BMJ Oncology

Clinical prediction models using machine learning in oncology: challenges and recommendations

Gary S Collins ,[1] Mae Chester-Jones,[2] Stephen Gerry,[2] Jie Ma,[2] Joao Matos,[2] Jyoti Sehjal,[2] Biruk Tsegaye,[2] Paula Dhiman[2]

To cite: Collins GS, Chester-Jones M, Gerry S, *et al*. Clinical prediction models using machine learning in oncology: challenges and recommendations. *BMJ Oncology* 2025;**4**:e000914. doi:10.1136/bmjonc-2025-000914

Received 20 June 2025
Accepted 22 September 2025

[1]Department of Applied Health Sciences, University of Birmingham, Birmingham, UK
[2]University of Oxford, Oxford, UK

Correspondence to
Professor Gary S Collins;
g.s.collins@bham.ac.uk

ABSTRACT

Clinical prediction models are widely developed in the field of oncology, providing individualised risk estimates to aid diagnosis and prognosis. Machine learning methods are increasingly being used to develop prediction models, yet many suffer from methodological flaws limiting clinical implementation. This review outlines key considerations for developing robust, equitable prediction models in cancer care. Critical steps include systematic review of existing models, protocol development, registration, end-user engagement, sample size calculations and ensuring data representativeness across target populations. Technical challenges encompass handling missing data, addressing fairness across demographic groups and managing complex data structures, including censored observations, competing risks or clustering effects. Comprehensive internal and external evaluation requires assessment of both statistical performance (discrimination and calibration) and clinical utility. Implementation barriers include limited stakeholder engagement, insufficient clinical utility evidence, a lack of consideration of workflow integration and the absence of post-deployment monitoring plans. Despite significant potential for personalising cancer care, most prediction models remain unimplemented due to these methodological and translational challenges. Addressing these considerations from study design through post implementation monitoring is essential for developing trustworthy tools that bridge the gap between model development and clinical practice in oncology.

INTRODUCTION

Prediction models have long been used to guide clinical decision making and fall broadly into one of two categories: diagnostic or prognostic.[1] Diagnostic prediction models estimate the probability of a specific condition (typically a disease) being present, while a prognostic prediction model estimates the probability of developing a specific health outcome over a specific time period. One of the primary benefits of multivariable prediction models is their superior predictive accuracy compared with simpler risk classification systems or single prognostic factors. By incorporating multiple predictors simultaneously, these models use all available relevant information to generate more precise individualised risk estimates. This approach is particularly important in oncology, where disease progression and treatment response are influenced by numerous interacting factors.

Cancer care is inherently prediction-driven, with countless clinical decisions, from screening and prevention to treatment selection and palliative care transitions, all fundamentally dependent on assessing the likelihood of future events.[2] Traditional approaches to prediction in oncology have heavily relied on categorical risk classifications such as cancer staging systems, which assign patients to broad risk groups based on limited clinical variables. However, these approaches are limited in their ability to capture the complex and multifactorial nature of cancer progression. Prediction modelling offers a more nuanced, data-driven approach to individualised risk estimation that better addresses the inherent complexities of cancer care.

Many prediction models have been developed for various cancers, including kidney,[3] breast,[4] oral,[5] gastric,[6] colorectal,[7] endometrial,[8] pancreas,[9] melanoma,[10] head and neck,[11] lung,[12] oesophageal[13] and cervical.[14] Well-known examples, within the field of oncology, include models such as the Gail model to estimate the 5-year and lifetime breast cancer risk,[15] the PREDICT tool for estimating survival following surgery for invasive breast cancer,[16] and the Prostate Biopsy Collaborative Group model for estimating the risk of high-grade prostate cancer on biopsy.[17]

Evidence demonstrates that multivariable clinical prediction models consistently outperform simpler risk classification methods across various cancer types.[18] For example, greater accuracy has been found using models than clinical staging alone for

predicting disease-specific survival in gastric cancer,[19] or better patient outcomes than pathological stage.[20] This improved accuracy translates directly to better clinical decision-making by providing more reliable information about individual patient risks and potential benefits of different treatment options. Another advantage of prediction models is their ability to incorporate novel predictors such as genomic data, imaging features and other emerging biomarkers. As our understanding of cancer biology advances, prediction models provide a structured framework for evaluating and integrating new predictors. The flexible nature of these models allows for regular refinement as new evidence emerges. When a promising new biomarker is identified, it can be evaluated for its incremental predictive value within existing models.[21] This process, in principle, ensures that clinical prediction incorporates the most current scientific knowledge while maintaining statistical rigour.

Clinical prediction models have traditionally been developed using regression approaches. For example, logistic regression is typically used to predict binary outcomes (such as the presence or absence of cancer) and the Cox proportional hazards model for time-to-event outcomes (such as survival or recurrence).[22] These statistical methods integrate multiple predictors to generate a mathematical formula that calculates individualised risk estimates. More recently, machine learning approaches have expanded the modeller's toolkit for developing prediction models, offering potential advantages in handling complex, multimodal and non-linear relationships between predictors and outcomes.[23] These techniques range from decision trees and random forests to neural networks and deep learning methods, each with specific strengths, complexities and limitations for clinical prediction.

Despite their potential benefits to improve care and outcomes for oncology patients, the development, evaluation and implementation of clinical prediction models face several challenges that can limit their implementation. There is overwhelming evidence that many prediction models, including those involving machine learning, suffer from poor design and methods,[24–26] incomplete reporting,[27 28] are at high risk of bias[29 30] and exhibit overinterpretation.[31 32] In this article, we outline some key steps researchers should consider when developing or validating a clinical prediction model (table 1).

DO WE NEED YET ANOTHER PREDICTION MODEL?

There are often many published models developed for the same purpose, for example, >900 models for supporting breast cancer decision-making (including >400 for predicting mortality and >200 for predicting recurrence),[33] or the >100 prognostic models for predicting overall survival in patients with gastric cancer.[34] Therefore, before developing a new model, consider conducting a (systematic) review of the literature to identify,[35 36] critically appraise[37] and, where appropriate, evaluate and update any existing and promising models.[38] If and only if none of the existing models can be used in the target setting, consider developing a new model. The findings from carrying out a review of existing models can also help inform model development, that is, identifying commonly used predictors, or ensuring no flaws or critical shortcomings in existing models are repeated.

END-USER INVOLVEMENT

At the outset, end-user involvement, including clinicians, patients and the public, is critical in developing a clinical prediction model to ensure it is relevant, usable and trustworthy in real-world settings.[39] Engaging clinicians helps clarify the clinical questions the model must address, informs the selection of meaningful predictors and guides how predictions will be integrated into clinical workflows, which is essential for implementation and impact.[40]

Involving patients and the public ensures that models reflect lived experiences, prioritise outcomes that matter most to those affected and address concerns about data use, fairness and transparency.[41] This collaborative approach also improves the accuracy, understandability and practical utility of prediction models, as end users can provide feedback on whether outputs are actionable and aligned with their needs. Ultimately, end-user involvement in determining whether a model is actually needed in a specific population and setting. Their involvement throughout model development enhances the likelihood that prediction tools will be effectively implemented, ethically sound and capable of improving patient care and decision-making.

PROTOCOL AND REGISTRATION

A study protocol is an essential document for any research project, detailing its purpose, design, data collection, analysis methods, patient and public involvement, data/code sharing, dissemination plans and overall organisation.[42] For certain studies, like randomised controlled trials, protocols are mandatory for funding, ethics approval and publication. In the context of machine learning in healthcare, where prediction model studies are rapidly increasing and systematic reviews often highlight methodological flaws, protocols are vital for reducing research waste and improving study quality. Developing a protocol forces researchers to plan thoroughly, reducing future problems and enhancing transparency, reproducibility and research integrity. Publicly available protocols also allow readers to compare planned versus actual methods and results and facilitate peer review. We acknowledge, however, that in the case of machine learning or other adaptive modelling approaches, a protocol may necessarily specify areas of uncertainty about the final modelling strategies to be employed. Rather than undermining protocol registration, this openness provides readers with

Table 1 Key considerations when developing and evaluating the performance of a clinical prediction model: challenges and recommendations

Step	Challenges	Recommendations
Clinical purpose and end-user involvement	Lack of early and meaningful stakeholder engagement may lead to models that are clinically irrelevant or misaligned with real-world workflows. Duplication of effort and model proliferation without added value.[80]	Engage end-users (eg, healthcare professionals, patients) early to align the model with a clinical need, the decision the model is intended to support, target population and setting. Avoid redundancy by reviewing and comparing (head-to-head) existing models. Consider updating any existing model before developing a new model.
Protocol and registration	Absence of protocol registration reduces transparency, increasing the risk of selective reporting and methodological inconsistency.	Register the study, for example, on clinicaltrials.gov, and prepare a study protocol, describing all aspects of the model development and evaluation, and make this publicly available.[42]
Study design and data	Retrospective, repurposed or non-representative data may compromise the validity and generalisability of the model.	Ideally, use designs with prospective data collection (cross-sectional for diagnostic, longitudinal for prognostic models). Data should be representative of the target population and setting of intended use.
Sample size	Calculating the sample size is required for developing clinical prediction models,[44] but this is rarely done in the literature.[26] Small sample sizes lead to unstable models (*development*) and unreliable performance metrics (*evaluation*).	*Model development*: Ensure the sample size is sufficiently large enough to develop a model so that overfitting is minimised.[44] *Model evaluation*: Ensure the sample size is sufficiently large enough to reliably and precisely estimate model performance.[45]
Missing data	Inadequate handling of missing data may introduce bias, reduce model generalisability and hinder clinical uptake.	Avoid excluding individuals with incomplete data; instead, explore methods for imputing those missing values. Carefully consider whether missing data will be when the model is used in practice and plan how to handle those missing values during both model development and evaluation.
Fairness	Biases in data or analysis can lead to unfair predictions, potentially creating or worsening existing health inequities.	Ensure aspects of fairness are embedded and addressed, for example, through diverse stakeholder involvement, design, data curation, analysis and reporting.[64]
Model complexities	Ignoring data-specific complexities may produce misleading inferences and limit clinical utility.	Ensure complexities in the data are appropriately analysed, for example, censored observations, competing risks,[67] clustering[81] and recurrent events.[82]
Model evaluation	Reliance on internal validation alone (using model development data) may mask poor generalisability. Limited external validation reduces trust and real-world applicability.	*Internal validation*: use bootstrapping or cross-validation. *Internal-external validation*: when possible (eg, large, clustered data), explore heterogeneity in model performance, where a prediction model is iteratively developed on data from multiple subsets (eg, hospitals, geographic locations) and validated on the remaining excluded subsets. *External validation*: evaluate the model in new data to assess generalisability.[74]
Model stability	The larger the instability in model outputs, the greater the threat that the developed model has poor internal or external validity.	Consider stability checks during model development to ensure predictions are reproducible and trustworthy.[83]

Continued

Table 1 Continued

Step	Challenges	Recommendations
Model performance	Selected performance metrics are not always proper or centred on patient outcomes, with relevance in clinical scenarios.	Assess model discrimination (eg, c-statistic), calibration (calibration plots) and clinical utility (net benefit).[75]
Translation and implementation barriers	Failure to account for implementation barriers (eg, interoperability, clinician burden, automation bias[84]) may limit real-world impact.	Consider how the model will be used in practice and how best to facilitate adoption.[85]
Post-deployment monitoring	Models need to be monitored after deployment to ensure they maintain their performance over time.	Include provisions for ongoing evaluation and refinement, consider statistical performance, practical utility and clinical impact.
Reporting	Incomplete or inconsistent reporting hinders trust, reproducibility, study comparison and critical appraisal of models.	Transparently report all aspects of the development and validation of the clinical prediction model by following the TRIPOD+AI reporting guideline.[78]

TRIPOD+AI, Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence.

important context to assess the planned flexibility and the rigour of the process.

While there is currently no mandatory requirement to register machine learning-driven prediction model research, registering can reduce selective reporting and inflated performance claims, fostering greater trust. Platforms like ClinicalTrials.gov and the Open Science Framework allow registration of studies and depositing of study protocols, with options to embargo details until publication to prevent being scooped. Registration promotes higher standards and trustworthiness in prediction model research.

STUDY DESIGN: SAMPLE SIZE

Study design is crucial, as it guides data collection, the choice of analysis methods and validity of findings. In health research, where findings impact patient care, public health policies and resource allocation, flawed design can lead to misleading conclusions, wasted resources and even harm, if implemented into a workflow. Sample size is a critical design component of all studies and a key consideration in predictive Artificial Intelligence (AI) research. Yet, despite its importance, the data used to train and test predictive AI, even before any data splitting, are often too small and rarely justified.[43]

Sample size calculations are crucial and arguably one of the most important considerations for studies developing and validating clinical prediction models, as inadequate sample sizes can lead to overfitted models with poor performance when applied to new patients. For regression-based approaches, guidance on minimum sample size has been proposed for model development[44] and validation,[45] which have been implemented in R and Stata statistical software 'pmsampsize' and 'pmvalsampsize', respectively. However, for other machine learning methods, sample size requirements often differ substantially and are more difficult to determine.

Machine learning algorithms typically require larger datasets due to their ability to model complex, non-linear relationships and interactions, with requirements varying based on the specific algorithm (eg, random forests, neural networks, support vector machines).[46] Unlike regression methods, many machine learning approaches lack established sample size formulas. Researchers will typically have to rely on simulation studies, learning curves[47] or performance plateaus to gauge whether the sample size is adequate. Some other emerging methods include simulation or Bayesian-driven approaches that estimate minimum sample sizes by examining model performance, degradation and stability.[48–50] However, this methodological gap highlights the need for further research to develop sample size calculation methods for prediction models using machine learning techniques that can align with established guidance available for regression-based prediction models.

STUDY DESIGN: DATA REPRESENTATIVENESS

It is important that the data used to train a clinical prediction model is representative of the target population in whom the model is intended to be used. Representative data ensures that the model accounts for the demographic, clinical and socioeconomic diversity of the population, reducing biases that could lead to inaccurate predictions or inequitable health outcomes for underrepresented groups (eg, ethnic minorities). Without such representativeness, models risk perpetuating disparities, as they will likely fail to generalise to real-world settings. Using existing data which were collected for purposes other than developing a prediction model, out of mere convenience, needs careful consideration as it may not reflect

contemporary target populations, have data quality issues or have not collected information on important predictors. Similarly, using data from a different geographic setting is important, as there is often substantial variation in patient demographics, diagnosis, care, treatment and health outcomes. For example, requiring a model for use in China but using training data from the USA (eg, SEER) needs careful consideration and justification.

Class imbalance refers to differences in the number of individuals with and without the outcome in a dataset. This imbalance is natural, expected and reflects the reality of clinical populations; it is not inherently problematic when developing a prediction model. Nevertheless, methods such as undersampling, oversampling or synthetic approaches like synthetic minority oversampling technique (SMOTE) are often promoted as solutions to supposed concerns about imbalance. These methods artificially manipulate the data to balance outcome frequencies, but in doing so, they distort the underlying distribution and undermine the representativeness of the dataset. Far from being corrective, such approaches are fallacious and can harm the validity and generalisability of the resulting model. We caution against their use in clinical prediction model studies and refer the reader to further discussion of other problems created by 'correcting' for class imbalance.[51]

STUDY DESIGN: DATA PROVENANCE

Detailing data provenance, the documented history of data sources, collection methods and handling procedures, serves as a crucial safeguard against research fraud. Detailed information on provenance enables verification of data authenticity, making fabricated or manipulated data more difficult to introduce undetected. Without detailed data provenance, the scientific community cannot effectively distinguish between genuine clinical findings and manufactured results, potentially leading to flawed models that could endanger patient safety if implemented in clinical practice. As high-profile cases of data fabrication continue to emerge in biomedical research, establishing transparent, auditable data lineage is an essential ethical requirement, ensuring that clinical prediction models are built on foundations of scientific integrity rather than fraudulent data.

MISSING DATA

A carefully designed study is crucial for minimising missing data. Careful planning of data collection procedures, clear definitions of variables and training of data collectors can reduce the occurrence of missing values. However, when missing data inevitably arise, they can introduce bias leading models with poor generalisability. Discarding individuals with any missing values and conducting a 'complete-case' analysis should generally be avoided. It can lead to bias under many missingness mechanisms, reduce sample size, discard important information increasing the risk of overfitting and affect the representativeness of the data, leading to a model with reduced or poor generalisability.

When missing data cannot be avoided, including using existing data which has missing values, it is therefore important to choose an appropriate method depending on the aim and setting. Importantly, the appropriate strategy depends on whether missing data will be permitted at model deployment and if so, how. Techniques such as multiple imputation (replacing missing values with multiple plausible estimates),[52] regression imputation (using fitted models to predict missing values), reference values (eg, mean or median age-sex values) and missing indicator methods offer viable solutions depending on the model that will be used in practice.[53 54] For some modelling approaches, handling missing data is built-in to the model training process (eg, XGBoost).

FAIRNESS

Fairness is increasingly emphasised in prediction model research, with recommendations for it to be considered at multiple stages of model development and evaluation. However, fairness considerations should begin at the stage of problem formulation. Researchers should question whether there are documented health inequities in disease incidence, screening, diagnosis, treatment or outcomes, leading to the essential first step of identifying clinically meaningful dimensions (ie, sensitive attributes) for fairness evaluation.

In oncology, where disease burden, access to care and clinical outcomes frequently differ across demographic and socioeconomic groups, addressing these disparities is critical to avoid exacerbating existing inequities. For example, prostate cancer incidence is higher in black patients.[55] Further, incidence, pathological characteristics and clinical outcomes in breast cancer vary across geographical, racial and ethnic groups, disproportionately affecting both black and Hispanic populations, who also remain underrepresented in breast cancer clinical trials.[56] Data representativeness is foundational. Models like the Khorana Score for estimating the risk of venous thromboembolism in patients with cancer (developed in a predominantly white population) may lack current representativeness.[57] Genomics-driven cancer care faces similar challenges, with databases like The Cancer Genome Atlas overrepresenting European ancestry and underrepresenting Asian, African and Hispanic populations, risking poor generalisability.[58]

Predictor selection also impacts fairness. Excluding sensitive attributes (eg, race/ethnicity) can affect model performance, especially where disease burden differs.[59] While 'race-aware models' might mitigate disparities,[60] race/ethnicity is a complex social construct, inconsistently recorded,[61] and intersects with other identities.[62] Frameworks such as *GUIDE* (Guidance for Unbiased Predictive Information for Healthcare Decision-making

and Equity) help navigate their use, balancing statistical validity and ethics.[59]

Fairness evaluation also requires more than aggregate metrics; population-level assessment can mask subgroup disparities. Stratified evaluation (eg, by cancer type, sex, age, socioeconomic status, race/ethnicity) and visual tools (calibration plots, decision curve analysis) are necessary to detect differential performance and guide mitigation (eg, subgroup-specific thresholds). See online supplemental 1 for an example.[63]

Finally, fairness cannot be achieved without diverse stakeholder engagement. Involving patients, clinicians, caregivers and policymakers is vital to ensure that models are aligned with the values and needs of the populations they are intended to serve.[64]

MODELLING COMPLEXITIES

Developing a clinical prediction model requires accounting for any complex data structures. Common complexities, particularly in the field of oncology, include censored observations, competing risks and clustering. A censored observation occurs when the exact event time (eg, death, progression) is unknown because the patient either exits the study before the event occurs or the event has not yet happened by the end of the observation period. If improperly handled (eg, excluding censored patients, or assuming they did not experience the event), it can lead to incorrect risk estimates and compromise clinical decision making. Common approaches for handling censoring when developing a prediction model include Cox proportional hazards regression, random survival forests or neural networks (eg, DeepSurv[65]).

Competing risks refer to events that preclude the event being predicted (eg, death may preclude disease recurrence) and require specific methods during model development to avoid overestimation of risks[66] and for model validation.[67] Recent examples implementing XGBoost and neural networks for breast cancer prognostication in the presence of competing risks can be found here.[68–70]

Clustering, for example, within hospital or geographical location, can also be important to address when developing a clinical prediction model since observations within the same cluster can be more similar (ie, in characteristics or care) to each other than to those in different clusters, resulting in cluster groups that are distinct yet still related.[71] If clustering is ignored, models may fail to account for any within-group similarities and between-group differences, leading to biased risk estimation, reduced predictive performance and poor calibration. Accounting for any clustering can offer advantages in prediction model research by allowing exploration of heterogeneity in model performance across different clusters.[72] By identifying the sources of this heterogeneity, it may facilitate better adaptation and tailoring of prediction models to specific clusters, ultimately creating models that are more generalisable and effective in diverse settings.

These complexities are not always isolated and can sometimes coexist; therefore, complex strategies are sometimes required to integrate the different components of the modelling approaches that deal with the complexities.

MODEL EVALUATION

Model evaluation is a multistage process that examines a model's predictive performance across different contexts and datasets. Initial assessment typically involves internal validation, using techniques such as bootstrapping or cross-validation to quantify overfitting and estimate performance within the development data. Data splitting is often done, but is not always wise or efficient.[73] When working with large or clustered datasets, internal-external validation can provide further insight by iteratively training the model on data from certain subsets (eg, hospitals or regions) and validating it on excluded subsets. External validation, performed on entirely independent data, is essential for assessing the model's generalisability to new populations and clinical environments.[74]

MODEL PERFORMANCE

Comprehensive evaluation of prediction models should also address multiple dimensions of performance, including statistical validity, clinical utility and implementation feasibility. Statistical validation should assess both discrimination (the model's ability to distinguish between patients who will and will not experience the outcome) and calibration (the agreement between predicted and observed outcomes). Common metrics include the area under the receiver operating characteristic curve (referred to as the area under the receiver operating characteristic curve or c-statistic) for discrimination and calibration curves for calibration.[75] Clinical utility assessment should evaluate whether model-guided decisions lead to improved patient outcomes compared with standard care.[76] This evaluation may involve impact studies that measure changes in clinical decisions, resource utilisation, patient satisfaction and health outcomes associated with model implementation. With the increase in developed AI models, other classification-based measures of model's performance are often reported. However, these should be interpreted with caution as they are often dependent on arbitrary risk thresholds and may lead users to treatment decisions without full consideration of the predicted risk.

TRANSPARENT REPORTING

Transparent reporting is fundamental in the development and evaluation of clinical prediction models, as it enables stakeholders, including clinicians, researchers, patients and policymakers, to critically appraise study design, assess risk of bias and determine the clinical usefulness of a model.[37] Further, where the model itself is not reported,

this impedes the model being evaluated and used in clinical practice. Incomplete or inaccurate reporting can conceal methodological flaws, hinder reproducibility and ultimately compromise patient safety if flawed models are implemented in clinical practice.

Transparent reporting fosters trust, enhances the credibility of research and supports the ethical obligation to communicate findings honestly and completely.[77] Guidelines such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD+AI)[78] provide structured recommendations to ensure comprehensive description of methods, data sources and results, thereby facilitating model evaluation, replication and responsible implementation in healthcare. Ultimately, transparent reporting is not optional but essential for advancing trustworthy, equitable and effective clinical prediction models.

TRANSLATIONAL AND IMPLEMENTATION BARRIERS

Even methodologically sound prediction models face significant barriers to clinical implementation. The gap between model development and clinical use remains substantial, with few models successfully translated into routine practice.[79] Several factors contribute to this translational gap, including:

1. Limited stakeholder engagement during model development.
2. Incomplete or poor statistical evaluation of model performance.
3. Insufficient evidence of clinical utility beyond statistical performance, with largely non-existent evaluation of prediction models as an intervention to guide treatment and improve patient outcomes.
4. No plan or intention to implement that model.
5. Lack of user-friendly interfaces that integrate seamlessly into clinical workflows.
6. Inadequate education and training for potential users.
7. Regulatory and implementation hurdles.
8. Lack of plans or framework for post implementation monitoring.

Addressing these barriers requires a more comprehensive approach to model development that considers implementation factors from the outset, during the design stage, rather than as an afterthought.

POST-DEPLOYMENT MONITORING

Prediction models require ongoing quality assurance and periodic updating to maintain their performance over time. Clinical practices, patient populations and treatment options evolve continuously, potentially affecting the relevance and accuracy of prediction models. Without systematic processes for monitoring and updating, models may become outdated and potentially misleading. The life cycle of a clinical prediction model should include provisions for ongoing evaluation, refinement and, when necessary, replacement. This quality assurance process should assess not only statistical performance but also practical utility and clinical impact. Regulatory and technical frameworks such as the Food and Drug Administration (FDA) Software as a Medical Device principles and operational approaches like Machine Learning Operations (MLOps) further support post-deployment surveillance, enabling structured monitoring, risk management and continuous improvement of prediction models in real-world practice.

CONCLUSIONS

Prediction models, using statistics and increasingly machine learning methods, are often developed for personalised diagnostic and prognostic assessments to aid clinical decision making. However, there are numerous design and methodological challenges, including lack of end-user involvement, small sample sizes, missing data, model unfairness, failure to account for complexities like censoring or competing risks and weak evaluation of model performance. Therefore, despite their potential, many models remain unimplemented. Addressing these challenges will ensure robust, equitable prediction tools that enhance decision-making and outcomes in oncology, bridging the gap between innovation and clinical practice.

Social media Gary S Collins, Bluesky @gscollins.bsky.social

Acknowledgements GSC is a National Institute for Health and Care Research (NIHR) senior investigator. The views expressed in this article are those of the author(s) and not necessarily those of the NIHR, or the Department of Health and Social Care.

Contributors GSC drafted the article, and all authors commented and approved the manuscript. GSC is the guarantor.

Funding GSC is supported by the EPSRC (Engineering and Physical Sciences Research Council) grant for Artificial intelligence innovation to accelerate health research (EP/Y018516/1) and MRC-NIHR Better Methods Better Research grant (MR/Z503873/1). JMat is funded by a Clarendon Fund Scholarship at University of Oxford. PD and BT are supported by CRUK (project grant: PRCPJT-Nov21\100021). MC-J is funded by a National Institute for Health Research (NIHR) doctoral research fellowship (NIHR302985). JS is funded by Cancer Research UK (CANCTA-2023/100008). The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report or decision to submit the article for publication.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Provenance and peer review Commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD
Gary S Collins http://orcid.org/0000-0002-2772-2316

REFERENCES

1 van Smeden M, Reitsma JB, Riley RD, *et al*. Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol* 2021;132:142–5.
2 Vickers AJ. Prediction models in cancer care. *CA Cancer J Clin* 2011;61:315–26.
3 Harrison H, Thompson RE, Lin Z, *et al*. Risk Prediction Models for Kidney Cancer: A Systematic Review. *Eur Urol Focus* 2021;7:1380–90.
4 Louro J, Posso M, Hilton Boon M, *et al*. A systematic review and quality assessment of individualised breast cancer risk prediction models. *Br J Cancer* 2019;121:76–85.
5 Espressivo A, Pan ZS, Usher-Smith JA, *et al*. Risk Prediction Models for Oral Cancer: A Systematic Review. *Cancers (Basel)* 2024;16:617.
6 Xu L, Lyu J, Zheng X, *et al*. Risk Prediction Models for Gastric Cancer: A Scoping Review. *J Multidiscip Healthc* 2024;17:4337–52.
7 Usher-Smith JA, Walter FM, Emery JD, *et al*. Risk Prediction Models for Colorectal Cancer: A Systematic Review. *Cancer Prev Res (Phila)* 2016;9:13–26.
8 Forder BH, Ardasheva A, Atha K, *et al*. Models for predicting risk of endometrial cancer: a systematic review. *Diagn Progn Res* 2025;9:3.
9 Ioannou LJ, Maharaj AD, Zalcberg JR, *et al*. Prognostic models to predict survival in patients with pancreatic cancer: a systematic review. *HPB (Oxford)* 2022;24:1201–16.
10 Kunonga TP, Kenny RPW, Astin M, *et al*. Predictive accuracy of risk prediction models for recurrence, metastasis and survival for early-stage cutaneous melanoma: a systematic review. *BMJ Open* 2023;13:e073306.
11 Smith CDL, McMahon AD, Ross A, *et al*. Risk prediction models for head and neck cancer: A rapid review. *Laryngoscope Investig Otolaryngol* 2022;7:1893–908.
12 Wu Z, Wang F, Cao W, *et al*. Lung cancer risk prediction models based on pulmonary nodules: A systematic review. *Thorac Cancer* 2022;13:664–77.
13 Li H, Sun D, Cao M, *et al*. Risk prediction models for esophageal cancer: A systematic review and critical appraisal. *Cancer Med* 2021;10:7265–76.
14 He B, Chen W, Liu L, *et al*. Prediction Models for Prognosis of Cervical Cancer: Systematic Review and Critical Appraisal. *Front Public Health* 2021;9:654454.
15 Costantino JP, Gail MH, Pee D, *et al*. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst* 1999;91:1541–8.
16 Wishart GC, Azzato EM, Greenberg DC, *et al*. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res* 2010;12:R1.
17 Ankerst DP, Straubinger J, Selig K, *et al*. A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts. *Eur Urol* 2018;74:197–203.
18 Jannello LMI, Morra S, Scheipner L, *et al*. Multivariable model versus AJCC staging system: cancer-specific survival predictions in adrenocortical carcinoma. *Endocr Relat Cancer* 2024;31:e230353.
19 Wang L, Ge J, Fang Y, *et al*. Construction and validation of a novel nomogram based on the log odds of positive lymph nodes to predict cancer-specific survival in elderly patients with gastric adenocarcinoma after radical surgery. *BMC Gastroenterol* 2025;25:215.
20 Vickers AJ, Cronin AM, Kattan MW, *et al*. Clinical benefits of a multivariate prediction model for bladder cancer: a decision analytic approach. *Cancer* 2009;115:5460–9.
21 Steyerberg EW, Vedder MM, Leening MJG, *et al*. Graphical assessment of incremental value of novel markers in prediction models: From statistical to decision analytical perspectives. *Biometrical J* 2015;57:556–70.
22 Royston P, Moons KGM, Altman DG, *et al*. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009;338:338/mar31_1/b604.
23 Hunter DJ, Holmes C. Where Medical Statistics Meets Artificial Intelligence. *N Engl J Med* 2023;389:1211–9.
24 Andaur Navarro CL, Damen JAA, van Smeden M, *et al*. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *J Clin Epidemiol* 2023;154:8–22.
25 Dhiman P, Ma J, Andaur Navarro CL, *et al*. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022;22:101.
26 Dhiman P, Ma J, Qi C, *et al*. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. *BMC Med Res Methodol* 2023;23:188.
27 Dhiman P, Ma J, Navarro CA, *et al*. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol* 2021;138:60–72.
28 Andaur Navarro CL, Damen JAA, Takada T, *et al*. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol* 2022;22:12.
29 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.
30 Andaur Navarro CL, Damen JAA, Takada T, *et al*. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021;375:n2281.
31 Dhiman P, Ma J, Andaur Navarro CL, *et al*. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *J Clin Epidemiol* 2023;157:120–33.
32 Andaur Navarro CL, Damen JAA, Takada T, *et al*. Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models. *J Clin Epidemiol* 2023;158:99–110.
33 Hueting TA, van Maaren MC, Hendriks MP, *et al*. The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias. *J Clin Epidemiol* 2022;152:238–47.
34 Feng Q, May MT, Ingle S, *et al*. Prognostic Models for Predicting Overall Survival in Patients with Primary Gastric Cancer: A Systematic Review. *Biomed Res Int* 2019;2019:5634598:5634598:.
35 Snell KIE, Levis B, Damen JAA, *et al*. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). *BMJ* 2023;381:e073538.
36 Moons KGM, de Groot JAH, Bouwmeester W, *et al*. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014;11:e1001744.
37 Moons KGM, Damen JAA, Kaul T, *et al*. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* 2025;388:e082505.
38 Collins GS, Moons KGM. Comparing risk prediction models. *BMJ* 2012;344:e3186.
39 Lekadir K, Frangi AF, Porras AR, *et al*. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 2025;388:e081554.
40 Hassan N, Slight R, Morgan G, *et al*. Road map for clinicians to develop and evaluate AI predictive models to inform clinical decision-making. *BMJ Health Care Inform* 2023;30:e100784.
41 Markham S. Patient perspective on predictive models in healthcare: translation into practice, ethical implications and limitations? *BMJ Health Care Inform* 2025;32:e101153.
42 Peat G, Riley RD, Croft P, *et al*. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. *PLoS Med* 2014;11:e1001671.
43 Tsegaye B, Snell KIE, Archer L, *et al*. Larger sample sizes are needed when developing a clinical prediction model using machine learning in oncology: methodological systematic review. *J Clin Epidemiol* 2025;180:111675.
44 Riley RD, Ensor J, Snell KIE, *et al*. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.
45 Riley RD, Snell KIE, Archer L, *et al*. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ* 2024;384:e074821.
46 van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14:137.
47 Christodoulou E, van Smeden M, Edlinger M, *et al*. Adaptive sample size determination for the development of clinical prediction models. *Diagn Progn Res* 2021;5:6.
48 Sadatsafavi M, Gustafson P, Setayeshgar S, *et al*. Bayesian sample size calculations for external validation studies of risk prediction models. 2025.

49 Riley RD, Whittle R, Sadatsafavi M, *et al*. A general sample size framework for developing or updating a clinical prediction model. 2025.
50 Kalaycıoğlu O, Pavlou M, Akhanlı SE, *et al*. Evaluating the sample size requirements of tree-based ensemble machine learning techniques for clinical risk prediction. *Stat Methods Med Res* 2025;34:1356–72.
51 van den Goorbergh R, van Smeden M, Timmerman D, *et al*. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc* 2022;29:1525–34.
52 White IR, Royston P. Imputing missing covariate values for the Cox model: IMPUTING MISSING COVARIATE VALUES FOR THE COX MODEL. *Statist Med* 2009;28:1982–98.
53 Sperrin M, Martin GP, Sisk R, *et al*. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol* 2020;125:183–7.
54 Sisk R, Sperrin M, Peek N, *et al*. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Stat Methods Med Res* 2023;32:1461–77.
55 Ehdaie B, Carlsson S, Vickers A. Racial Disparities in Low-Risk Prostate Cancer. *JAMA* 2019;321:1726–7.
56 Hirko KA, Rocque G, Reasor E, *et al*. The impact of race and ethnicity in breast cancer-disparities and implications for precision oncology. *BMC Med* 2022;20:72.
57 Ghassemi M, Gusev A. Limiting bias in AI models for improved and equitable cancer care. *Nat Rev Cancer* 2024;24:823–4.
58 Dankwa-Mullan I, Weeraratne D. Artificial Intelligence and Machine Learning Technologies in Cancer Care: Addressing Disparities, Bias, and Data Diversity. *Cancer Discov* 2022;12:1423–7.
59 Ladin K, Cuddeback J, Duru OK, *et al*. Guidance for unbiased predictive information for healthcare decision-making and equity (GUIDE): considerations when race may be a prognostic factor. *NPJ Digit Med* 2024;7:290.
60 Zink A, Obermeyer Z, Pierson E. Race adjustments in clinical algorithms can help correct for racial disparities in data quality. *Proc Natl Acad Sci U S A* 2024;121:e2402267121.
61 Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. *N Engl J Med* 2020;383:874–82.
62 Nielsen MW, Gissi E, Heidari S, *et al*. Intersectional analysis for science and technology. *Nature New Biol* 2025;640:329–37.
63 Overvad TF, Ording AG, Nielsen PB, *et al*. Validation of the Khorana score for predicting venous thromboembolism in 40 218 patients with cancer initiating chemotherapy. *Blood Adv* 2022;6:2967–76.
64 Wawira Gichoya J, McCoy LG, Celi LA, *et al*. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021;28:e100289.
65 Katzman JL, Shaham U, Cloninger A, *et al*. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018;18:24.
66 Kantidakis G, Putter H, Litière S, *et al*. Statistical models versus machine learning for competing risks: development and validation of prognostic models. *BMC Med Res Methodol* 2023;23:51.
67 van Geloven N, Giardiello D, Bonneville EF, *et al*. Validation of prediction models in the presence of competing risks: a guide through modern methods. *BMJ* 2022;377:e069249.
68 Clift AK, Dodwell D, Lord S, *et al*. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. *BMJ* 2023;381:e073800.
69 Wynants L, Collins G, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG* 2017;124:423–32.
70 Clift AK, Collins GS, Lord S, *et al*. Predicting 10-year breast cancer mortality risk in the general female population in England: a model development and validation study. *Lancet Digit Health* 2023;5:e571–81.
71 de Jong VMT, Moons KGM, Eijkemans MJC, *et al*. Developing more generalizable prediction models from pooled studies and large clustered data sets. *Stat Med* 2021;40:3533–59.
72 Riley RD, Ensor J, Snell KIE, *et al*. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140.
73 Steyerberg EW, Harrell FE Jr, Borsboom GJ, *et al*. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;54:774–81.
74 Collins GS, Dhiman P, Ma J, *et al*. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024;384:e074819.
75 Calster B, Collins GS, Vickers AJ, *et al*. Performance evaluation of predictive AI models to support medical decisions: Overview and guidance. *arXiv* 2025;10288.
76 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;352:i6.
77 World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants. *JAMA* 2025;333:71–4.
78 Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378.
79 Craddock M, Crockett C, McWilliam A, *et al*. Evaluation of Prognostic and Predictive Models in the Oncology Clinic. *Clin Oncol* 2022;34:102–13.
80 Moher D, Glasziou P, Chalmers I, *et al*. Increasing value and reducing waste in biomedical research: who's listening? *The Lancet* 2016;387:1573–86.
81 Wynants L, Vergouwe Y, Van Huffel S, *et al*. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. *Stat Methods Med Res* 2018;27:1723–36.
82 Bonnett LJ, Spain T, Hunt A, *et al*. Guide to evaluating performance of prediction models for recurrent clinical events. *Diagn Progn Res* 2025;9:6.
83 Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. 2022.
84 Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012;19:121–7.
85 Bonnett LJ, Snell KIE, Collins GS, *et al*. Guide to presenting clinical prediction models for use in clinical settings. *BMJ* 2019;365:l737.