Journal of Clinical Epidemiology 180 (2025) 111675

ORIGINAL RESEARCH

Larger sample sizes are needed when developing a clinical prediction model using machine learning in oncology: methodological systematic review

Biruk Tsegaye[a,*], Kym I.E. Snell[b,c], Lucinda Archer[b,c], Shona Kirtley[a], Richard D. Riley[b,c], Matthew Sperrin[d], Ben Van Calster[e,f,g], Gary S. Collins[a], Paula Dhiman[a]

[a]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK*

[b]*Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK*

[c]*Institute of Translational Medicine, National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, UK*

[d]*Division of Imaging, Informatics and Data Science, Manchester Academic Health Science Centre, University of Manchester, Manchester M13 9PL, UK*

[e]*Department of Development and Regeneration, KU Leuven, Leuven, Belgium*

[f]*Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, The Netherlands*

[g]*Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium*

Accepted 7 January 2025; Published online 13 January 2025

Abstract

Background and Objectives: Having a sufficient sample size is crucial when developing a clinical prediction model. We reviewed details of sample size in studies developing prediction models for binary outcomes using machine learning (ML) methods within oncology and compared the sample size used to develop the models with the minimum required sample size needed when developing a regression-based model (N_{min}).

Methods: We searched the Medline (via OVID) database for studies developing a prediction model using ML methods published in December 2022. We reviewed how sample size was justified. We calculated N_{min}, which is the N_{min}, and compared this with the sample size that was used to develop the models.

Results: Only one of 36 included studies justified their sample size. We were able to calculate N_{min} for 17 (47%) studies. 5/17 studies met N_{min}, allowing to precisely estimate the overall risk and minimize overfitting. There was a median deficit of 302 participants with the event ($n = 17$; range: −21,331 to 2298) when developing the ML models. An additional three out of the 17 studies met the required sample size to precisely estimate the overall risk only.

Conclusion: Studies developing a prediction model using ML in oncology seldom justified their sample size and sample sizes were often smaller than N_{min}. As ML models almost certainly require a larger sample size than regression models, the deficit is likely larger. We recommend that researchers consider and report their sample size and at least meet the minimum sample size required when developing a regression-based model.

Keywords: Machine learning; Oncology; Sample size; Prediction model; Systematic review; Methodology

Registration: The study protocol was registered with PROSPERO (CRD42023394388) and Open Science Framework (osf.io/kce76/).

Funding: Biruk Tsegaye and Paula Dhiman are supported by Cancer Research UK (project grant: PRCPJT-Nov21\100021). Gary Collins and Shona Kirtley are supported by Cancer Research UK (programme grant: C49297/A27294). Biruk Tsegaye is also supported by the NIHR Blood and Transplant Research Unit in Data Driven Transfusion Practice (NIHR203334). Gary Collins, Richard Riley, Lucinda Archer, Kym Snell and Paula Dhiman are supported by an EPSRC grant (number: EP/Y018516/1). Kym Snell, Lucinda Archer and Richard Riley are supported by funding from the NIHR Birmingham Biomedical Research Centre (BRC) at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. Ben Van Calster is supported by the Research Foundation − Flanders (grant: G097322 N) and Internal funds KU Leuven (grant: C24 M/20/064). This publication presents independent research funded by Cancer Research UK, and the National Institute for Health Research (NIHR). GSC and RDR are National Institute for Health and Care Research (NIHR) Senior Investigators. The views expressed are those of the author(s) and not necessarily those of the Cancer Research UK, the NHS, the NIHR or the Department of Health and Social Care.

* Corresponding author. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK.

E-mail address: biruk.tsegaye@csm.ox.ac.uk (B. Tsegaye).

https://doi.org/10.1016/j.jclinepi.2025.111675

What is new?

Key findings

- A sample size calculation or justification was only reported in one of the included studies.
- 17/36 (47%) studies reported enough information for the minimum required sample size to be calculated, of which five studies met the recommended minimum sample size to precisely estimate the overall risk and minimize overfitting, and an additional three studies only met the recommended minimum sample size to precisely estimate the overall risk for regression-based approaches.
- Studies developing a prediction model using machine learning (ML) were a median 302 events smaller than the minimum required sample size for regression-based approaches.

What this adds to what is known?

- We build on existing research that has highlighted poor reporting of sample size calculations for prediction models developed using regression approaches and found similar results in studies developing prediction models using ML methods.
- Our study provides evidence that sample size calculations are rarely considered in ML studies, and that the actual sample sizes are often much smaller than minimum sample size requirements even for regression-based approaches.

What is the implication and what should change now?

- Formal sample size guidance is needed for studies that develop and validate a prediction model using ML methods. Until guidance is available, sample size calculations for regression-based models provide a suitable lower bound for any ML method.
- We urge researchers to consider and justify their sample size at the design and protocol stage of their study.
- Researchers should transparently report their available and used sample sizes (total number and number of events) throughout their study flow (eg, before and after data splitting or modifications to address class imbalance).

1. Introduction

Prediction models are used in health care for the diagnosis and prognosis of health-related outcomes. Though clinical prediction models are developed in many clinical areas, they are frequently developed and used in oncology to help diagnose cancer [1,2], assess the risk of developing cancer in the future [3], and determine patient prognosis after diagnosis or treatment [4,5]. Cancer prediction models are often used to plan health-care delivery, using patients' predicted risks to help determine treatment plans or the need for further tests. They are used in clinical practice as well as for research purposes.

Use of machine learning (ML) methods to predict cancer related outcomes has increased over recent years and continues to rise given their potential to model complex datasets [6,7]. ML is often used with a view to improve the accuracy of predictions (frequently over statistical approaches) and to help guide clinical decision making. However, many studies have found inherent limitations in the studies that develop ML prediction models in oncology [8−11], and specifically with the too-small amount of data that is used to develop them. Dhiman et al. found that only five (8%) of 62 studies developing prediction models using ML in oncology justified their sample size [8], but as sample size guidance does not exist for ML models, details of the 'deficit' in sample size remains unclear.

ML models are complex and known to be data hungry [12]. They typically require larger sample sizes than regression-based approaches as they, by default, are more flexible. If enough data is not used to develop them, it can increase instability in model predictions [13]. Though there is a lack of sample size guidance for ML models, guidance is available for regression-based prediction models [14,15], including when predicting binary outcomes, based on precisely estimating the overall risk and minimizing overfitting. As both criteria are applicable to ML, these sample size calculations provide a useful lower bound for the required sample size for ML models.

The aim of this study was to review sample size calculations or justifications reported in studies developing a prediction model using ML methods for a binary outcome in the field of oncology. We investigated whether and how sample size calculations are reported or justified. We also compared the reported sample size with the minimum required sample size needed when developing a regression-based model minimum recommended for regression-based approaches (N_{min}), as calculated using the Riley et al. formulae [16].

2. Methods

We carried out a methodological systematic review of sample size calculations or justifications for studies developing a ML prediction model in oncology. The study protocol was registered with The International Prospective Register of Systematic Reviews (CRD42023394388). The review is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 [17] and PRISMA-Search [18] checklists.

2.1. Eligibility

We included studies that described the development of a multivariable (including at least two predictors) prediction model (diagnostic or prognostic) using any standard (non-regression based) ML modeling method, in oncology for individualized risk predictions in humans. Though we excluded studies that developed only regression-based models, such as logistic regression alone, studies were included if ML models were also developed in the same study. Studies could use any study design (eg, case-control, cohort and registry studies, and randomized trials) to develop a model predicting a binary patient-related health outcome. We included studies that developed a model using clinical predictors alone, or in combination with radiomic features extracted from imaging datasets, or in combination with genetic predictors.

We excluded studies that developed a prediction model using only regression methods (eg, least absolute shrinkage and selection operator [LASSO] logistic regression), that were predicting a time-to-event outcome, and were (external) validation only studies (without a model development component). We excluded studies developing a model using only lab-based (eg, in vitro), molecular, imaging and genetics predictors/features (indicating high dimensional data − data with an extensive number of features or predictors that are also often correlated). We also excluded conference abstracts and reviews and studies where the full text was unavailable (authors were not contacted) or published in a non-English language.

2.2. Information sources and search strategy

We searched the MEDLINE® ALL (via Ovid) database on January 24, 2023, using a search strategy developed in consultation with a senior information specialist (SK). We limited the search to one database to obtain a convenience sample of studies to review. A publication date limit for articles published between December 1, 2022, and December 31, 2022, was applied to the search (including Epub Ahead of Print and In-Process & Other Nonindexed Citations articles) to obtain a contemporary sample of studies.

Search terms included Medical Subject Headings headings and free-text terms related to ML (eg, ‘‘supervised’’, ‘‘classification’’, ‘‘boosting’’, ‘‘ensemble’’), prediction (eg, ‘‘risk’’, ‘‘prognosis’’, ‘‘probability’’), model performance (eg, ‘‘accuracy’’, ‘‘discrimination’’, ‘‘AUC’’, ‘‘decision curve’’, ‘‘validation’’, ‘‘calibration’’) and oncology (eg, ‘‘neoplasm’’, ‘‘oncology’’, ‘‘carcinoma’’, ‘‘adenocarcinoma’’, ‘‘malignant’’, ‘‘metastasis’’). The terms within each of the four search facets (ML, prediction, model performance, oncology) were combined with ‘OR’ and then the facets were combined with ‘AND’. No other limits were applied to the literature search. The full search strategy is provided in Supplementary Table 1.

2.3. Data management and selection process

The references identified by the MEDLINE (OVID) search (including abstracts and titles) were imported as a complete reference into EndNote Citation Manager [19] where they were deduplicated. The remaining records were then uploaded to Rayyan Web Application [20] for title and abstract screening by two reviewers (BT and PD) in line with the eligibility criteria.

The same reviewers then assessed the full text publications for studies that met the inclusion criteria. Any uncertainties and disagreements throughout the screening processes were discussed and adjudicated through another independent reviewer (GSC). Reasons for the exclusion of studies were recorded during full text screening.

2.4. Data extraction form and data collection

The data extraction form was informed by the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies [21] and Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) checklists [22]. We also extracted necessary information to calculate the minimum required sample size based on formulae by Riley et al for binary outcome measures (described below) [16].

The data extraction form was piloted on three randomly selected studies to achieve consistency on the data extraction among the reviewers. Three reviewers (BT, KS, and LA) then independently extracted data using the standardized data extraction form. BT extracted from all the articles included, and KS and LA were each allocated half of the studies. Any conflicts through the extraction processes were discussed and adjudicated with another independent reviewer (PD).

If more than one model was developed per study, we noted the total number of developed models, but only extracted data on one model based on the following order of preference:

1. The model that was suggested as the best model to predict the outcome of interest by the study authors,
2. The model with best performance (highest c-statistic) compared to other models, or
3. The first reported model in the abstract, results, or discussion sections of the paper.

If a study develops multiple models and includes a regression-based model, we extracted on the best performing ML model, excluding the performance of the regression-based model.

2.5. Data items and outcome

Data extraction items included study characteristics (cancer type, study design, data source, prognostic or diagnostic outcome type), details on any sample size calculations or justifications, study type (model development

only, or development and validation), type of ML approaches or algorithms used for the developed model, number of models developed, total number predictors and predictor parameters considered during the modeling process, sample sizes (total participants available, and number used for model development), number of outcome events (total number of events, and number of events used for model development) and model performance measures (for example, c-statistic/area under the receiver operating characteristics curve, calibration-in-the-large, calibration slope). We also collected detail about the model performance measures and if they were apparent, split sample, bias corrected or external validation performance measures, depending on the validation type. We judged whether a study outcome was prognostic or diagnostic based on the timing of the outcome and the time lapsed between predictor and outcome measurement.

When available we used the extracted information to calculate the minimum sample size to develop a regression-based (N_{min}) model using Riley et al formulae, which is based on the following three criteria:

1. Small overfitting (where a developed model's predictions are more extreme than they should be in new individuals from the same target population) - defined by an expected global shrinkage (penalization/regularization) of predictor effects by 10% or less,
2. Small optimism ($\leq$0.05) in the model's apparent Nagelkerke's R-squared value (a measure of overall model fit).
3. Precise estimation (within a small margin of error) of the average outcome risk in the target population,

A sample size was calculated to meet each criterion and the largest of the three sample sizes was taken as N_{min}. The calculation was implemented using the 'pmsampsize' package [23], which requires specification of the outcome type (binary for all cases), Cox-Snell R-squared (calculated from the c-statistic, if not reported [24]), number of predictor parameters, and the outcome prevalence. The value of the c-statistic used for the sample size calculation was chosen in order of preference, from the reported c-statistic from (i) bias-corrected (derived from either bootstrapping or cross-validation methods) or (ii) split sample, where multiple values were reported. An example scenario for a typical sample size calculation is provided in Box 1.

No formal analysis on reporting adherence was conducted; however, we evaluated whether any statement about sample size was reported (including a calculation or justification) by original authors and the reporting quality of these sample size statements, if provided. The total available sample size and sample size used to develop the prediction models was compared to N_{min}, calculated using recommended formulae by Riley et al. [16]. The difference between the actual sample size used and N_{min} was calculated and reported for each sample size criteria and overall. Studies that met N_{min} to precisely estimate the overall risk and minimize overfitting (criteria 1−3), only precisely estimate the overall risk (criterion 3), and to minimize overfitting (criteria 1 and 2) were identified.

Box 1 Implementing the Riley et al. minimum sample size criteria for model development: an example scenario

Consider a study wanting to develop a prediction model using a random forest to predict an incident lung cancer diagnosis in a 'healthy' population. Previous literature estimates that the prevalence of lung cancer in the population is 5% and the anticipated c-statistic is 0.68. The study aims to predict incident lung cancer using data on 37 candidate predictors.

Of these 37 candidate predictors:

- 20 are binary (equating to 20 predictor parameters)
- 7 are categorical with three categories each (equating to 14 predictor parameters)
- 5 are continuous and have a linear relationship with the outcome (equating to five predictor parameters); and
- 5 are continuous and have a non-linear relationship with the outcome, such that each introduces at most one additional term in the modeling process (equating to 10 predictor parameters).

Thus, in total, a minimum of 49 candidate predictor parameters will be estimated during the modeling process.

For this scenario and using the 'pmsampsize' statistical package available in R and Stata R code:

pmsampsize (type = "b", cstatistic = 0.68, parameters = 49, prevalence = 0.05) to estimate the overall risk precisely and minimize overfitting, the minimum sample size required for developing the new model is 21,586 participants (1080 incident lung cancer events), with an events per predictor parameter of about 22.

2.6. Data analysis

We summarized results of the review using descriptive statistics and a narrative synthesis. Numbers and percentages are used to describe categorical data, median, 25th and 75th percentiles and range for continuous data. The difference between the actual and N_{min} was calculated and visually presented using a scatterplot. The scatterplot also indicated studies that met the N_{min} to precisely estimate the overall risk and minimize overfitting (criteria 1−3), studies that met the N_{min} to only precisely estimate the overall risk (criterion 3), and studies that met the N_{min} to minimize overfitting (criteria 1 and 2). Data analysis was performed using R (version: R.4.3.1) statistical software package [25].

2.7. Sensitivity analysis

As we calculate post hoc sample size calculations for the included studies and use their reported results to estimate their N_{min}, our sample size estimates may be optimistic. We, therefore, also conducted a sensitivity analysis using the reported outcome prevalence and the number of candidate predictors for the included studies and took this information with 15% of the maximum Cox-Snell R^2 possible for each prediction scenario − a recommended approach when the anticipated Cox-Snell R^2 or c-statistic is unavailable [16].

3. Results

The search identified 175 published articles on Ovid MEDLINE (R) ALL between December 1, 2022, and December 31, 2022. After title and abstract screening, 126 studies were excluded leaving 49 studies for full text screening. After full text screening, 36 studies were eligible and included in the review for data extraction. A flowchart of the study selection is provided in Figure 1.

3.1. Study characteristics

Of the 36 included studies, 28 were model development only (78%; Table 1) [26−53] and eight were development with validation (22%) [54−61]. Over half of the studies were diagnostic (n = 19/36; 53%) [26,27,30,32,38−42,48,49,51,53−56,58,60,61]. Most studies used a retrospective cohort study design (n = 31/36, 86%) [28−31,34−39,41−61] and half of the studies used electronic health records (n = 18/36, 50%) [27,32,36−39,41,42,44,46,48−50,52,53,57,59,60].

A median of eight models (lower quartile to upper quartile: 4−16; range: 1−64) were developed per study and only two studies developed a single model [30,42]. Of the 34 studies that developed multiple models, 30 studies [26−28,31−36,39−41,43−47,49−61] used different modeling methods (30/34; 88%), including 11 studies that also used different predictor sets [26,27,36,39,41,46,50−52,58,59], and three studies that also developed models for different outcomes [35,56,61]. Two studies developed different models using only different predictor sets [29,38], one study reported three separate Light Gradient-Boosting Machine framework forecasting models depending on different classification thresholds [48], and one study developed models with different predictor sets and predicting multiple outcomes [37]. Random forests were used to develop models in most studies (n = 26/36, 72%) [26−29,32,34−39,43−47,51−59,61] and was the commonly developed model type (n = 26/133 models; Supplementary Table 2).

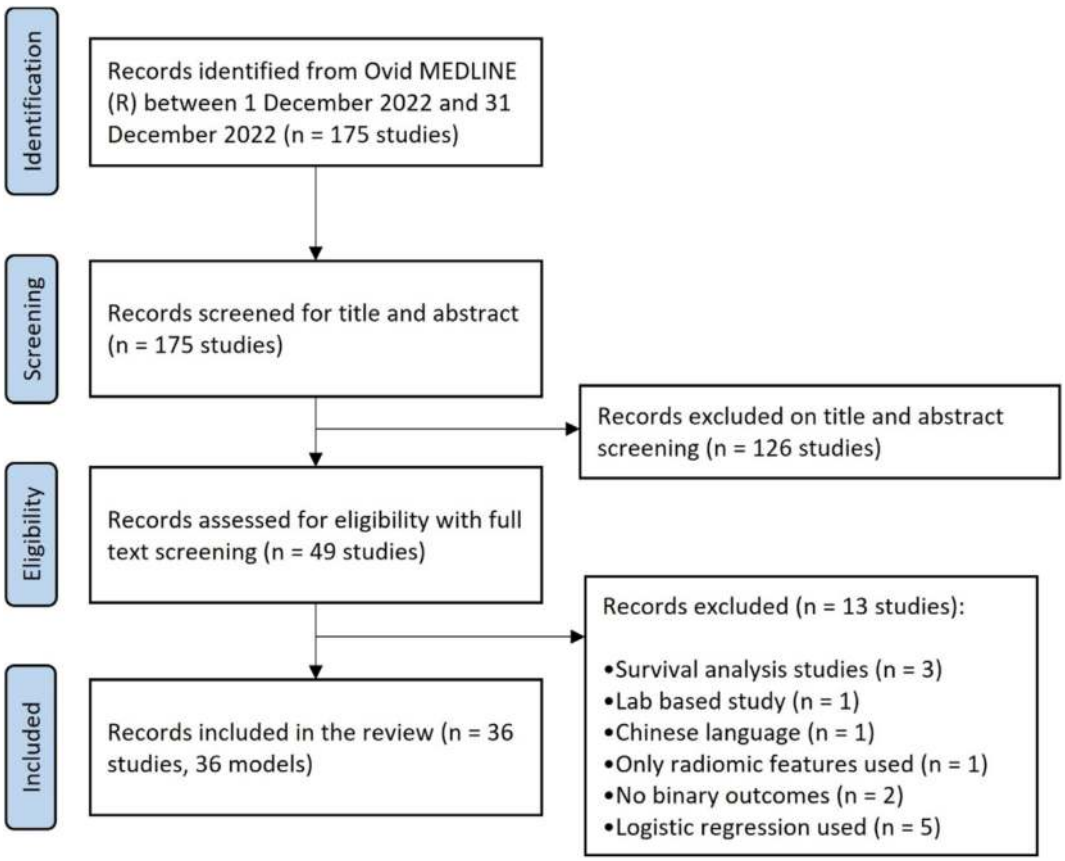

Figure 1. Studies selection flowchart (PRISMA). PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Table 1. Study characteristics for the 36 included studies

Characteristics	*n* (%) studies (*N* = 36)
Study type	
Development only	28 (77.8)
Development and validation	8 (22.2)
Study design	
Retrospective cohort	31 (86.0)
Case-control	2 (5.6)
Prospective cohort	1 (2.8)
Cross-sectional	1 (2.8)
Unclear	1 (2.8)
Data source	
EHR	19 (52.8)
SEER	6 (16.7)
Other[a]	11 (30.6)
Model type (prediction type)	
Diagnostic	19 (53.0)
Prognostic	17 (47.0)
Cancer type	
Lung	6 (17.0)
Liver	3 (8.3)
Kidney	3 (8.3)
Nasopharyngeal	3 (8.3)
Colorectal	3 (8.3)
Oral	2 (5.6)
Ovarian	2 (5.6)
Breast	2 (5.6)
Prostate	2 (5.6)
Esophageal	2 (5.6)
Gastric	2 (5.6)
Cervical	2 (5.6)
Any cancer	2 (5.6)
Brain	1 (2.8)
Abdominal lymph node	1 (2.8)
Thyroid	1 (2.8)
Outcome type	
Metastases	8 (22.2)
Death/survival	10 (27.8)
Cancer diagnosis	4 (11.1)
Cancer screening	2 (5.6)
Recurrence	2 (5.6)
Treatment response	5 (13.9)
Other[b]	5 (13.9)
Developed models[c]	
Random forest	26 (72.0)
Logistic regression[d]	22 (61.1)
Extreme Gradient Boosting	20 (55.6)
Support vector machine	20 (55.6)
Decision tree	10 (27.8)
Naïve Bayes	9 (25)
K-nearest neighbor	8 (22.2)
Gradient boost tree	4 (11.1)
Ensemble model	4 (11.1)
Light gradient boosting machine	3 (8.3)
Artificial neural network	3 (8.3)
AdaBoost	2 (5.6)
Bayesian network	2 (5.6)
Imaging features	
Image features used	14 (39.0)
Image features not used	22 (61.0)

SEER, surveillance, epidemiology, and end results, EHR, electronic health record.

[a] Data sources: cancer registry, Cancer Genome Atlas Lung Adenocarcinoma dataset, existing cohort from other publications, cross-sectional survey, Chinese clinical trial registry, from different institutions – by University of Pennsylvania, community oral cancer program and hospital clinic.

[b] Presence of simultaneous symptoms with advanced cancer patients, External/internal respiratory motion correlation in lung cancer patients, Presence of immune Checkpoint Inhibitors – Acute Kidney Injury, High risk of nodular thyroid disease and Select patients who are not suitable for Transarterial chemoembolization as the first treatment.

[c] Values do not add up to $n = 36$ (100%) as more than one model developed per study, and we describe the number of times each model was developed.

[d] We include the number of logistic regression models that were develop in additional to the machine learning models for descriptive purposes only. These were not included when evaluating sample size statements.

3.2. Observed sample size for model training

All but one study (n = 35/36; 97%) failed to report rationale for their sample size (given as a calculation or some justification) for their model development or validation analysis in their studies. One study reported that a learning curve (a visual display of model performance with increasing sample size) was used to assess whether their sample size was sufficient [47] – "*A learning curve is used to diagnose if the sample size is adequate for modeling and if an overfitting or underfitting problem occurs. It comprises two lines that represent the errors of the training set and the validation set, respectively, in relation to the sample size.*" A plot of the learning curve was provided in the results, which illustrated *"that with 25 features being considered, the gap between the train and the validation error became steady as the sample size exceeded 200"* – this study used 399 individuals to train their model, of which 242 experienced the outcome.

The median available sample size before any splitting was 426 (range: 67–71,414) and 147 events (range: 18–3749) (Table 2). After data modification through omission of missing data, imputation, or data-splitting (into train and test sets), the median sample size used to train the model was 310 (range: 54–57,134) and a median of 87 events (range: 25–2544).

3.3. Candidate predictors

The number of candidate predictors was clearly reported in only five studies (n = 5/36; 14%) but could be counted (by counting from reported tables in the publication) or estimated (by adding counts from multiple sources in the publication) in an additional 30 studies (n = 30/36; 83%). The median total number of candidate predictors in these 35 studies was 38 (range: 5, 3244). The median number of predictors in the final models reporting in 33studies (n = 33/36, 92%) was 10 (range: 4, 119) (Supplementary Table 3).

3.4. Calculated minimum required sample size for model training

The minimum required sample size was calculable for 17 studies (n = 17/36; 47%) that provided all the information (prevalence rate of the outcome of interest, number of candidate predictor parameters and reported c-statistic) needed for the sample size calculation. In the remaining 19 studies there was insufficient information about the prevalence of the outcome of interest or the number of events used to develop the model, thus comparison of sample size to recommendation was not possible in these 19 studies.

Of the 17 studies where sample size could be calculated, we used the reported bias-corrected c-statistic from 10 studies (59%) [38,42,44,50,55−60] and the reported c-statistic from a split-sample internal validation from seven studies (41%) [26,28,30,31,39,43,46]. Only five studies (n = 5/17; 29%) met N_{min} to estimate overall risk precisely (criterion 3) and minimize overfitting (criteria 1 and 2) [31,38,43,55,56], with a further three studies (n = 3/17; 18%) that only met the N_{min} to estimate overall risk precisely [39,44,60]. The remaining nine studies (n = 9/17; 53%) studies did not meet the N_{min} for any criteria (Fig 2).

Overall, there was a median deficit of 302 events [LQ to UQ: −13 to 1031] in the number used to develop (train) the models compared the minimum required sample size, which reduced to a deficit of 265 events [LQ to UQ: −60 to 1020] if the total available sample size had, hypothetically, been used (before any data modification such as split sample or excluding missing data) (Table 3).

3.5. Methodological conduct affecting sample size

3.5.1. Missing data

Missing data were mentioned or reported in 25 studies (n = 25/36; 69%). Approaches used to handle missing data varied; 17 (n = 17/25; 68%) studies conducted complete case analysis, where those with missing data were excluded during patient screening, data cleaning or during data preprocessing [29,31,34,38,43−47, 51,53−56,58−60]. Of these studies, six studies (n = 6/17, 35%) reported the number of patients excluded, with a median 1464 individuals (range: 24−217,885) excluded from the study due to missing data, equating to a total of 69% of data being excluded (proportion of total excluded individuals in six studies over total patients before any exclusions) [38,43,45,49,56,59]. Imputation (single, multiple, random forest, and multiple imputations by chain equation) were used in five studies (20%) [26,28,33,40,57]. One study reported there was no missing data [61]. Two (8%) studies used the equivalent to a missing indicator approach which is inherent to the eXtreme Gradient Boosting (XGBoost) modeling method. For example, one study mentioned "XGBoost uses parallel tree boosting and handles missing data well" and "an algorithm like XGBoost that can tolerate lots of missing data" [27]; while another study reported "[XGBoost] can make full use of missing data without filling in the data" [49].

3.5.2. Model testing and hyperparameter tunning

All studies tested (internally validated) their models. Fifty-eight percent of studies used a random split sample approach to internally validate their models (n = 21/36) [26−28,30−33,35−37,39,41,43,45−47,49,51,52,56,61], 22% used cross-validation (n = 8/36) [34,40,42,48,55,57, 58,60] and 11% used a random split sample approach with cross-validation (n = 4/36) [44,50,53,54]. Two studies used bootstrapping to internally validate their models [29,38] and one study used a non-random split sample approach (holding out one center for internal validation) [59]. One study that used a random split sample approach, also reported averaging the performance index and calculating 95% confidence intervals by 'repeating 100 tests' [36].

Of the 25 studies that included a random split sample, a median of 75% of the data was used to train the models (range: 60−90). Of the 12 studies that included a cross-validation approach for internal validation, a median of 5-fold cross-validation was used to train the models (range: 4 to 10).

Hyperparameter tunning was reported in 75% of studies (n = 27/36). Of these, 10 studies reported using cross-validation (n = 10/27, 37%) [30,37−40,45−48,59] and three studies used random grid search (n = 3/27, 11%) [33,34,43]. Eight studies reported using cross-validation with grid search to tune hyperparameters (n = 8/27, 30%) [32,35,41,44,56,58,60,61] and one study reported using cross-validation with 'random hyperparameter tuning' [28]. One study reported using 'default hyperparameters' [57], two studies reported prespecified hyperparameters [36,42], one study used random split sampling to allocate a portion of the data to tune hyperparameters and one study specified using the 'training set' [29].

3.5.3. Class imbalance

Class imbalance was reported in 14 (n = 14/36; 39%; Table 4) studies with Synthetic Minority Oversampling Technique (SMOTE) or SMOTE with Edited Nearest Neighbor (SMOTE-ENN) most often used to balance data (n = 12/36; 28%) [28,30,33−35,39,42,

Table 2. Sample sizes presented and used for development and internal validation of prediction models

Sample sizes	Reported (%) ($N = 36$)	Median [LQ, UQ]	Range
Available for model training			
Number of participants	36 (100%)	426 [180, 1646]	[67, 71,414]
Number of events	29 (80.6%)	147 [58, 743]	[18, 3749]
Used for model training			
Number of participants	36 (100%)	310 [159, 5123]	[54, 57,134]
Number of events	17 (47.2%)	87 [48, 190]	[25, 2544]
Used to tune hyperparameters			
Number of participants	1 (2.8%)	561 [NA]	[NA]
Number of events	-	-	-
Used for model testing (internal validation)			
Number of participants	34 (94.4%)	176 [59.5, 527.0]	[21, 42,827]
Number of events	15 (41.7%)	42 [25.5, 87.0]	[8, 404]

[LQ, UQ], lower and upper quartiles (25th and 75th percentiles).

49,53,54,58,61]. One study [46] reported using under sampling and another one reported using oversampling class and under sampling majority class [48]. Of 14 studies that applied SMOTE, only five studies reported their new sample size [30,34,46,49,54], and only four also provided the new number of events [30,46,49,54].

The initial sample size, number of events and percentage of events had a median of 1708 individuals (range: 102−42,827),

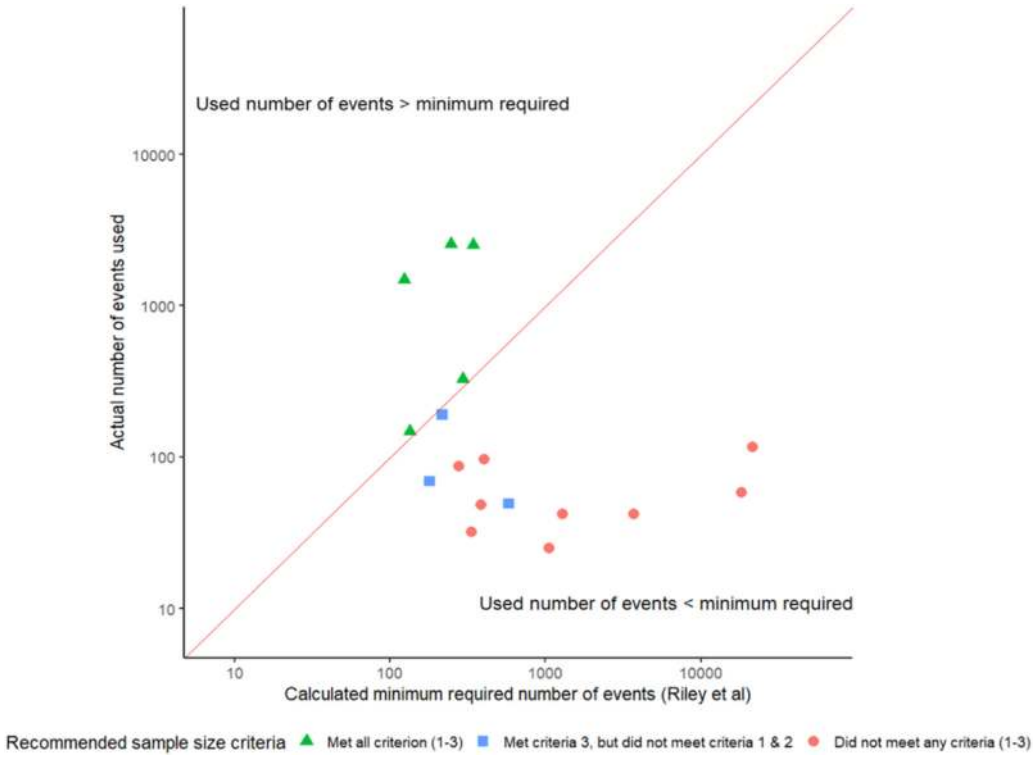

Figure 2. Scatter plot of calculated minimum required number of events against the number of events used for model development. Reference line (red) with slope 1 differentiates the studies that meet/exceed the minimum required sample size and those did not meet the minimum required sample size (Riley et al). Points below the red reference line indicate studies that use smaller sample sizes compared to calculated minimum required sample sizes; points above the red reference line indicate studies that use sample sizes higher than the minimum required. Green triangles represent studies that met recommended sample size criteria (criteria 1, 2 and 3), blue squares indicate studies that met precise overall risk estimation (criteria 1) but did not meet overfitting (criteria 2 and 3) and red circles indicate studies that did not meet any of the sample size calculation criteria.

Table 3. Minimum required sample size (N_{min}) and the difference between actual and calculated estimates where the observed value is the sample size used to develop a prediction model and the total available sample size; the values are presented median and (25% and 75% percentiles)

Sample size calculations	Median required sample size [LQ, UQ]	Median required events [LQ, UQ]	Median required events per predictor parameter (EPP) [LQ, UQ]
Calculated minimum required sample size			
Criterion 1 − small overfitting defined by an expected shrinkage of predictor effects by 10% or less (n = 17)	1981 [1170, 3818]	342 [241, 1056]	8 [5,12]
Criterion 2 − small absolute difference of 0.05 in the model's apparent and adjusted Nagelkerke's R^2 value (n = 17)	1462 [648, 4708]	253 [201, 1042]	7 [5,7]
Criterion 3 − precise estimation (within ± 0.05) of the overall risk (n = 17)	289 [234, 343]	72 [44, 115]	1 [0.2, 4]
Minimum required sample size (N_{min}) to meet all the three criteria for model training (n=17)	**1981 [1170, 4708]**	**342 [246, 1056]**	**8 [6, 12]**
Difference in observed and calculated minimum required sample size (used sample size)			
Difference between sample size used for model training and the N_{min} required to meet all criteria (n = 17)	−1010 [−3054, 45]	−302 [−1031,13]	−5 [−7, 4]
Difference between sample size used for model training and the N_{min} required to meet criterion 3 (n = 17)	−10 [−134, 224]	−3 [−37, 75]	0.1 [−0.1, 8.1]
Difference in observed and calculated minimum required sample size (total available sample size)			
Difference between available sample size for model training and the N_{min} required to meet all criteria (n = 17)	−419 [−3004, 214]	−265 [−1020, 60]	−5 [−7, 9.8]
Difference between available sample size for model training and the N_{min} required to meet criterion 3 (n = 17)	110 [−108, 1308]	8 [−23, 162]	0.2 [0.0, 13.4]

Bold indicates maximum of the minimum required sample size for the criterion 1−3.

323 events (range: 56−1631) and 29% (4% to 55%), respectively. After SMOTE under (and over) sampling methods were applied, the sample size, number of events and percentage of events increased to a median of 3390 individuals (range: 248−73,890), 665 events (range: 68−4932) and 42% (27%−50%), respectively (Supplementary Table 4).

Of the 17 studies that provided enough information to calculate the minimum required sample size using Riley et al formulae, six studies [28,30,39,42,46,58] also reported having used methods to handle class imbalance (Supplementary Table 5). Only one of these studies provided its new sample size for model development [30]; however, none provided their new number of events for model development.

3.5.4. Handling of overfitting

Nineteen studies reported their strategies to reduce risk of overfitting their models (n = 19/36, 53%) [27,28,30,32−34,37,41,42,44−48,53,54,56,57,59]. Seven studies referred to a form of regularization when training their models to reduce overfitting − these included five studies using the LASSO to screen and select predictors [37,41,45,51,59], one study using the early-stop method [56] and one study reporting that the 'objective function of XGBoost is regularized' [54]. Seven studies reported using a cross-validation method when training their models to reduce overfitting [32−34,42,44,46,48] − which in fact does not reduce overfitting as it does not change your model, rather it gives you a more realistic estimate of your model performance. These studies included one study that also used feature selection to reduce overfitting [33]. One study reported using a large amount of data to reduce overfitting [28], and another study used feature selection [57]. One study reported that a 'receiver operating characteristics (ROC) curve was built, and its area under the curve (AUC) was considered to discard all trained models most prone to overfitting' [35]. One study used a learning curve to diagnose overfitting [47] and one study used a reliability

Table 4. Methodological characteristics of the included 36 studies

Methodology	*n* (%) (*N* = 36)
Class imbalance reported	14 (38.9)
SMOTE[a]	12 (85.7)
Under/over sampling	2 (14.3)
Hyperparameter tuning reported	27 (75)
Grid search	3 (11.1)
Cross-validation	10 (37.0)
Cross-validation with grid search	8 (29.6)
Cross-validation with random hyperparameter tuning	1 (3.7)
Prespecified	2 (7.4)
Random split sample	1 (3.7)
Training set	1 (3.7)
Default hyperparameters	1 (3.7)
Internal validation reported	36 (100)
Random split sample	21 (58.3)
Non-random split sample	1 (2.8)
Cross-validation	8 (22.2)
Bootstrapping	2 (5.6)
Random split sample and cross-validation	4 (11.1)

[a] Includes two studies performing SMOTE-ENN (Synthetic Minority Oversampling Technique – Edited Nearest Neighbor).

curve (a plot that checks calibration – difference between the predicted and actual probabilities) to check and prevent overfitting [27].

In addition to the five studies using LASSO to select their features to reduce overfitting, six studies also used LASSO or regularized Ridge regression to select their predictors but did not explicitly specify using these methods to reduce overfitting [26,38,49,50,55,60].

3.6. Model performance

All studies reported the model discrimination (c-statistic). The apparent c-statistic was reported in eight studies (n = 8/36; 22%) [26,30,31,37,39,42,49,50] with a median apparent c-statistic of 0.89 (range: 0.76–0.98) (Supplementary Table 6). The twenty-five studies that randomly split their data into model training and testing reported a median c-statistic of 0.86 (range: 0.62–1.0). A c-statistic was reported in 15 studies out of 21 studies (n = 15/21, 71%) that used cross-validation and two studies used bootstrapping (n = 2/21, 10%) with a median value of 0.84 (range: 0.65–0.96) and 0.77 (range: 0.77–0.78), respectively.

Calibration was poorly reported, with only nine studies (25%) presenting a calibration plot or reporting calibration-in-the-large and calibration slope estimates [31,38,40,43,44,50,56,59,60]. Other model performance measures such as sensitivity, specificity, accuracy, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and F1-score were reported in some studies; however, almost all studies reported at least two of these performance measures. The number of studies that reported sensitivity, specificity, accuracy, PPV, NPV and F1-score were 26 (72.2%), 20 (55.6%), 24 (66.7%), 17 (47.2%), 9 (25%), 17 (47.2%), respectively and decision curve analysis was performed in 10 studies (n = 10/36, 28%) [36,38,43–45,50,56,59–61].

3.7. Reporting guidelines

Only seven studies mentioned using any reporting guideline (n = 7/36, 19.4%). Of these, five studies used the recommended reporting guideline for prediction model studies - TRIPOD [28,34,42,48,61], one study used a reporting guideline recommended for diagnostic accuracy studies - Standards for Reporting Diagnostic accuracy studies (STARD) [51], and one study used a risk of bias tool as a reporting guideline - Prediction Model Risk Of Bias ASsessment Tool [39]. However, of the studies that used a reporting guideline, none reported any rationale for their sample size (either calculation or justification), despite item five of the TRIPOD checklist asking authors to 'Explain how the study size was arrived at', and STARD ''Intended sample size and how it was determined'.

The 36 included studies were published in 27 journals. Of these 27 journals, reporting guidelines were mentioned in the author instructions of 10 journals (n = 10/27; 37%), and TRIPOD was specifically mentioned in three of these journals (n = 3/10; 30%). Of the seven studies that used a reporting guideline, only one published in a journal where reporting guidelines (including TRIPOD) was mentioned. Of the remaining 29 studies that did not use a reporting guideline, 11 were published in journals that mentioned reporting guidelines in their author instructions (Supplementary Table 7).

3.8. Sensitivity analysis

We repeated calculations to derive the minimum required sample sizes using the reported outcome prevalence and the number of candidate predictors for the included studies and took this information with 15% of the maximum Cox-Snell R^2 possible for each prediction scenario. Taking this more conservative approach led to only three (18%) studies meeting N_{min} for all criteria [43,55,56] and the deficit between the used and calculated sample sizes was 1758 [LQ to UQ: −10,370 to −378]. Full results are provided in Supplementary Figure 1 and Supplementary Table 8.

4. Discussion

4.1. Summary of findings

We evaluated 36 studies that included the development of a clinical prediction model for cancer risk prediction

using supervised ML approaches (excluding regression-based approaches). We found that almost all the studies did not report a sample size calculation or provide any justification of their sample size. Further, over half of the studies did not report the sample size that was eventually used to develop their models, after accounting for any omission of missing data and splitting of the data into testing and training sets. In turn, when applying the Riley et al. formulae to 17 studies, only five met the minimum recommended sample size. There was a large deficit in the sample size that was used to develop a model compared to what was minimally required. Given the Riley et al. formulae calculates the minimum required sample size for logistic regression, and that more flexible ML approaches typically require more data to achieve the same stability [12], the true deficit will likely be much greater, depending on the modeling approach used.

The deficit in the sample size used to develop a model was influenced by carrying out complete case analyses and arbitrarily splitting the data into training and testing datasets. However, we also found that, should the entire available dataset be used to develop the models, the shortfall in sample size reduces but remains substantial. When we focused our analysis on whether studies met the minimal sample size criterion of precisely estimating the overall average outcome risk only (the least stringent of the Riley et al. criteria), we found, though much smaller, there was still a deficit in the sample size that was used compared to the minimum required to meet this criterion. Collectively, our findings indicate that research teams are not adequately considering sample size when designing their study.

The sample used to develop the prediction models was also affected by methods to handle class imbalance, namely using under and over sampling methods such as SMOTE. Though these methods artificially increase the sample size available to develop models, the natural events fraction will be distorted. In the most extreme case, a study developing a prediction model to predict high-grade squamous intraepithelial lesions, reported an event rate of 25% which increased to 50% when over sampling was performed. No recalibration methods were used to recalibrate any of the models back to the natural event rates, leading to increased risk of developing miscalibrated models [62].

Many studies did not report using a reporting guideline despite reporting guidelines being specified in the author instructions of the journal. This emphasizes that more work is needed to ensure author adhere to reporting guidelines and also that journals enforce their author guidelines.

4.2. Context and implications

Generally, our findings contribute to a growing body of evidence of poor reporting and conduct in prediction model research as indicated in several studies [11,63−66]. There has been limited research that specifically calculates and evaluates the sample size requirements for prediction models using ML. Recently, Dhiman et al. reviewed the sample size requirements for regression-based prediction models in oncology and found that only 8% of studies provided sample size calculation or justification. Though a similar proportion of studies did not meet the minimum required sample size (73%) compared to our review (71%), the deficit in the sample size that was used to develop the prediction models was much higher in our study (median deficit of 75 events compared to median deficit of 302 events in this study). A study by Collins et al also reviewed the sample sizes in prediction model studies in prostate cancer and found a low rate of reporting of sample size justification (2%), though the proportion meeting the minimum required sample size was higher (51%) [67].

The implications of using an insufficient sample size to develop a model include imprecise measurement of any parameter estimates, increased risk of overfitting and instability in the model and its predictions (ie, the developed model and its predictions could be different if a different sample from the same population and of the same size was used) if used in clinical practice [13,16]. Some studies included in our review used penalization and shrinkage methods when selecting their predictors for the model building and it is tempting to assume that these methods would limit the risk of overfitting. However, studies have shown that when the sample size is too small, these methods do not overcome the problem and the shrinkage parameter would also be estimated with great uncertainty [68−70].

4.3. Strengths and limitations

We provide a comprehensive review of sample size considerations and reporting of the justification (if provided) and the reporting of numbers informing analyses along the study pathway (eg, before and after data modifications). We build on existing literature by focusing our review of studies specifically on those using ML methods to develop their prediction models and not limiting our study to any specific cancer type. We also include imaging studies where a prediction model was developed using extracted features from images in combination with clinical data to provide a broader and more relevant view of conduct around sample size in cancer prediction research.

Though we provide a contemporary view by limiting our search to studies published between December 1, 2022, and December 31, 2022, and limited it to search only one database, it is likely we will have missed some eligible studies. However, our findings are in line with existing research, and it is likely that additional studies would not change the conclusions of this review. Further, we only assess the sample size requirements for binary outcomes and not continuous or time to event outcomes. We decided to focus on binary outcomes as they are predominantly the outcome of interest for prediction, especially within ML models. We

do, however, recommend further research to assess sample size requirements for other types of outcomes.

We used recommended Riley et al. formulae to calculate the minimum required sample size for each study (where possible). It can be argued that as these formulae are focused on regression-based models, that they are not completely relevant for studies using ML methods. However, regression and ML should not be viewed as two separate entities, rather a spectrum of inherent modeling flexibility − a modeling flexibility that can also be added to regression models through using methods such as fractional polynomials and restricted cubic splines, and thus requiring a higher sample size as additional predictor parameters need to be estimated. We, therefore, believe these formulae are relevant to the ML field as ML methods typically require larger sample sizes than regression; thus, if studies do not achieve this minimum sample size for regression it is unlikely to be large enough to use ML methods. As such, these calculations also provide the minimum required sample for studies and so our review reflects a conservative estimate of the sample size requirements in this area. For example, if the sample size is not even large enough to estimate the overall risk precisely, developing a model to generate individual-level predictions is futile.

Further, we used the outcome proportion, number of predictor parameters, and c-statistics reported in the included studies to inform the minimum required sample size calculations in our review. As such these are post hoc calculations and may differ if they were conducted before the study was conducted. However, in the absence of existing information, our estimates provide the 'best case scenario' for each study.

As our calculations were reliant on the reported model performance measures and as poor reporting and conduct was found in the reviewed studies, it is likely that the reported model performance estimates are optimistic and higher than what may have been anticipated should a sample size calculation been done at the outset of the study. As such, this may have introduced some bias in our findings and a lower minimum sample size requirement may have been needed. To mitigate this, we assessed and report the sensitivity of our results by using recommended conservative approach when the anticipated R^2 and c-statistic are unavailable and took 15% of the max Cox-Snell R^2 approach. This reduced the number of studies meeting the minimum required sample size from five studies to three.

4.4. Recommendations

We strongly recommend that researchers estimate the minimum required sample size for their prediction model scenario before conducting the study, document this in a study protocol and report it in the final study article. In cases where existing data is being used, estimating the required sample size is still important to inform the analysis, to potentially reduce the number of candidate predictors where necessary, to minimize the risk of any overfitting. We appreciate that guidance for sample size in prediction model studies using ML is limited but we recommend using the Riley et al. formulae to at least estimate the minimum required sample size needed to develop a regression-based model, knowing that a larger sample is likely to be needed. In turn, we recommend further research to develop bespoke guidance for sample size estimations for ML prediction model studies.

Further, we found that many studies did not report using a reporting guideline despite reporting guidelines being specified in the author instructions of the journal. This emphasizes that more work is needed by journals to not only include the use of reporting guidelines in their author instructions, but to also check adherence to these instructions and authors adherence to the reporting guideline itself.

5. Conclusion

Studies that develop a prediction model using ML methods rarely report any justification of their sample size and there is a large deficit in the used and required sample size to minimize overfitting and precisely estimate the average outcome risk. By using a sample that is too small, study authors risk overfitting and large instability of their model and its predictions. We recommend researchers consider sample size requirements before conducting the study.

CRediT authorship contribution statement

Biruk Tsegaye: Writing − review & editing, Writing − original draft, Formal analysis, Data curation, Conceptualization. **Kym I.E. Snell:** Writing − review & editing, Data curation. **Lucinda Archer:** Writing − review & editing, Data curation. **Shona Kirtley:** Writing − review & editing, Conceptualization. **Richard D. Riley:** Writing − review & editing, Supervision. **Matthew Sperrin:** Writing − review & editing, Supervision. **Ben Van Calster:** Writing − review & editing, Supervision. **Gary S. Collins:** Writing − review & editing, Conceptualization. **Paula Dhiman:** Writing − review & editing, Writing − original draft, Supervision, Methodology, Data curation, Conceptualization.

Declaration of competing interest

There are no competing interests for any author.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2025.111675.

Data availability

Data that has informed the analysis can be found on the Open Science Framework (https://osf.io/kce76/).

References

[1] Overview | ColonFlag for identifying people at risk of colorectal cancer | Advice. NICE 2018. Available at: https://www.nice.org.uk/advice/mib142. Accessed October 10, 2023.

[2] Oke JL, Pickup LC, Declerck J, Callister ME, Baldwin D, Gustafson J, et al. Development and validation of clinical prediction models to risk stratify patients presenting with small pulmonary nodules: a research protocol. Diagn Progn Res 2018;2:22.

[3] Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. BMJ Open 2015;5:e007825.

[4] Battersby NJ, Bouliotis G, Emmertsen KJ, Juul T, Glynne-Jones R, Branagan G, et al. Development and external validation of a nomogram and online tool to predict bowel dysfunction following restorative rectal cancer resection: the POLARS score. Gut 2018;67:688−96.

[5] Candido dos Reis FJ, Wishart GC, Dicks EM, Greenberg D, Rashbass J, Schmidt MK, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res 2017;19:58.

[6] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349:255−60.

[7] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436−44.

[8] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol 2022;22:101.

[9] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. J Clin Epidemiol 2023;157:120−33.

[10] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. Diagn Progn Res 2022;6:13.

[11] Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, 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.

[12] 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.

[13] Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J 2023;65:e2200302.

[14] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I − continuous outcomes. Stat Med 2019;38:1262−75.

[15] Riley RD, Snell KI, Ensor J, Burke DL, Jr FEH, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38:1276−96.

[16] Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441.

[17] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.

[18] Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021;10:39.

[19] EndNote T. EndNote. Philadelphia: Clarivate; 2013.

[20] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016;5:210.

[21] Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11:e1001744.

[22] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med 2015;13:1.

[23] Ensor J, Martin EC, Riley RD. pmsampsize: Calculates the Minimum Sample Size Required for Developing a Multivariable Prediction Model. R package version 1.1.2. 2022. Available at: https://CRAN.R-project.org/package=pmsampsize. Accessed January 29, 2025.

[24] Riley RD, Van Calster B, Collins GS. A note on estimating the Cox-Snell R2 from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. Stat Med 2021;40:859−64.

[25] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available at: https://www.R-project.org/. Accessed January 29, 2025.

[26] Bi L, Guo Y. Development and validation of the random forest model via combining CT-PET image features and demographic data for distant metastases among lung cancer patients. J Healthc Eng 2022;2022:7793533.

[27] Chen A, Lu R, Han R, Huang R, Qin G, Wen J, et al. Building practical risk prediction models for nasopharyngeal carcinoma screening with patient graph analysis and machine learning. Cancer Epidemiol Biomarkers Prev 2023;32:274−80.

[28] Costantino A, Sampieri C, Pirola F, De Virgilio A, Kim S-H. Development of machine learning models for the prediction of positive surgical margins in transoral robotic surgery (TORS). Head Neck 2023;45:675−84.

[29] DeVries DA, Lagerwaard F, Zindler J, Yeung TPC, Rodrigues G, Hajdok G, et al. Performance sensitivity analysis of brain metastasis stereotactic radiosurgery outcome prediction using MRI radiomics. Sci Rep 2022;12:20975.

[30] Gaudiano C, Mottola M, Bianchi L, Corcioni B, Cattabriga A, Cocozza MA, et al. Beyond multiparametric MRI and towards radiomics to detect prostate cancer: a machine learning model to predict clinically significant lesions. Cancers 2022;14:6156.

[31] Hu J, Gong N, Li D, Deng Y, Chen J, Luo D, et al. Identifying hepatocellular carcinoma patients with survival benefits from surgery combined with chemotherapy: based on machine learning model. World J Surg Oncol 2022;20:377.

[32] Li H, Wang F, Huang W. A novel, simple, and low-cost approach for machine learning screening of kidney cancer: an eight-indicator Blood test panel with predictive value for early diagnosis. Curr Oncol 2022;29:9135−49.

[33] Sidey-Gibbons CJ, Sun C, Schneider A, Lu S-C, Lu K, Wright A, et al. Predicting 180-day mortality for women with ovarian cancer using machine learning and patient-reported outcome data. Sci Rep 2022;12:21269.

[34] Sorayaie Azar A, Babaei Rikan S, Naemi A, Bagherzadeh Mohasefi J, Pirnejad H, Bagherzadeh Mohasefi M, et al. Application of machine learning techniques for predicting survival in ovarian cancer. BMC Med Inf Decis Making 2022;22:345.

[35] Tan JY, Adeoye J, Thomson P, Sharma D, Ramamurthy P, Choi SW. Predicting overall survival using machine learning algorithms in oral cavity squamous cell carcinoma. Anticancer Res 2022;42:5859−66.

[36] Tang M, Gao L, He B, Yang Y. Machine learning based prognostic model of Chinese medicine affecting the recurrence and metastasis of I-III stage colorectal cancer: a retrospective study in China. Front Oncol 2022;12:1044344.

[37] Tankyevych O, Trousset F, Latappy C, Berraho M, Dutilh J, Tasu JP, et al. Development of radiomic-based model to predict clinical

outcomes in non-small cell lung cancer patients treated with immunotherapy. Cancers (Basel) 2022;14:5931.
[38] Tao Y, Chen S, Yu J, Shen Q, Ruan R, Wang S. Risk factors of lymph node metastasis or lymphovascular invasion for superficial esophageal squamous cell carcinoma: a practical and effective predictive nomogram based on a cancer hospital data. Front Med (Lausanne) 2022;9:1038097.
[39] Tsai HY, Tsai TY, Wu CH, Chung WS, Wang JC, Hsu JS, et al. Integration of clinical and CT-based radiomic features for pretreatment prediction of pathologic complete response to neoadjuvant systemic therapy in breast cancer. Cancers (Basel) 2022;14:6261.
[40] van der Stap L, van Haaften MF, van Marrewijk EF, de Heij AH, Jansen PL, Burgers JMN, et al. The feasibility of a Bayesian network model to assess the probability of simultaneous symptoms in patients with advanced cancer. Sci Rep 2022;12:22295.
[41] Wang M, Liu L, Dai Q, Jin M, Huang G. Developing a primary tumor and lymph node 18F-FDG PET/CT-clinical (TLPC) model to predict lymph node metastasis of resectable T2-4 NSCLC. J Cancer Res Clin Oncol 2023;149:247−61.
[42] Wang T, Hu J, Huang Q, Wang W, Zhang X, Zhang L, et al. Development of a normal tissue complication probability (NTCP) model using an artificial neural network for radiation-induced necrosis after carbon ion re-irradiation in locally recurrent nasopharyngeal carcinoma. Ann Transl Med 2022;10:1194.
[43] Xiong F, Cao X, Shi X, Long Z, Liu Y, Lei M. A machine learning-Based model to predict early death among bone metastatic breast cancer patients: a large cohort of 16,189 patients. Front Cell Dev Biol 2022;10:1059597.
[44] Xu J, Zhou J, Hu J, Ren Q, Wang X, Shu Y. Development and validation of a machine learning model for survival risk stratification after esophageal cancer surgery. Front Oncol 2022;12:1068198.
[45] Yu W, Lu Y, Shou H, Xu H, Shi L, Geng X, et al. A 5-year survival status prognosis of nonmetastatic cervical cancer patients through machine learning algorithms. Cancer Med 2023;12:6867−76.
[46] Yu X, Wu R, Ji Y, Huang M, Feng Z. Identifying patients at risk of Acute kidney Injury among patients receiving immune Checkpoint Inhibitors: a machine learning approach. Diagnostics 2022;12:3157.
[47] Zhang C, Zhang Y, Yang YH, Xu H, Zhang XP, Wu ZJ, et al. Machine learning models for predicting one-year survival in patients with metastatic gastric cancer who experienced upfront radical gastrectomy. Front Mol Biosci 2022;9:937242.
[48] Zhang X, Song X, Li G, Duan L, Wang G, Dai G, et al. Machine learning radiomics model for external and internal respiratory motion correlation prediction in lung tumor. Technol Cancer Res Treat 2022;21:15330338221143224.
[49] Zhao F, Zhang H, Cheng D, Wang W, Li Y, Wang Y, et al. Predicting the risk of nodular thyroid disease in coal miners based on different machine learning models. Front Med 2022;9:1037944.
[50] Zhu Z, Chen M, Hu G, Pan Z, Han W, Tan W, et al. A pre-treatment CT-based weighted radiomic approach combined with clinical characteristics to predict durable clinical benefits of immunotherapy in advanced lung cancer. Eur Radiol 2023;33:3918−30.
[51] Mahmoudi S, Koch V, Santos DPD, Ackermann J, Grünewald LD, Weitkamp I, et al. Imaging biomarkers to stratify lymph node metastases in abdominal CT − is radiomics superior to dual-energy material decomposition? Eur J Radiol Open 2023;10:100459.
[52] Park SB, Kim K-U, Park YW, Hwang JH, Lim CH. Application of 18 F-fluorodeoxyglucose PET/CT radiomic features and machine learning to predict early recurrence of non-small cell lung cancer after curative-intent therapy. Nucl Med Commun 2023;44:161−8.
[53] Shu X, Liu Y, Qiao X, Ai G, Liu L, Liao J, et al. Radiomic-based machine learning model for the accurate prediction of prostate cancer risk stratification. Br J Radiol 2023;96:20220238.
[54] Chen M, Wang J, Xue P, Li Q, Jiang Y, Qiao Y. Evaluating the feasibility of machine-learning-based predictive models for precancerous cervical lesions in patients referred for colposcopy. Diagnostics (Basel) 2022;12:3066.
[55] Feng X, Hong T, Liu W, Xu C, Li W, Yang B, et al. Development and validation of a machine learning model to predict the risk of lymph node metastasis in renal carcinoma. Front Endocrinol 2022;13:1054358.
[56] Ji L, Zhang W, Huang J, Tian J, Zhong X, Luo J, et al. Bone metastasis risk and prognosis assessment models for kidney cancer based on machine learning. Front Public Health 2022;10:1015952.
[57] Prayongrat A, Srimaneekarn N, Thonglert K, Khorprasert C, Amornwichet N, Alisanant P, et al. Machine learning-based normal tissue complication probability model for predicting albumin-bilirubin (ALBI) grade increase in hepatocellular carcinoma patients. Radiat Oncol 2022;17:202.
[58] Wang D, Lee SH, Geng H, Zhong H, Plastaras JP, Wojcieszynski AP, et al. Interpretable machine learning model for predicting pathologic complete response in patients with rectal adenocarcinoma treated with chemoradiation therapy. Int J Radiat Oncol Biol Phys 2022;114:e115−6.
[59] Wang D-D, Zhang J-F, Zhang L-H, Niu M, Jiang H-J, Jia F-C, et al. Clinical-radiomics predictors to identify the suitability of transarterial chemoembolization treatment in intermediate-stage hepatocellular carcinoma: a multicenter study. Hepatobiliary Pancreat Dis Int 2022;S1499-3872(22):00273-9.
[60] Yang T, Martinez-Useros J, Liu J, Alarcón I, Li C, Li W, et al. A retrospective analysis based on multiple machine learning models to predict lymph node metastasis in early gastric cancer. Front Oncol 2022;12:1023110.
[61] Adeoye J, Zheng L-W, Thomson P, Choi S-W, Su Y-X. Explainable ensemble learning model improves identification of candidates for oral cancer screening. Oral Oncol 2023;136:106278.
[62] Van Den Goorbergh R, Van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inf Assoc 2022;29:1525−34.
[63] Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, 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.
[64] Dhiman P, Ma J, Qi C, Bullock G, Sergeant JC, Riley RD, 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.
[65] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12−22.
[66] Wynants L, Calster BV, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328.
[67] Collins SD, Peek N, Riley RD, Martin GP. Sample sizes of prediction model studies in prostate cancer were rarely justified and often insufficient. J Clin Epidemiol 2021;133:53−60.
[68] Riley RD, Snell KIE, Martin GP, Whittle R, Archer L, Sperrin M, et al. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. J Clin Epidemiol 2021;132:88−96.
[69] Van Calster B, van Smeden M, deCock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. Stat Methods Med Res 2020;29:3166−78.
[70] Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J 2023;65:2200302.