
663 Scale items for predicting pressure ulcer risk in older adults receiving home health care.
664 *Journal of wound, ostomy, and continence nursing : official publication of The Wound,*
665 *Ostomy and Continence Nurses Society* 2001;28(6):279-89. doi: 10.1067/mjw.2001.119012
- 666 77. Song M, Choi KS. Factors predicting development of decubitus ulcers among patients admitted
667 for neurological problems. *The Journal of Nurses Academic Society* 1991;21(1):16-26.
- 668 78. Halfens R, Van Achterberg T, Bal R. Validity and reliability of the Braden scale and the influence of
669 other risk factors: a multi-centre prospective study. *International Journal of Nursing Studies*
670 2000;37(4):313-19.
- 671 79. Choi KS, Song MS. Test of predictive validity for the new pressure risk assessment scale. *Journal of*
672 *Korean Academy of Adult Nursing* 1991;3(1):19-28.
- 673 80. Kwong E, Pang S, Wong T, et al. Predicting pressure ulcer risk with the modified Braden, Braden,
674 and Norton scales in acute care hospitals in Mainland China. *Appl Nurs Res* 2005;18(2):122-8.
675 doi: 10.1016/j.apnr.2005.01.001
- 676 81. Pang SM, Wong TK. Predicting pressure sore risk with the Norton, Braden, and Waterlow scales in
677 a Hong Kong rehabilitation hospital. *Nursing Research* 1998;47(3):147-53.
- 678 82. Schue RM, Langemo DK. Pressure ulcer prevalence and incidence and a modification of the
679 Braden Scale for a rehabilitation unit. *Journal of wound, ostomy, and continence nursing :*
680 *official publication of The Wound, Ostomy and Continence Nurses Society* 1998;25(1):36-43.
681 doi: 10.1016/s1071-5754(98)90011-0
- 682 83. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different
683 strategies for estimating the out-of-sample validity of logistic regression models. *Stat*
684 *Methods Med Res* 2017;26(2):796-808. doi: 10.1177/0962280214558972 [published Online
685 First: 2014/11/19]
- 686 84. Smith GC, Seaman SR, Wood AM, et al. Correcting for optimistic prediction in small data sets. *Am*
687 *J Epidemiol* 2014;180(3):318-24. doi: 10.1093/aje/kwu140 [published Online First:
688 2014/06/24]
- 689 85. Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: what, why, how,
690 when and where? *Clin Kidney J* 2021;14(1):49-58. doi: 10.1093/ckj/sfaa188 [published Online
691 First: 2020/11/24]
- 692 86. de Hond AAH, Shah VB, Kant IMJ, et al. Perspectives on validation of clinical predictive
693 algorithms. *npj Digital Medicine* 2023;6(1):86. doi: 10.1038/s41746-023-00832-9
- 694 87. Binuya MAE, Engelhardt EG, Schats W, et al. Methodological guidance for the evaluation and
695 updating of clinical prediction models: a systematic review. *BMC Med Res Methodol*
696 2022;22(1):316. doi: 10.1186/s12874-022-01801-8 [published Online First: 2022/12/12]
- 697 88. Qaseem A, Mir TP, Starkey M, et al. Risk Assessment and Prevention of Pressure Ulcers: A Clinical
698 Practice Guideline From the American College of Physicians. *Annals of Internal Medicine*
699 2015;162(5):359-69. doi: 10.7326/m14-1567
- 700 89. Pei J, Guo X, Tao H, et al. Machine learning-based prediction models for pressure injury: A
701 systematic review and meta-analysis. *Int Wound J* 2023 doi: 10.1111/iwj.14280 [published
702 Online First: 2023/06/20]
- 703 90. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for
704 reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527. doi: 10.1136/bmj.h5527
705 [published Online First: 2015/10/28]
- 706 91. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction
707 model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann*
708 *Intern Med* 2015;162(1):W1-73. doi: 10.7326/m14-0698
- 709 92. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, et al. MINIMAR (MINimum Information for
710 Medical AI Reporting): Developing reporting standards for artificial intelligence in health
711 care. *Journal of the American Medical Informatics Association* 2020;27(12):2011-15. doi:
712 10.1093/jamia/ocaa088

713 93. Collins G, Dhiman P, Logullo P, et al. TRIPOD+AI. OSF, 2023.