Factors influencing clinician and patient interaction with machine learning-based risk prediction models: a systematic review

Rebecca Giddings, Anabel Joseph, Thomas Callender, Sam M Janes, Mihaela van der Schaar, Jessica Sheringham, Neal Navani

Machine learning (ML)-based risk prediction models hold the potential to support the health-care setting in several ways; however, use of such models is scarce. We aimed to review health-care professional (HCP) and patient perceptions of ML risk prediction models in published literature, to inform future risk prediction model development. Following database and citation searches, we identified 41 articles suitable for inclusion. Article quality varied with qualitative studies performing strongest. Overall, perceptions of ML risk prediction models were positive. HCPs and patients considered that models have the potential to add benefit in the health-care setting. However, reservations remain; for example, concerns regarding data quality for model development and fears of unintended consequences following ML model use. We identified that public views regarding these models might be more negative than HCPs and that concerns (eg, extra demands on workload) were not always borne out in practice. Conclusions are tempered by the low number of patient and public studies, the absence of participant ethnic diversity, and variation in article quality. We identified gaps in knowledge (particularly views from under-represented groups) and optimum methods for model explanation and alerts, which require future research.

Lancet Digit Health 2024; 6: e131–44

Lungs for Living Research Centre, UCL Respiratory (R Giddings MFPH, A Joseph BSc, T Callender MFPH, Prof S M Janes PhD, N Navani PhD) **and Department of Applied Health Research** (J Sheringham FFPH)**, University College London, London, UK; Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK** (Prof M van der Schaar PhD)**; The Alan Turing Institute, London, UK** (Prof M van der Schaar)

Correspondence to:
Rebecca Giddings, Lungs for Living Research Centre, UCL Respiratory, University College London, London WC1E 6JF, UK
r.giddings@ucl.ac.uk

Introduction

Machine learning (ML)-based risk prediction models are increasingly being developed across all areas of medical care.[1] Such models can support the health-care setting in several ways, including pattern recognition in digital imaging[2] and aiding decision making for patients and clinicians. Given the current emphasis on personalised medicine, there is a growing need for increasingly complex risk prediction tools, illustrated by ongoing research developing polygenic risk scores via ML methods.[3] Despite their potential however, use of such ML models is restricted within health-care settings.[4]

ML has the potential to enhance the performance of risk prediction models; however, there are challenges associated with it. For instance, one article found a high risk of bias in oncology-related ML-based risk models, contraindicating use in clinical practice, and highlighted the necessity for sufficient data in model development.[5] Another challenge is that many types of ML models are not readily interpretable, which might lead clinicians to use less accurate standard models rather than more accurate ML models that they do not trust.[6]

Given the current concerns, it is important to understand patient and health-care staff perceptions of ML-based predictive models, which might help explain the gap between model potential and use. To our knowledge, no previous reviews have synthesised the existing literature on this topic. Reviews have considered perceptions of decision aids,[7–10] clinician perspective of risk prediction tools not specific to ML,[11] perspectives of artificial intelligence (AI) in general,[12–14] and predictive analytics using electronic health record (EHR) data only.[15,16] Our aim is to better understand health-care staff and patient perceptions of ML risk prediction models to inform the development and deployment of future risk prediction models. Our objectives in the study were to: (1) understand perceptions of ML-based risk prediction models, (2) understand user recommendations for model improvement, and (3) identify gaps in current research.

Methods

This Review of published literature was registered with PROSPERO (CRD42022330042) and adhered to PRISMA systematic review guidance.[17]

Search strategy and selection criteria

We aimed to identify articles assessing perceptions of ML risk prediction models in the health-care setting. Inclusion and exclusion criteria were documented with the PICO framework (panel). We searched MEDLINE, Embase, Web of Science, and Scopus on June 13, 2022, with search terms developed by considering previous reviews[9,12,18–20] and permutations of the following: "machine learning", "prediction", and "user perception" (appendix pp 2–8). Following duplicate removal, RG performed article screening against inclusion and exclusion criteria (panel) by sequentially assessing the title, abstract, and full text. To supplement the database searches, for included articles forward and backward citation searches (forward citation search screens all articles that cite the included article, and backward search screens articles referenced in the included article itself) were performed, along with screening similar articles (using word-weighted algorithms) identified by PubMed or Web of Science. We repeated the database search on July 21, 2023 to update publications.

See **Online** for appendix

Data analysis

Data were extracted by RG—from main text and supplementary material—on study design, study population, and model details (appendix pp 9–10). Due to

***Panel:* PICO framework inclusion and exclusion criteria**

Population

- Inclusion criteria: staff working in a health-care setting, patients, members of the public, and parents and relatives
- Exclusion criteria: not meeting inclusion criteria

Intervention

- Inclusion criteria: machine learning (ML) risk prediction models assessing risk per individual of disease or health condition; we considered any modelling approach except standard linear models (eg, linear, logistic, and Cox's proportional hazards regression) as ML; models must include clinicodemographic data as input variables; and studies must consider either developed ML models, hypothetical ML models (ie, providing model screenshots, stills, or exerts without access to an underlying model), clinical decision support systems based on eligible risk prediction models, or articles garnering input before and following ML risk prediction model development
- Exclusion criteria: we excluded models diagnosing a disease or condition rather than predicting risk due to focus on risk predictions, models aiding administrators in health-care resource and staff allocations, models predicting epidemiological outcomes, and cost-based models identifying high-cost patients

Comparison

- Inclusion criteria: not applicable
- Exclusion criteria: not applicable

Outcome

- Inclusion criteria: user perceptions of models
- Exclusion criteria: evaluation of the prognostic or predictive value of different factors; determining intervention effectiveness, optimal treatment dosages, in vitro outcomes, or cost-effectiveness; and model development

Date

- Inclusion criteria: published in the past 10 years
- Exclusion criteria: not applicable

Location

- Inclusion criteria: no restrictions
- Exclusion criteria: no restrictions

Language

- Inclusion criteria: English
- Exclusion criteria: none

Study type

- Inclusion criteria: mixed methods, qualitative, and quantitative studies
- Exclusion criteria: not applicable

the diversity of study design and outcomes, a narrative synthesis was undertaken. Following author familiarisation with published literature,[12,21] we took a deductive approach and identified a priori themes and applied as predetermined codes, with codes further refined during analysis. During coding, RG assessed whether an article presented, per theme, predominantly positive, negative, or mixed views as expressed by respondents (or alternatively general thoughts).

In instances where additional technologies were assessed in a study, article results were considered unless clearly irrelevant to ML risk prediction models. Due to study variation in providing respondents with tool access and differing ML experience levels of the responders, article findings varied between the respondents' general perceptions and thoughts versus details from lived experiences. Unless relevant, no distinction was made between these viewpoints during coding.

We adapted two widely used quality assessment tools to form an appraisal tool for included studies: the mixed methods appraisal tool (MMAT)[22] and Critical Appraisal Skill Program (CASP).[23] MMAT was appropriate given its relevance to mixed method articles;[24] however, the criteria poorly differentiated the included qualitative studies, and some questions were deemed unsuitable by the first author RG (eg, assessing patient intervention adherence). Integration with relevant CASP criteria was therefore performed, with efforts to facilitate consistency across study types, resulting in five criteria per study type and five additional questions for mixed methods studies maintained as per MMAT (appendix pp 11–20).

Quality appraisal was undertaken concurrently by two reviewers (RG and AJ) following an initial screening of three articles and discussion of the ratings for calibration. Following assessment against recommended indicators, studies were assigned a high, medium, or low rating for each criterion by each reviewer. Inter-rater reliability was assessed with Cohen's Kappa. Disagreements in ratings were discussed by the two reviewers until a consensus was agreed.

Results

After removing duplicates, 3515 articles were identified from database searches performed on June 13, 2022, and screened. Articles that were excluded following title review (n=2001) and abstract review (n=941) generally only identified prognostic or predictive factors or did not consider a risk prediction model. Of the remaining 562 articles, nine met inclusion criteria and the main reason for exclusion was because of not meeting outcome inclusion criteria (see appendix pp 21–109 for table of exclusion reasons for full texts). Forward and backward citation search, along with reviewing similar articles identified another 28 articles for inclusion, while additional database searches performed on July 21, 2023 identified an additional four articles for inclusion (figure 1).[25]

Most included articles had a quantitative survey (26 of 41 [63%]) or an interview of participants (19 [46%] of 41; table 1; appendix pp 109–36). The most frequently considered specialty or settings were emergency departments (nine [22%] of 41).[27,34,37,44,47,49,56,57,62] Health conditions commonly considered were sepsis[40,43,48,56] (four [10%] of 41), pneumonia[32,34,47,62] (four, 10%), and suicide[29,30,64,65] (four, 10%). Most articles analysed ML risk prediction models alone (23 [56%] of 41); five (12%) analysed risk prediction models in general[27,29,43,48,64] (ML and non-ML models); and five (12%)[31,50,54,55,59] analysed AI in general (risk prediction

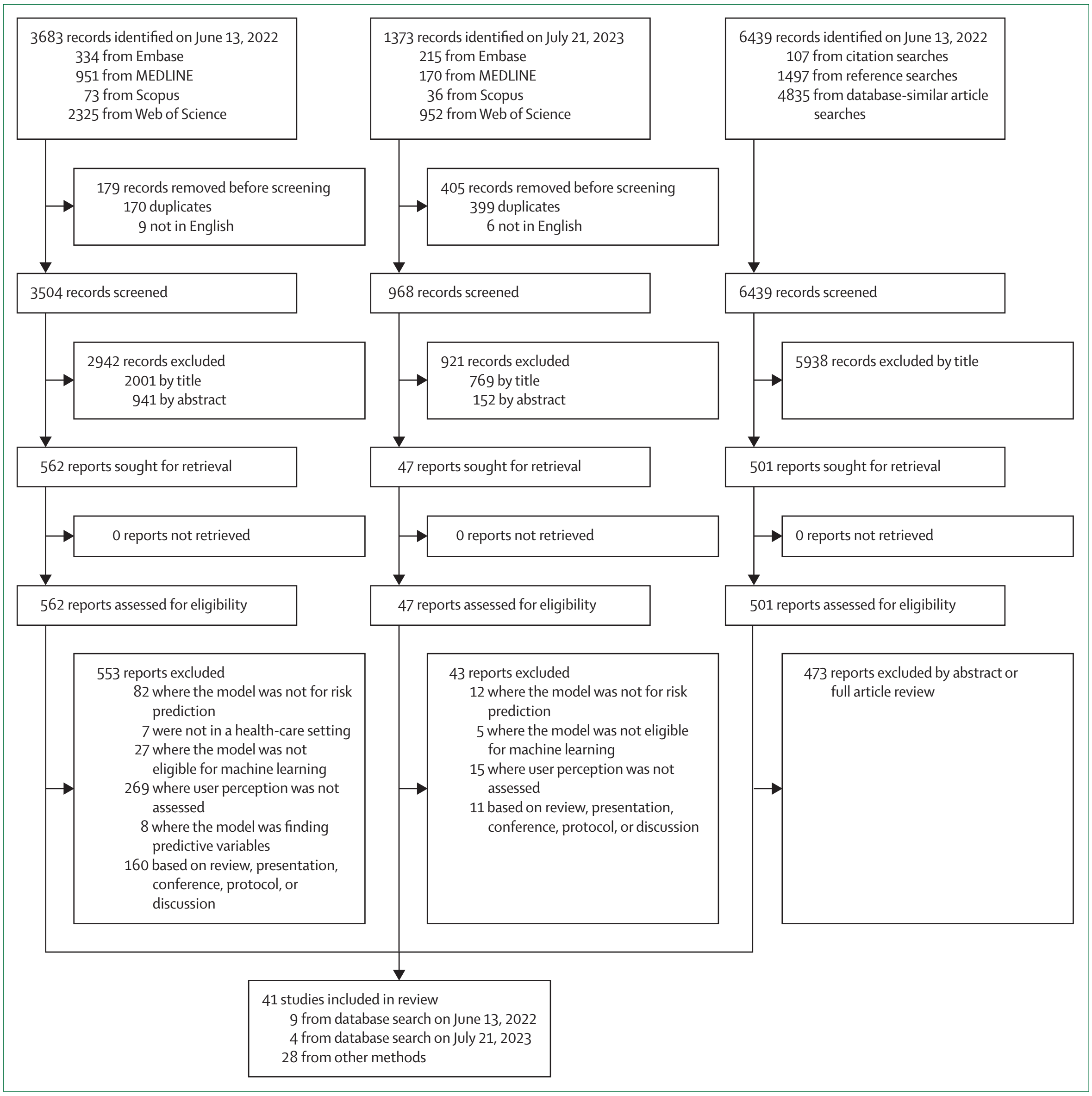
3683 records identified on June 13, 2022
334 from Embase
951 from MEDLINE
73 from Scopus
2325 from Web of Science
1373 records identified on July 21, 2023
215 from Embase
170 from MEDLINE
36 from Scopus
952 from Web of Science
6439 records identified on June 13, 2022
107 from citation searches
1497 from reference searches
4835 from database-similar article searches
179 records removed before screening
170 duplicates
9 not in English
405 records removed before screening
399 duplicates
6 not in English
3504 records screened
968 records screened
6439 records screened
2942 records excluded
2001 by title
941 by abstract
921 records excluded
769 by title
152 by abstract
5938 records excluded by title
562 reports sought for retrieval
47 reports sought for retrieval
501 reports sought for retrieval
0 reports not retrieved
0 reports not retrieved
0 reports not retrieved
562 reports assessed for eligibility
47 reports assessed for eligibility
501 reports assessed for eligibility
553 reports excluded
82 where the model was not for risk prediction
7 were not in a health-care setting
27 where the model was not eligible for machine learning
269 where user perception was not assessed
8 where the model was finding predictive variables
160 based on review, presentation, conference, protocol, or discussion
43 reports excluded
12 where the model was not for risk prediction
5 where the model was not eligible for machine learning
15 where user perception was not assessed
11 based on review, presentation, conference, protocol, or discussion
473 reports excluded by abstract or full article review
41 studies included in review
9 from database search on June 13, 2022
4 from database search on July 21, 2023
28 from other methods

Figure 1: **PRISMA flow diagram**

alongside other AI models). Eight (20%) of 41 articles analysed tools with multiple elements used in addition to ML risk-prediction.[28,32–34,46,47,52,61]

Study participant numbers ranged from four[33] to 1357 individuals,[65] with convenience sampling (where individuals are chosen on the basis of being readily available) commonly used. Most studies recruited individuals from the USA (22 [54%] of 41 articles) or Canada[28,39,52,61] (four, 10%), with response rates ranging from 4·5%[65] (23 of 510 participants) to 80%[47] (72 of 90). Where recorded, approximately half of studies recruited predominantly female participants (11 [48%] of 23 studies), three-quarters were majority White ethnicity (nine [75%] of 12), with an average age between 30 and 40 years (seven [47%] of 15). The majority (25 [61%] of 41 articles) recruited health-care professionals (HCPs) only; six (15%) members of the public (four articles with patients,[46,54,64,65] one with parents,[59] and one with members from patient and public involvement groups);[50] one (2%) hospital executive only;[63] and one (2%) recruited clinical leaders and executives only.[48] The remaining eight articles recruited physicians, the public and patients,[28,44,52] domain experts,[66] IT and ML experts,[26,44,49] and social workers.[30,55]

Study participants generally interacted with a risk prediction model in a test environment or clinical practice (24 [59%] of 41); of those articles where

	Study design	Setting	Clinical specialty	Number of participants	Number invited to participate	Participant role	Nature of the model	ML technique	Duplicate model
Alabi et al (2022)[26]	Quantitative descriptive	··	Oncology	11	16	Physicians and experts in IT or ML	Prognostic	Deep learning neural network	No
Ballard et al (2013)[27]	Quantitative descriptive	USA	Emergency departments	339	499	Physicians	Prognostic	Binary recursive partitioning	PECARN prediction rules
Barda et al (2020)[21]	Mixed methods	USA	Paediatrics	21	··	Physicians	Prognostic	Random forest model	No
Benrimoh et al (2021)[28]	Mixed methods	Canada	Psychiatry	20	··	Physicians and public or patients	Prognostic	Deep learning neural network	Aifred risk prediction model
Bentley et al (2022)[29]	Qualitative	USA	Urban hospitals	40	60	Physicians	Prognostic	Naive Bayesian classifiers	No
Brown et al (2020)[30]	Quantitative descriptive	··	Psychiatry	139 (68 with complete data)	139	Physicians and social workers	Prognostic	Not specifically testing one model	No
Buck et al (2022)[31]	Qualitative	··	Primary care	18	··	Physicians	··	Not specifically testing one model	No
Carr et al (2022)[32]	Mixed methods	USA	Resource-limited settings	31	62	Physicians	Diagnostic	Probabilistic Bayesian network	ePNa detection system
Chari et al (2023)[33]	Qualitative	··	Renal medicine	4	··	Physicians	Prognostic	Suite of machine learning models	No
Dean et al (2019)[34]	Quantitative descriptive	USA	Emergency departments	90	169	Physicians	Diagnostic	Probabilistic Bayesian network	ePNa detection system
Diprose et al (2020)[35]	Quantitative descriptive	New Zealand	··	170	1315	Physicians	Diagnostic	Hypothetical	No
Elahi et al (2020)[36]	Mixed methods	Uganda	Hospitals in low-income and middle-income settings	25 (use-cases survey), 11 (interview), and 9 (attitude survey)	··	Physicians	Prognostic	Bayesian generalised linear model	No
Fujimori et al (2022)[37]	Mixed methods	Japan	Emergency departments	14	··	Physicians	Diagnostic	XGBoost	No
Ghanzouri et al (2022)[38]	Qualitative	··	Cardiology, vascular medicine, and primary care	12	··	Physicians	Diagnostic	Deep learning neural network	No
Gilbank et al (2020)[39]	Qualitative	Canada	Oncology	10	··	Physicians	Prognostic	Model in development	No
Ginestra et al (2019)[40]	Quantitative non-randomised	USA	Non-intensive care unit admissions	180 (first survey of nurses), 43 (second survey of nurses), 107 (first survey of providers), and 44 (second survey of providers)	360 (first survey of nurses), 180 (second survey of nurses), 356 (first survey of providers), and 107 (second survey of providers)	Physicians	Prognostic	Random forest model	No
Greenberg et al (2022)[41]	Mixed methods	USA	Paediatrics	20	··	Physicians	Prognostic	Binary recursive partitioning	No
Gu et al (2020)[42]	Quantitative descriptive	China	Oncology	32	··	Physicians	Prognostic	XGBoost	No
Henry et al (2022)[43]	Qualitative	USA	Acute clinical care	20	··	Physicians	Prognostic	Unclear ML models	No
Jacobsohn et al (2022)[44]	Mixed methods	USA	Emergency departments	25 (patients from online survey) and 32 (emergency department clinicians from online survey); usability testing unknown	65 (online survey)	Physicians, patients, and experts in IT or ML	Prognostic	Unclear ML models	No
Jauk et al (2021)[45]	Mixed methods	Austria	Inpatients	47 (survey) and 15 (group sessions)	21 (survey of physicians) and 67 (survey of nurses)	Physicians and domain experts	Prognostic	Random forest model	No
Jayakumar et al (2021)[46]	Randomised control trial	USA	Orthopaedics	129	162	Patients	Prognostic	Unclear ML models	No

(Table 1 continues on next page)

	Study design	Setting	Clinical specialty	Number of participants	Number invited to participate	Participant role	Nature of the model	ML technique	Duplicate model
(Continued from previous page)									
Jones et al (2019)[47]	Quantitative descriptive	USA	Emergency departments	72	90	Physicians	Diagnostic	Probabilistic Bayesian network	ePNa detection system
Joshi et al (2022)[48]	Qualitative	USA	Hospitals	21	18 (hospital sites)	Clinical leaders and executives	Prognostic	Not specifically testing one model	No
Masterson et al (2018)[49]	Qualitative	USA	Emergency departments	38	··	Physicians and experts in IT or ML	Prognostic	Binary recursive partitioning	PECARN prediction rules
Musbahi et al (2021)[50]	Mixed methods	UK	··	28	··	Public	Diagnostic	XGBoost	No
Parikh et al (2022)[51]	Qualitative	USA	Oncology	29	72	Physicians	Prognostic	Not specifically testing one model	No
Popescu et al (2021)[52]	Mixed methods	Canada	Psychiatry	7 (staff) and 14 (patients)	20	Physicians and public or patients	Prognostic	Deep learning neural network	Aifred risk prediction model
Rho et al (2022)[53]	Quantitative descriptive	Korea	Oncology	86	1100	Physicians	Prognostic	Random forest model and k-nearest neighbours	No
Richardson et al (2021)[54]	Qualitative	USA	··	87	946	Patients	Prognostic	Not specifically testing one model	No
Romero-Brufau et al (2020)[55]	Quantitative non-randomised	USA	Primary care	45 (pre-implementation) and 38 (post-implementation)	81	Physicians and social workers	Prognostic	Unclear ML models	No
Sandhu et al (2020)[56]	Qualitative	USA	Emergency departments	15	··	Physicians	Prognostic	Deep learning neural network	No
Sax et al (2022)[57]	Mixed methods	USA	Emergency department	8 (semi-structured interviews) and 67 (survey)	Unknown from interview and 103 (survey)	Physicians	Prognostic	Not specifically testing one model	No
Schwartz et al (2022)[58]	Qualitative	USA	Inpatient setting (medical, surgical, and intensive care units)	17	··	Physicians	Prognostic	Unclear ML models	No
Sisk et al (2020)[59]	Quantitative descriptive	··	Paediatrics	404	··	Parents	··	Not specifically testing one model	No
Soliman et al (2023)[60]	Qualitative	Sweden	Cardiology	12 (semi-structured interviews), 12 (user testing), and 3 (discussion)	··	Physicians	Prognostic	CatBoost (resembles gradient-boosting decision trees)	No
Tanguay-Sela et al (2022)[61]	Mixed methods	Canada	Psychiatry	20	··	Physicians	Prognostic	Deep learning neural network	Aifred risk prediction model
Tsai et al (2022)[62]	Quantitative descriptive	Taiwan	Emergency department	10	··	Physicians	Numerous models, prognostic, and diagnostic	Multiple ML and deep learning algorithms	No
Watson et al (2020)[63]	Qualitative	USA	Academic medical centres	33	51	Executives	··	Not specifically testing one model	No
Yarborough et al (2022)[64]	Qualitative	USA	Psychiatry	62	146	Patients	Prognostic	Hypothetical and logistic regression	No
Yarborough and Stumbo (2021)[65]	Mixed methods	USA	Non-veterans	23 (focus groups) and 1357 (survey)	510 (focus groups) and 11 293 (survey)	Patients	Prognostic	Not specifically testing one model	No

ePNa=electronic clinical decision support tool for diagnosis and treatment of pneumonia. ML=machine learning. PECARN=Pediatric Emergency Care Applied Research Network.

Table 1: **Details of the included studies**

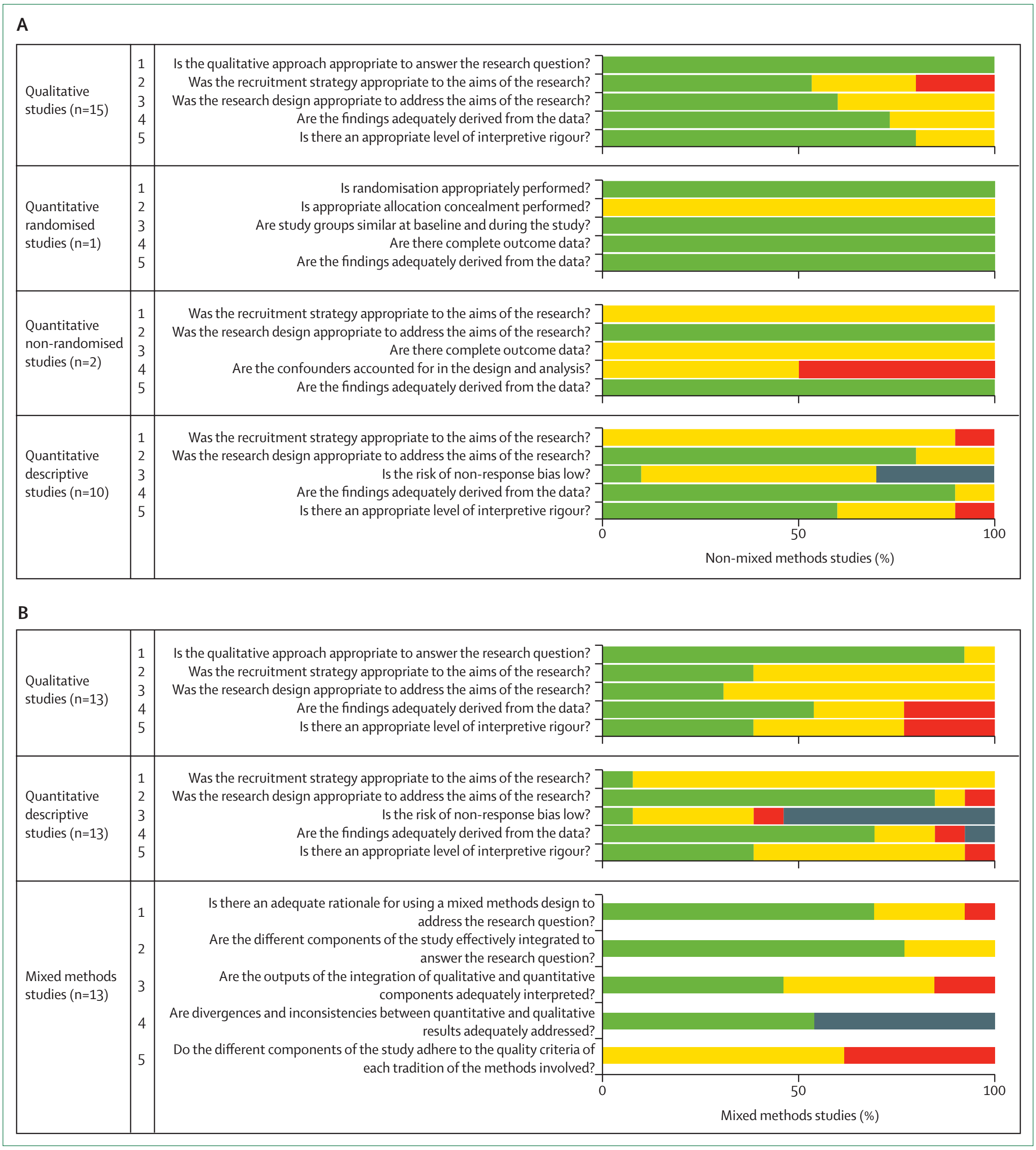

Figure 2: **Quality assessment of included articles**
Quality assessment of (A) non-mixed method studies and (B) mixed methods studies. Article quality were scored as: green (high quality or where the criteria were met), yellow (medium quality or criteria were somewhat met), red (low quality or criteria were not met), and grey (unable to determine).

individuals interacted with a risk prediction model, one (4%) of 24 studies provided model access for user testing but not interviews.[60] Five (21%) of 24 articles were unclear on the ML methods used,[43,44,46,55,58] with 15 unique models included across the remaining articles, the majority of which were prognostic (16, [20%] of 80). Frequently used ML methods included neural networks,[26,28,38,56] binary recursive partitioning,[41,49] random forest,[40,66] and XGBoost and CatBoost techniques.[37,42,60] Common model parameters included patient demographics and clinical information. Many models used EHR data (12 [60%] of 20) and used thresholds highlighting individuals at risk (15, 71%). Where participants did not interact with an ML model, methodologies included asking participants questions regarding ML and risk prediction,[30,31,48,51,54,57,59,63,65] or providing model screenshots or exerts, most of which were prognostic (12, [71%] of 17).

Critical appraisal

Figure 2 provides the percentage of studies where criteria were scored high (green), medium (yellow), or low (red) quality by study type. The single randomised control trial scored 4 of 5 criteria as green, with 1 yellow rating for a lack of blinding.[46] Across all remaining studies, recruitment strategies appeared inadequate, with one study providing no recruitment information[39] and some studies generally missing sample size justification.

Qualitative papers (n=15) scored highly, with a median number of green criteria per article of 4 (range 1–5); shortcomings included the absence of detail regarding interview design[38,39,60,63,64] and poor description of coding and analysis.[38,39,56] Quantitative descriptive studies (n=10) performed less well, with a median green criteria per article of 3 (range 0–3); weaknesses included the absence of information on non-responders[26,27,34,35,46,47] or not mentioning non-response rate.[42,59,62] A similar issue was identified in quantitative non-randomised studies (n=2), where confounding factors were additionally poorly considered.[40,55]

Mixed methods articles (n=13) scored lower than qualitative and quantitative studies in their respective criteria, largely as details pertaining to research design were poorly reported (9 [69%] of 13) or had a poor level of interpretative rigour (9, 69%). Articles additionally scored poorly in criteria specific to mixed methods, with a median number of green criteria per article of 3 (range 0–4). Weaknesses included inadequate rationale for using mixed methods[21,44,50,52] and poor integration of

	Key theme	Exemplar quote
Contextual barriers and facilitators to use		
HCP	Concerns felt before implementation—that models would take up too much time and negatively affect workload—were not borne out in practice	"[...] in routine cases, [AI] would not be a time saver for me."[31]
HCP	It is important to sympathetically integrate new tools into the existing workflow and consider current work processes and priorities	"It's an interruption…Every single call we get is completely disruptive to workflow."[56]
HCP	It is important to provide education on tool use and feedback on tool outcomes	"But I think you need to loop it back and [...] say: what was the performance on my patients? How many did I send to radiation nursing clinic and how much [did] it drop the actual rate of emergency admissions?"[39]
Hospital executives	Alert fatigue is a major concern, although no universal solution appears to be known at present	"To optimise the alert, an issue is there is no consensus for what optimal means: there is no clear consensus for setting thresholds"[48]
	Consideration is needed regarding model costs alongside personnel and technical requirements	"Don't give up. You got to just keep chugging. Sometimes it's a lot of little steps that sometimes feel like you're climbing a mountain that doesn't end."[48]
Desirability of models in the health-care setting		
HCP	Feeling that models are easy to use and are useful in the health-care setting	"It provides you clinical support to provide better quality of care to the patient."[49]
HCP	View ML risk prediction models are superior to alternatives, although HCPs have skills and abilities that cannot be matched by models	"[The system] can't help you with what it can't see."[43]
HCP	Suggestion that models should recommend next steps and actions	"Understanding how the predicting part comes in, I think would give me more confidence…some sort of like if/then tool, so if the score is greater than this, then you should take this kind of action."[58]
HCP and executives	Recommendation that clinicians should be involved in the model development process	"I think just with anything, having someone who's actually been there done that is way more, makes it, makes whatever you're developing way more accurate, way more useful, way more diligent."[58]
Patients or public	Feeling that models can improve care and are acceptable to reduce the burden on the health-care workforce	"Using AI can just, reduce the burden on the health work force, meaning doctors can do what they're supposed to do."[50]
Patients or public	Feeling an individuals' unique situation must be considered alongside tool recommendations	"[I]t's important to take into account that people, depending on what the AI comes out with, people might not be willing to go with what that is, they might need alternates."[54]
Model development and outputs		
HCP	It is important to understand how models have produced their predictions	"What's the weight of the data that is available from the moment they did transfer them to the ICU and how does that carry into this predictive model?"[21]
HCP	There is no consensus regarding whether false positives or negatives are preferable	"We get so many [I] prompts for drug interactions and other things that I often ignore them. My fear is if there were too many false positives, I would start to ignore them."[51]
HCP	Models should be user friendly and simple	"It took me three clicks to get here; it would have been better if it was simpler. It does affect my usage and efficiency and maybe satisfaction."[38]
Patients or public	Concerns were raised regarding data privacy	"We need to be aware that we know nothing about who these people that are creating these AI algorithms, they can be anyone and they'd have access to all our data."[50]
HCPs and patients or public	There are concerns regarding quality of data used to build models and inbuilt bias in data and model development	"I just think there's a lot of under and over-reporting of co-morbidities. I'm not sure what [the algorithm] would capture from [the electronic medical record], but that would be my biggest worry is that the data it is using to make its calculation is not accurate. I guess that would be my biggest worry."[32] "There's a lot of discrepancies in the medical record I must say, especially now that you can see your portal. I know I've seen things saying that certain things were done or about myself and procedures that were totally not true. So I've had a lot of different things in my medical chart that are inaccurate, very inaccurate, so if they're training an artificial intelligence that this is facts, it's like, well no."[54]

(Table 2 continues on next page)

	Key theme	Exemplar quote
(Continued from previous page)		
Potential effects for patients and HCPs		
HCP	Models seen to positively influence clinician behaviours and knowledge	"The initiative…just creates a lot more vigilance…I almost feel like I'm very cognizant of sepsis and almost like, imagining the Sepsis Watch people upstairs like, looking down on me…I'm honestly like, just waiting for their call."[56]
HCP	Model use can aid communication with patients and within teams, although the effects on the clinician–patient relationship was unclear	"I think that the visual aid would be really great for families to understand, to completely understand, and bring it down into concrete terms"[49]
HCP	HCPs should retain autonomy over clinical decisions	"Once it falls on the lap of the neurosurgeon, I'd want to not be forced to walk down this aisle and no other alternative…In other words, I want to have the option of using my own judgment as well."[41]
HCP	The patients, staff, and setting that could see the most positive effects of these tools was unclear	"It's probably a way more useful tool, not in the ED. In the ED, all we think about all the time is sepsis [be]cause it's such a big part of our practice. So, that's why I think it doesn't apply well to us, but it would apply well in other settings where they don't think about or see or miss the bundle more often."[56]
Patients or public	Concerns regarding unintended consequences, particularly regarding use of model outputs by insurance companies	"I mean, …that information is wonderful, but who's gonna get it after the doctors look at it is my big thing. Is the insurance company gonna take it, and now all of a sudden…my premium doubles for health insurance?"[54]
HCPs and patients or public	There are fears HCPs might become reliant on such tools	"I think [that] there are a lot of people, frankly, that will quickly default to having a tool tell them what to do and stop assessing, and I hope that's not true, but I've seen it happen"[43] "If they were to get hacked or a system goes down…like what's the contingency plan, but what is the contingency plan? If you have all these doctors who are so used to having this artificial intelligence read all these, and they don't have the skill of reading it, then what happens?"[54]

AI=artificial intelligence. ED=emergency department. HCP=health-care professional. ICU=intensive care unit.

Table 2: **Summary of key findings regarding perspectives of machine learning risk prediction models in the health-care setting by stakeholder perspective**

the inferences from the differing methods.[21,28,32,36,44,50,52] The critical appraisal inter-rater reliability Cohen's kappa value was 0·57 (95% CI 0·49–0·65): before the discussion, the total number of agreements was 241, with 133 agreements likely to be caused by chance (see appendix pp 137–39 for critical appraisal score per included article).

Synthesis

Our deductive approach established four broad themes: (1) contextual barriers and facilitators to ML risk prediction model use in the health-care setting (identifying setting-dependent factors affecting uptake), (2) desirability of such models within the health-care setting (uncovering whether ML risk prediction models can potentially add value within the health-care setting and the factors affecting acceptance), (3) model development and outputs (perceptions of ML risk prediction model development techniques and design), and (4) the potential effect for patients and HCPs (possible consequences of model use to inform clinical practice). Alongside each broad theme, recommendations to overcome barriers were recorded.

Coding was applied with information on whether positive, negative, or mixed respondent views dominated the articles (appendix pp 140–46). With this information, we found that for articles including only patient views or public views (n=6), on average 34% (95% CI 3–65%) of the themes considered per article were positive views and 39% (95% CI 18–59) were negative views, compared with 48% (95% CI 34–62) with positive and 23% (95% CI 13–34) negative views for articles with HCP participants only (n=25). Articles where participants had access to an ML risk prediction tool as part of the study (n=24) showed on average 56% (95% CI 43–70) of the themes considered per article were positive views and 15% (95% CI 7–23) were negative, compared with 34% (95% CI 17–52) positive and 43% (95% CI 27–59) negative for articles with no model access (n=17). Table 2 summarises the key themes per participant group and the appendix (pp 147–65) contains themes identified and exemplar quotes.

Contextual barriers and facilitators

At an organisational level, a supportive culture positively influenced uptake, with tool use reinforced by a strong match with the mission of the institution.[49] Patients, HCPs, and health-care executives reported concerns regarding financial costs[31,54,63] and technical issues.[31,36,37,49,50,54,63] HCPs and hospital executives believed that competing priorities[21,44,49] and difficulty implementing models into the clinical workflow[29,45,48,56] were barriers to success (ie, barriers to use in the setting). There appeared to be a conflict between the expected extra demand on health service resources and staff when using such tools[29,36] versus potential resource reduction.[66] Hospital executives expressed the importance of the connection between HCPs and IT,[48,63] while HCPs[26] felt no need for technical support.

At the individual level, HCP knowledge pertaining to AI and ML techniques appeared mixed,[21,27,31] with training deemed as important.[21,29,48,49,57,58,60] Both HCPs and hospital executives identified alert fatigue[29,48,63] (ie, desensitisation to risk prediction alerts) as a barrier to success; however, one article reported this issue was lower with ML models,[48] whereas models with appropriate alerts had a positive effect on ML model perception.[21] Some HCPs had concerns before implementation (eg, demands on workload)[21,29,31,37,44] that were not borne out in practice.[36,45,46,52]

To maximise ML model acceptance and uptake, recommendations included providing feedback to users,[39,48,49] revising procedures,[60] endorsement by peers or professional bodies,[27,31,43,58] and appointing model champions.[48,49,56,60] It was important to optimise model integration into the workflow with ease,[55] minimising interruptions[37] and similarity in existing processes[27,39,49] positively received; however, articles did not build consensus on a solution for reducing alert fatigue.[29,44,48,56,57]

Desirability of models in the health-care setting

Articles largely reported positive HCP and patient views regarding ML models being useful or helpful[27,31,37,41,42,44–47,50,60,62] (12 of 14 positive views), having the potential to improve care[27,46,49,50,56,61,65] (eight of 11 positive; albeit with improvements only verified following long-term follow-up),[62] and improving patient safety[41,62] and aiding clinical diagnosis and decision making[26–28,34,36,38,39,41–44,46,50,54] (14 of 18 positive). HCP and patient views were all positive regarding models being easy to use,[26,27,36,41,45,49,52,62] understandable,[26,35,37,45] and holding potential to add value in the health-care setting,[39,29,51,54] with HCP perceptions improving with use.[32,61] However, model trust was affected by agreement with previous beliefs of HCPs[46] and their own predictions;[35,58] with predictions at risk of being disregarded when in disagreement with clinical judgement.[39,51] Perceived actionability of model recommendations also affected tool usefulness.[21,29,33,57,58,63]

HCPs believed they had abilities and capabilities not covered by models[31,43,56,58] and patients recommended considering an individual's unique situation when applying such models to avoid harm.[54] Suggestions to enhance trust included ensuring that models suggest realistic and actionable recommendations[63] and including HCPs in model development.[58,63] HCPs and hospital executives largely believed ML risk prediction models were superior to other tools;[37,43,48,56] however, one article described HCP hesitancy compared with simpler models suggesting that discomfort would be mitigated if ML models were "to bias towards a more conservative pathway of care".[57]

HCPs suggested providing clear evidence of model validation,[39,43] accuracy,[29,31,39,58] evidence of safety,[57] and effectiveness in improving patient outcomes or outperforming clinical judgement.[29,31,38,48,49,56–58] One article suggested undertaking model evaluation to measure its effect,[63] while patients requested sufficient model testing and a careful transition of tools to practice.[54]

Model development and outputs

Respondents largely perceived model outputs as reasonable or accurate for predicting disease or outcomes:[28,31,45,52,56,58,61] seven (78%) of nine articles expressed positive views on accuracy rather than negative views with potentially increased accuracy comparative to HCPs.[58] Fewer articles reported concerns regarding accuracy of models,[48,51,56] inherent bias,[54,57] or potential to miss cases or recommend inappropriate treatments.[34,55] HCPs did not reach consensus regarding whether false positives or negatives were more concerning.[27,30,48,51,65]

In all studies where it was considered, concerns were raised by patients, HCPs, and hospital executives around data quality for model development[21,29,31,48,51,54,58,63,64] and biases therein.[54,58] Concerns included under-reporting and over-reporting,[51] missing data,[58,63] data representation of real patients,[50] timeliness of data availability,[58] accuracy, and consistency of data coding.[63] The majority of articles expressing negative data quality views did not provide participants with access to an ML-based risk prediction model (eight of ten).[21,29,31,50,51,54,63]

Data privacy was considered in five of nine articles involving patients and the public[50,54,59,64,65] and only one involving HCPs;[31] the topics of concern ranged from data privacy identified as the most common public concern regarding AI use[50] to data use not being considered an invasion of privacy.[64]

Recommendations included ensuring a user-friendly model interface[29,38,56] with simplicity key[31,35,36,38,48] and the consideration of risk terminology.[21,38,44] Patients expressed the need for oversight and a regulatory framework to protect against harm.[54] HCPs and patients reported concerns regarding model input variables—missing[41,64] or counterintuitive variables[21,61]—with a suggestion for model customisation to exhibit parameter control.[48] HCPs stated the importance in understanding how models are developed and validated,[21,39,41,58] their relevance to a particular patient,[39] local and national average risks for comparison,[57] and how they make their predictions,[21,27,29–31,35,38,58,61] although there was no consensus regarding the information required[21,39,61] or preferred explanation format.[29,35,58] Patients and the public did not comment on this requirement for model explanations in any included article.

Potential effects for patients and HCPs

Positive aspects of ML models reported by HCPs included providing additional insight,[26,28,45,61] positively influencing behaviour,[30,36,40,41,51,56,58] aiding communication with patients,[27,28,49,51,61] and improved team dynamics.[40,41,55] Concerns included over-reliance,[31,43,50,51,54] use of models leading to clinicians becoming redundant,[31] unintended consequences (eg, raising stigma directed at patients),[29,51,57,64,65] models affecting disparities in care,[57] models used by regulators to standardise and monitor care,[31,43] and models used by insurance companies to limit access to care.[54,64,65] Patient views heavily contributed to these negative perceptions of potential model effects (four of 11 articles had negative views),[50,64,65] with little positive patient feedback identified overall.

Views regarding the effect of the patient–HCP relationship were mixed.[28,31,46,52,61] Agreement between

HCPs and ML model recommendations had potential to reassure HCPs and offer a level of medico-legal protection,[27,31] whereas disagreement was feared to lead to negative implications when the model recommendations were not followed.[29,41,56] Few studies considered the effect of ML models on patient empowerment.[46,55]

Some HCPs expressed the view that ML models offer the largest benefit when used by less experienced HCPs.[37,41,49,56] We found an absence of consensus around the clinical setting to potentially benefit the most[29,36,49,56] and whether ML models were suitable for complex patients (ie, have difficult to predict disease or diagnostic uncertainty,[31,37,38,51] have rare diseases,[31] have unique or complex patient characteristics,[58] or are patients at risk).[57] There was a strong feeling that HCPs should retain autonomy,[41,43] with ML models supplementing and not replacing them.[39,43,50,54] Patients wished to control model use during consultations[54] and desired only trusted physicians to access risk information.[65]

Discussion

This Review develops understanding regarding HCP and patient perceptions of ML risk prediction models and their uses. We identified that where respondents do not have access to an ML model, they hold more negative views; implying fears might be lowered following actual tool use. Patients and members of the public might also hold more negative views regarding models, although more patient research is required to confirm the significance of this finding.

Perceptions were largely positive in terms of added value; however, responders identified many barriers and factors negatively affecting trust, particularly pertaining to contextual factors and methodological concerns. In terms of potential effects, HCP views appeared positive compared with patients and the public.

Many emerging themes had both positive and negative perspectives. Participants, for example, felt models aided decision making yet were concerned about over-reliance, the ability for models to protect from or lead to legal litigation, and the possibility that models might increase health service demands versus reducing resource need. Reservations that were identified potentially explain the low level of model use within health care. This Review suggests recommendations to improve uptake.

We identified contextual influences acting as barriers to model integration. However, given some concerns felt before implementation were not borne out in practice, consideration of longer-term effects might allay some fears. Broader research also emphasises opportunities not identified here—for example, decreasing financial costs with reducing inefficiencies and reducing overhead costs with time.[67] Careful implementation oversight[68] might also control for many systemic barriers identified, while further research in the field appears essential regarding workflow integration and alert fatigue.

ML risk prediction models were thought to be largely superior to alternatives and had the potential to add value in the health-care setting. However, clinician capabilities were deemed unmatched by models, while evidence supporting models and concordance between models and physician predictions was important for uptake, as was perceived actionability from model recommendations. The review by Kennedy and colleagues analysing HCP perspectives of clinical prediction rules supports many of these findings.[11] It is evident that models and physicians offer differing strengths, with benefit maximised when models are designed to add value for users. We identified the importance of physician involvement in model design; other studies also suggest involving patients.[69]

ML model outputs were generally deemed accurate; however, concerns regarding model development negatively affected perception, in particular data privacy, developmental data quality, and biases. Such data concerns echo findings elsewhere.[12,51,69] However, although data bias concerns were raised, there was little discussion within the included articles regarding the potential for tools to reinforce ethnic inequalities as stressed elsewhere,[70–73] with other studies highlighting the need to proactively design and use ML tools to advance health equity.[70] In line with other research,[9] we found that the explanations behind model predictions were important. Furthermore, model relevance to real-life patients was desired; recent research might assist with this requirement by enabling adaptation of ML models to accommodate regional variations.[74] Models with missing risk factors were viewed sceptically here: wider research shows that ML outperforms statistical models when either high or low variable counts are included,[75] although developers should consider what a complex model adds, for example the trade-off between variable inclusion and simplicity.[11]

We found autonomy important to HCPs and patients, while the role of predictive models should be clearly defined as individuals feared negative consequences of non-compliance as seen elsewhere.[11] Although the effect on patient–clinician relationship was unclear, models were perceived to provide additional insight and improve communication. Users reported fears of unintended consequences as identified in other studies, which recommend weighing potential benefits and harms before implementation.[12] Model suitability for complex patients appeared unclear here, with wider research indicating ML models offer improved prediction in relevant subpopulations comparative to statistical techniques.[75]

Our findings pertaining to ML risk prediction models largely align with perceptions of clinical decision support systems (CDSS) and risk prediction models (non-ML and in general) in the health-care setting. With analogous potential to add value,[11] there were similar concerns regarding factors such as alert fatigue,[8,76] data quality,[9] and unintended consequences[11] along with model design requirements[8,11,76] and retention of clinician autonomy.[11] Our Review did not, however, identify concerns regarding lack of model maturity to fully accommodate a problem—a concern with CDSS algorithms[8]—potentially implying more user faith in ML model capabilities. Indeed, we reported perceptions of superiority of ML models over rule-based alternatives in capturing complex disease dynamics.[48] However, the importance of model explanations identified here appears heightened for ML models, likely due to unfamiliarity, and the non-linear relationships used by such models comparative to non-ML techniques.

There has been an upsurge in standards and regulations relevant to health-care ML models, alongside best practice research and guidance, with importance placed on themes, including data security, transparency, and accountability.[77] Alignment with standards and guidelines might allay some fears outlined here and increase trust.

The strengths of this Review include that only peer-reviewed published literature were included; the study considered a broad range of settings, model uses, and stakeholders; was not restricted to EHR embedded models, and multiple study designs were included. Review of study quality was undertaken with a bespoke assessment tool and although not validated, its construction combined two widely accepted tools[22,23] and enabled critique of the included studies that was not possible by either tool independently.

This Review amalgamated results using a narrative synthesis, with inclusion of quotes to illustrate findings. Although a meta-analysis might potentially strengthen findings, study heterogeneity made this unsuitable. The inclusion of quantitative, qualitative, and mixed research methods enabled inclusion of complementary perspectives from different study types.

The Review has several limitations. First, article selection and analysis were performed by one author. Second, the included articles were identified predominantly via citation searches. However, previous research suggests challenges in developing searches to identify prognostic factor studies and impact prognostic studies[19] (including the absence of indexing of prediction studies and rarity of impact studies), the need for citation searching was anticipated. Last, we included only articles published in English (during article screening only 15 articles were removed from inclusion due to the language being non-English).

Articles included a range of ML techniques; thus, findings based on one model might not have relevance to more or less complex models; however, this breadth is helpful to develop understanding across the range of ML techniques. Furthermore, distinguishing models as either ML or statistical is challenging,[78] and the taxonomy used will vary across ML reviews. In instances where articles assessed additional tools to ML risk prediction, study results were considered relevant unless clearly irrelevant. However, such comments cannot be guaranteed to refer to ML risk prediction models, potentially restricting result validity; nevertheless, a level of relevance can be assumed due to accordance of our results with broader literature.

Of the included studies, participants were predominantly recruited from high-income countries, restricting the relevance of the findings elsewhere. Additionally, only nine of 41 (22%) articles recruited patients or the public, limiting patient perspectives. Further, five of the nine articles including patients and members of the public recruited individuals solely from the USA; use of the USA health-care system might explain why this group held largely negative perceptions regarding potential effects of ML risk prediction models (eg, due to concerns regarding insurance and cost of health care). Further analysis of the audience perspective outside the USA might present differing views. The search undertaken also focused on identifying patients and HCPs, rather than hospital leaders, which might explain why requirements for oversight and governance along with broader stakeholder involvement were not identified.

Although we analysed how respondent perceptions varied per article by respondents' role (and access to a model), we were unable to ascertain effects of other study factors (eg, respondent location and ML technique) due to low article numbers per category. More research in this field might enable further patterns and associations to be identified.

Article quality varied, with qualitative studies performing well compared with mixed method studies. Weaknesses included poor recruitment strategies with a high risk of non-response bias, meaning representativeness of findings across the intended populations was questionable. Furthermore, included mixed methods papers were unable to justify the inclusion of multiple study methods, with integration of results failing to enhance the analysis; given this is pivotal to mixed method research,[79] articles should improve methods to optimise their value.

There is recognition that health care is not currently taking full advantage of AI technology.[80] Our Review assists by facilitating acceptable ML risk prediction model development. Furthermore, we found concerns for before the implementation of models were not borne out in practice; further implementation and impact studies in this field might help raise expectations and increase trust. ML model developers should also monitor appropriate guidance, with adherence potentially allaying fears.

Physician and other stakeholder involvement is important in model design and implementation. Institutions should also make clear the anticipated interaction between models and users, to clarify responsibilities and capitalise on human and machine capabilities. We found conflicting perspectives relating to many of the themes considered; anticipated benefits and potential negative consequences of models should therefore be considered.

This Review highlighted several areas warranting further research, to provide clarification on ambiguous aspects. For model development, there is a need to understand user expectations regarding model explanations. Further research for ML model testing in health-care settings is needed to assess strategies to reduce alert fatigue and investigate the settings and patients who would benefit greatest from model use. Given the potential negative effect on patient–clinician relationship, it would be useful to identify optimal approaches to model use in consultations. Further investigation is needed for model applicability to minority populations inadequately represented in training and testing datasets and the potential effects of models and more feedback from more diverse ethnic backgrounds and developing countries should be collected. More patients and members of the public in studies investigating perceptions of ML risk prediction models should be recruited. More research on patients and public perceptions outside of the USA are needed, particularly regarding views on model potential effects. With further evaluation, there can be better understanding of physician needs for ongoing model technical support. Furthermore, it would be beneficial to review health-care leaders' perspectives to understand governance and oversight requirements, and understand other barriers to implementation (eg, financial cost).

Conclusion

This Review evaluated patient and health-care staff perceptions of ML-based predictive models, alongside recommendations to improve their uptake and trust. Generally, perceptions were positive that models have the potential to add benefit in the health-care setting; however, reservations remain. We have identified gaps in knowledge, particularly views from under-represented groups, and optimal methods for model explanation and alerts, which require future research.

Contributors

RG and NN were responsible for study conceptualisation. Development of the search strategy along with record screening and data extraction was performed by RG. Both RG and AJ performed the critical appraisal of included articles. All coding and formal analysis was performed by RG, while drafting of the manuscript was performed by RG and AJ. All authors were responsible for manuscript review and editing. All authors had full access to all data in the study and accept responsibility to submit for publication.

Declaration of interests

RG is funded by the National Institute for Health and Care Research (NIHR) for this study (NIHR302604). TC is a co-founder of and owns stock in Mortimer Health and is supported by the Wellcome Trust via a Wellcome Clinical PhD Training Fellowship (REF 22890/Z/21/Z). SMJ is supported by Cancer Research UK (CRUK; EDDCPGM\100002) and Medical Research Council (MRC; MR/W025051/1) programme grants; received fees for advisory board membership in the last three years from Bard1 Lifescience; received grant income from Owlstone and GRAIL; is an unpaid member of a GRAIL advisory board; has received lecture fees for academic meetings from Cheisi and AstraZeneca; and receives support from the CRUK Lung Cancer Centre and the CRUK City of London Centre, the Rosetrees Trust, the Roy Castle Lung Cancer foundation, the Longfonds BREATH Consortia, MRC UKRMP2 Consortia, the Garfield Weston Trust, and University College London (UCL) Hospitals Charitable Foundation where this work was partly undertaken and who also received a proportion of funding from the Department of Health's NIHR Biomedical Research Centre's funding scheme. SMJ's wife works for AstraZeneca. MvdS is a Director at the Cambridge Centre for AI in Medicine, which receives funding from AstraZeneca and GSK. JS is funded by NIHR Applied Research Collaboration North Thames, NIHR (policy research programme, programme grant, health service and delivery grant, and capacity building funding), and the UK Prevention Research Partnership. NN is supported by an MRC Clinical Academic Research Partnership (MR/T02481X/1); received a grant from the CRUK early diagnosis grant for the Real-time cancer analytics (REACT) study (EICEDAAP\100012); reports honoraria for non-promotional educational talks or advisory boards from Amgen, AstraZeneca, Boehringer ingelheim, Bristol Myers Squibb, EQRx, Fujifilm, Guardant Health, Intuitive, Janssen, Lilly, MSD, Olympus, OncLive, PeerVoice, Pfizer, Roche, and Takeda outside of the current work; holds positions as the Director of UK Lung cancer coalition, member of the Steering Committee of British Thoracic Oncology Group, and Clinical Director of National Lung Cancer Audit; and reports stock and stock options in OneWelbeck, The Physicians' Clinic (HCA), and Pharmacierge. AJ declares no competing interests. The views expressed in this publication are those of the authors and not necessarily those of the NIHR, National Health Service, or the UK Department of Health and Social Care.

Data sharing

The search strategy is available in the appendix (pp 2–8). Additional data are available upon request from the corresponding author.

Acknowledgments

We thank David Green for his kind assistance with formulating the search strategy for this Review. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

References

1 Sun H, Depraetere K, Meesseman L, et al. Machine learning-based prediction models for different clinical risks in different hospitals: evaluation of live performance. *J Med Internet Res* 2022; **24:** e34295.

2 Sundström J, Schön TB. Machine learning in risk prediction. *Hypertens* 2020; **75:** 1165–66.

3 Badré A, Zhang L, Muchero W, Reynolds JC, Pan C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J Hum Genet* 2021; **66:** 359–69.

4 Masum S, Hopgood A, Khan J. Artificial intelligence and machine learning for risk prediction in surgery. *J Cancer Sci Clin Ther* 2022; **6:** 358–59.

5 Dhiman P, Ma J, Andaur Navarro CL, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. *Diagn Progn Res* 2022; **6:** 13.

6 van der Schaar M. Interpretability: from black boxes to white boxes. April 14, 2020. https://www.vanderschaar-lab.com/from-black-boxes-to-white-boxes/ (accessed Oct 12, 2021).

7 Hassan N, Slight RD, Bimpong K, et al. Clinicians' and patients' perceptions of the use of artificial intelligence decision aids to inform shared decision making: a systematic review. *Lancet* 2021; **398:** S80 (abstr).

8 Miller A, Moon B, Anders S, Walden R, Brown S, Montella D. Integrating computerized clinical decision support systems into clinical work: a meta-synthesis of qualitative research. *Int J Med Inform* 2015; **84:** 1009–18.

9 Knop M, Weber S, Mueller M, Niehaves B. Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence-enabled clinical decision support systems: literature review. *JMIR Hum Factors* 2022; **9:** e28639.

10 Hogg HDJ, Al-Zubaidy M, Talks J, et al. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. *J Med Internet Res* 2023; **25:** e39742.

11 Kennedy G, Gallego B. Clinical prediction rules: a systematic review of healthcare provider opinions and preferences. *Int J Med Inform* 2019; **123:** 1–10.

12 Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *Lancet Digit Health* 2021; **3:** e599–611.

13 Seibert K, Domhoff D, Bruch D, et al. Application scenarios for artificial intelligence in nursing care: rapid review. *J Med Internet Res* 2021; **23:** e26522.

14 Wu C, Xu H, Bai D, Chen X, Gao J, Jiang X. Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis. *BMJ Open* 2023; **13:** e066322.

15 Lee TC, Shah NU, Haack A, Baxter SL. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. *Informatics* 2020; **7:** 25.

16 Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017; **24:** 198–208.

17 Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; **339:** b2535.

18 Xie A, Carayon P. A systematic review of human factors and ergonomics (HFE)-based healthcare system redesign for quality of care and patient safety. *Ergonomics* 2015; **58:** 33–49.

19 Geersing G-J, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KGM. Search filters for finding prognostic and diagnostic prediction studies in MEDLINE to enhance systematic reviews. *PLoS One* 2012; **7:** e32844.

20 Vasey B, Ursprung S, Beddoe B, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. *JAMA Netw Open* 2021; **4:** e211276.

21 Barda AJ, Horvat CM, Hochheiser H. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Med Inform Decis Mak* 2020; **20:** 257.

22 Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Educ Inf* 2018; **34:** 285–91.

23 Critical Appraisal Skills Programme. CASP checklists. 2023. https://casp-uk.net/casp-tools-checklists/ (accessed Jan 12, 2022).

24 Hong QN, Gonzalez-Reyes A, Pluye P. Improving the usefulness of a tool for appraising the quality of qualitative, quantitative and mixed methods studies, the Mixed Methods Appraisal Tool (MMAT). *J Eval Clin Pract* 2018; **24:** 459–67.

25 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; **372:** n71.

26 Alabi RO, Almangush A, Elmusrati M, Leivo I, Mäkitie A. Measuring the usability and quality of explanations of a machine learning web-based tool for oral tongue cancer prognostication. *Int J Environ Res Public Health* 2022; **19:** 8366.

27 Ballard DW, Rauchwerger AS, Reed ME, et al. Emergency physicians' knowledge and attitudes of clinical decision support in the electronic health record: a survey-based study. *Acad Emerg Med* 2013; **20:** 352–60.

28 Benrimoh D, Tanguay-Sela M, Perlman K, et al. Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician-patient interaction. *BJPsych Open* 2021; **7:** e22.

29 Bentley KH, Zuromski KL, Fortgang RG, et al. Implementing machine learning models for suicide risk prediction in clinical practice: focus group study with hospital providers. *JMIR Form Res* 2022; **6:** e30946.

30 Brown LA, Benhamou K, May AM, Mu W, Berk R. Machine learning algorithms in suicide prevention: clinician interpretations as barriers to implementation. *J Clin Psychiatry* 2020; **81:** 19m12970.

31 Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General practitioners' attitudes toward artificial intelligence-enabled systems: interview study. *J Med Internet Res* 2022; **24:** e28916.

32 Carr JR, Jones BE, Collingridge DS, et al. Deploying an electronic clinical decision support tool for diagnosis and treatment of pneumonia Into rural and critical access hospitals: utilization, effect on processes of care, and clinician satisfaction. *J Rural Health* 2022; **38:** 262–69.

33 Chari S, Acharya P, Gruen DM, et al. Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes. *Artif Intell Med* 2023; **137:** 102498.

34 Dean NC, Vines CG, Rubin J, et al. Implementation of real-time electronic clinical decision support for emergency department patients with pneumonia across a healthcare system. *AMIA. Annu Symp proceedings AMIA Symp* 2019; 2019: 353–62.

35 Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *J Am Med Inform Assoc* 2020; **27:** 592–600.

36 Elahi C, Spears CA, Williams S, et al. An attitude survey and assessment of the feasibility, acceptability, and usability of a traumatic brain injury decision support tool in Uganda. *World Neurosurg* 2020; **139:** 495–504.

37 Fujimori R, Liu K, Soeno S, et al. Acceptance, barriers, and facilitators to implementing artificial intelligence-based decision support systems in emergency departments: quantitative and qualitative evaluation. *JMIR Form Res* 2022; **6:** e36501.

38 Ghanzouri I, Amal S, Ho V, et al. Performance and usability testing of an automated tool for detection of peripheral artery disease using electronic health records. *Sci Rep* 2022; **12:** 13364.

39 Gilbank P, Johnson-Cover K, Truong T. Designing for physician trust: toward a machine learning decision aid for radiation toxicity risk. *Ergon Des* 2020; **28:** 27–35.

40 Ginestra JC, Giannini HM, Schweickert WD, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. *Crit Care Med* 2019; **47:** 1477–84.

41 Greenberg JK, Otun A, Kyaw PT, et al. Usability and acceptability of clinical decision support based on the KIIDS-TBI tool for children with mild traumatic brain injuries and intracranial injuries. *Appl Clin Inform* 2022; **13:** 456–67.

42 Gu D, Su K, Zhao H. A case-based ensemble learning system for explainable breast cancer recurrence prediction. *Artif Intell Med* 2020; **107:** 101858.

43 Henry KE, Kornfield R, Sridharan A, et al. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. *NPJ Digit Med* 2022; **5:** 97.

44 Jacobsohn GC, Leaf M, Liao F, et al. Collaborative design and implementation of a clinical decision support system for automated fall-risk identification and referrals in emergency departments. *Healthcare* 2022; **10:** 100598.

45 Jauk S, Kramer D, Avian A, Berghold A, Leodolter W, Schulz S. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. *J Med Syst* 2021; **45:** 48.

46 Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an artificial intelligence-enabled patient decision aid *vs* educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. *JAMA Netw Open* 2021; **4:** e2037107.

47 Jones BE, Collingridge DS, Vines CG, et al. CDS in a learning health care system: identifying physicians' reasons for rejection of best-practice recommendations in pneumonia through computerized clinical decision support. *Appl Clin Inform* 2019; **10:** 1–9.

48 Joshi M, Mecklai K, Rozenblum R, Samal L. Implementation approaches and barriers for rule-based and machine learning-based sepsis risk prediction tools: a qualitative study. *JAMIA Open* 2022; **5:** ooac022.

49 Masterson Creber RM, Dayan PS, Kuppermann N, et al. Applying the RE-AIM framework for the evaluation of a clinical decision support tool for pediatric head trauma: a mixed-methods study. *Appl Clin Inform* 2018; **9:** 693–703.

50 Musbahi O, Syed L, Le Feuvre P, Cobb J, Jones G. Public patient views of artificial intelligence in healthcare: a nominal group technique study. *Digit Health* 2021; **7:** 20552076211063682.

51 Parikh RB, Manz CR, Nelson MN, et al. Clinician perspectives on machine learning prognostic algorithms in the routine care of patients with cancer: a qualitative study. *Support Care Cancer* 2022; **30:** 4363–72.

52 Popescu C, Golden G, Benrimoh D, et al. Evaluating the clinical feasibility of an artificial intelligence-powered, web-based clinical decision support system for the treatment of depression in adults: longitudinal feasibility study. *JMIR Form Res* 2021; **5:** e31862.

53 Rho MJ, Park J, Moon HW, et al. Dr. Answer AI for prostate cancer: intention to use, expected effects, performance, and concerns of urologists. *Prostate Int* 2022; **10:** 38–44.

54 Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ Digit Med* 2021; **4:** 140.

55 Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. *Int J Med Inform* 2020; **137:** 104072.

56 Sandhu S, Lin AL, Brajer N, et al. Integrating a machine learning system into clinical workflows: qualitative study. *J Med Internet Res* 2020; **22:** e22421.

57 Sax DR, Sturmer LR, Mark DG, Rana JS, Reed ME. Barriers and opportunities regarding implementation of a machine learning-based acute heart failure risk stratification tool in the emergency department. *Diagnostics* 2022; **12:** 2463.

58 Schwartz JM, George M, Rossetti SC, et al. Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: qualitative descriptive study. *JMIR Hum Factors* 2022; **9:** e33960.

59 Sisk BA, Antes AL, Burrous S, DuBois JM. Parental attitudes toward artificial intelligence-driven precision medicine technologies in pediatric healthcare. *Children* 2020; **7:** 145.

60 Soliman A, Nair M, Petersson M, et al. Interdisciplinary human-centered AI for hospital readmission prediction of heart failure patients. *Stud Health Technol Inform* 2023; **302:** 556–60.

61 Tanguay-Sela M, Benrimoh D, Popescu C, et al. Evaluating the perceived utility of an artificial intelligence-powered clinical decision support system for depression treatment using a simulation center. *Psychiatry Res* 2022; **308:** 114336.

62 Tsai W-C, Liu C-F, Lin H-J, et al. Design and implementation of a comprehensive AI dashboard for real-time prediction of adverse prognosis of ED patients. *Healthcare* 2022; **10:** 1498.

63 Watson J, Hutyra CA, Clancy SM, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open* 2020; **3:** 167–72.

64 Yarborough BJH, Stumbo SP, Schneider JL, Richards JE, Hooker SA, Rossom RC. Patient expectations of and experiences with a suicide risk identification algorithm in clinical practice. *BMC Psychiatry* 2022; **22:** 494.

65 Yarborough BJH, Stumbo SP. Patient perspectives on acceptability of, and implementation preferences for, use of electronic health records and machine learning to identify suicide risk. *Gen Hosp Psychiatry* 2021; **70:** 31–37.

66 Jauk S, Kramer D, Avian A, Berghold A, Leodolter W, Schulz S. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. *J Med Syst* 2021; **45:** 48.

67 Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. *JMIR Med Inform* 2016; **4:** e38.

68 Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff* 2014; **33:** 1148–54.

69 Belard A, Buchman T, Forsberg J, et al. Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care. *J Clin Monit Comput* 2017; **31:** 261–71.

70 Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018; **169:** 866–72.

71 Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci* 2021; **4:** 123–44.

72 Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digit Med* 2020; **3:** 99.

73 O'Reilly-Shah VN, Gentry KR, Walters AM, Zivot J, Anderson CT, Tighe PJ. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. *Br J Anaesth* 2020; **125:** 843–46.

74 Qin Y, Alaa A, Floto A, Schaar MV. External validity of machine learning-based prognostic scores for cystic fibrosis: a retrospective study using the UK and Canadian registries. *PLoS Digit Health* 2023; **2:** e0000179.

75 Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK biobank participants. *PLoS One* 2019; **14:** e0213653.

76 Olakotan OO, Mohd Yusof M. The appropriateness of clinical decision support systems alerts in supporting clinical workflows: a systematic review. *Health Informatics J* 2021; **27:** 14604582211007536.

77 European Commission. High-level expert group on artificial intelligence. Ethics guidelines for trustworthy AI. April 8, 2019. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf (accessed Jan 17, 2023).

78 Smith H, Sweeting M, Morris T, Crowther MJ. A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagn Progn Res* 2022; **6:** 10.

79 Halcomb EJ. Mixed methods research: the issues beyond combining methods. *J Adv Nurs* 2019; **75:** 499–501.

80 Parliament UK. Artificial intelligence committee AI in the UK: ready, willing and able? Chapter 8: mitigating the risks of artificial intelligence. 2017. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10012.htm#_idTextAnchor119 (accessed Jan 17, 2023).