



OPEN ACCESS

EDITED BY

Sharon R. Pine,
University of Colorado Anschutz Medical
Campus, United States

REVIEWED BY

Devesh U. Kapoor,
Gujarat Technological University, India
Miaomiao Yang,
Yantai Yuhuangding Hospital, China

*CORRESPONDENCE

Joseph Finkelstein
✉ joseph.finkelstein@utah.edu

RECEIVED 03 January 2025

ACCEPTED 18 March 2025

PUBLISHED 14 April 2025

CITATION

Smiley A, Villarreal-Zegarra D,
Reategui-Rivera CM, Escobar-Agreda S
and Finkelstein J (2025) Methodological
and reporting quality of machine
learning studies on cancer diagnosis,
treatment, and prognosis.
Front. Oncol. 15:1555247.
doi: 10.3389/fonc.2025.1555247

COPYRIGHT

© 2025 Smiley, Villarreal-Zegarra,
Reategui-Rivera, Escobar-Agreda and
Finkelstein. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Methodological and reporting quality of machine learning studies on cancer diagnosis, treatment, and prognosis

Aref Smiley¹, David Villarreal-Zegarra¹,
C. Mahony Reategui-Rivera¹, Stefan Escobar-Agreda²
and Joseph Finkelstein^{1*}

¹Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States,

²Telehealth Unit, Universidad Nacional Mayor de San Marcos, Lima, Peru

This study aimed to evaluate the quality and transparency of reporting in studies using machine learning (ML) in oncology, focusing on adherence to the Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models (CREMLS), TRIPOD-AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis), and PROBAST (Prediction Model Risk of Bias Assessment Tool). The literature search included primary studies published between February 1, 2024, and January 31, 2025, that developed or tested ML models for cancer diagnosis, treatment, or prognosis. To reflect the current state of the rapidly evolving landscape of ML applications in oncology, fifteen most recent articles in each category were selected for evaluation. Two independent reviewers screened studies and extracted data on study characteristics, reporting quality (CREMLS and TRIPOD+AI), risk of bias (PROBAST), and ML performance metrics. The most frequently studied cancer types were breast cancer (n=7/45; 15.6%), lung cancer (n=7/45; 15.6%), and liver cancer (n=5/45; 11.1%). The findings indicate several deficiencies in reporting quality, as assessed by CREMLS and TRIPOD+AI. These deficiencies primarily relate to sample size calculation, reporting on data quality, strategies for handling outliers, documentation of ML model predictors, access to training or validation data, and reporting on model performance heterogeneity. The methodological quality assessment using PROBAST revealed that 89% of the included studies exhibited a low overall risk of bias, and all studies have shown a low risk of bias in terms of applicability. Regarding the specific AI models identified as the best-performing, Random Forest (RF) and XGBoost were the most frequently reported, each used in 17.8% of the studies (n = 8). Additionally, our study outlines the specific areas where reporting is deficient, providing researchers with guidance to improve reporting quality in these sections and, consequently, reduce the risk of bias in their studies.

KEYWORDS

cancer, artificial intelligence, diagnosis, prognosis, therapy

1 Introduction

Cancer is one of the leading causes of disease burden and mortality worldwide. Early detection is crucial for improving clinical outcomes, yet approximately 50% of cancers are diagnosed at an advanced stage (1). Artificial Intelligence (AI), particularly machine learning (ML) models, has shown great potential in cancer diagnosis, prognostic, and treatment, standing out for its high accuracy, sensitivity, and ability to integrate into clinical workflows (2). This progress has led to a significant increase in studies and publications dedicated to the development and evaluation of these models in recent years (3).

The quality of reporting in these studies is critical, as clear and comprehensive documentation allows for the validation of results and their replication in different contexts (2). However, growing concerns have emerged regarding the completeness and accuracy of reports in ML-based research. Indeed, a meta-review of fifty systematic reviews, revealed that most reports of primary diagnostic accuracy studies were incomplete, limiting their utility and reproducibility (4). Furthermore, systematic reviews on treatment response in cancer patients have identified high heterogeneity and reporting issues (5, 6).

In response to these deficiencies, specific guidelines, such as the Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Models (CREMLS) and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD+AI), have been developed to standardize the key aspects that should be included in ML study reports (7, 8). Nevertheless, the widespread adoption of these tools depends on their dissemination and proper training, which may delay their integration into recent scientific publications (9). In addition to ensuring the quality of reporting, assessing the risk of bias in studies using AI models for cancer patients is essential. The Prediction Model Risk of Bias Assessment Tool (PROBAST) is a tool designed to evaluate the risk of bias across four domains: participants, predictors, outcomes, and analysis (10).

Our study evaluated PubMed publications that applied ML for the diagnosis, treatment, and prognosis of cancer patients. We assessed reporting quality (CREMLS and TRIPOD+AI), risk of bias (PROBAST), and ML performance metrics.

2 Materials and methods

2.1 Eligibility criteria

We included primary studies that applied AI techniques to predict cancer diagnosis, treatment, and prognosis in patients. Studies were eligible regardless of participants' age, sex, race,

ethnicity, or other sociodemographic or clinical characteristics. Articles written in any language and employing observational or experimental study designs were considered. Reviews, viewpoints, and conference papers were excluded. Our study included research that evaluates the first case of a specific cancer, specific recurrent cancers, groups of cancers, or metastases.

2.2 Search strategy and sources

Our search strategy included terms related to oncology, AI, prognosis, diagnosis, and treatment. A search was conducted in PubMed, restricted to studies published between February 1, 2024, and January 31, 2025. The results were sorted by “most recent,” and the 15 most recent studies on prognosis, diagnosis, and treatment were selected, yielding a total of 45 articles. The most recent articles were chosen in order to reflect the current state of the rapidly evolving landscape of AI applications in oncology. Details of the search strategy are provided in [Supplementary Material 1](#).

2.3 Selection process

The selection process was conducted using Rayyan (11). Two reviewers independently screened the title and abstract of each record identified to determine if the inclusion criteria were met. Subsequently, both reviewers independently evaluated the full text of records that passed the title and abstract screening. In cases of disagreement, a third reviewer made the final decision regarding inclusion.

2.4 Data collection process and data items

Data extraction was performed using a standardized Excel form. Two independent reviewers conducted the data collection process, and a third reviewer made the final decision in cases of disagreement. Each reviewer independently collected information using this form, which captured details such as the name of the first author, year of publication, article title, patient characteristics, type of AI tested or developed, clinical contexts, reported outcomes (AI performance metrics), reporting quality (CREMLS and TRIPOD+AI) (7, 8), and risk of bias (PROBAST) (10).

The extraction process was conducted in duplicate, and the level of agreement between independent evaluators was assessed in the first five included studies. The extracted information was highly consistent, achieving an agreement level above 80%.

2.5 Synthesis methods

2.5.1 Report quality

The TRIPOD+AI checklist consists of 27 key items designed to ensure the quality and transparency of reporting in prediction model studies (8). This updated version of TRIPOD extends the recommendations to include models developed using ML methods.

Abbreviations: AI, Artificial Intelligence; ML, Machine Learning; TRIPOD+AI, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis – Artificial Intelligence extension; CREMLS, Checklist for Reporting Machine Learning Studies; PROBAST, Prediction Model Risk of Bias Assessment Tool.

The primary objective of TRIPOD+AI is to promote clear, comprehensive, and reproducible reporting in studies that develop or validate prediction models in healthcare. TRIPOD+AI results were categorized using the following response options: Yes, No/Not reported, or Not applicable.

The CREMLS guideline was developed to enhance the transparency and quality of reporting in ML studies for diagnostic and prognostic applications in healthcare (7). It comprises 37 items organized into five categories: study details, data, methodology, model evaluation, and model explainability. CREMLS results were assessed using the following response options: Yes, No/Not reported, or Not applicable.

Since our objective was to assess the reporting quality of studies using ML for diagnostic, prognostic, and treatment outcomes in cancer patients, no studies were excluded based on a negative evaluation of reporting quality.

2.5.2 Risk of bias

The PROBAST tool assessed the risk of bias (RoB) in studies on prediction models for diagnosis, prognosis, and treatment. PROBAST provides a structured framework for evaluating the internal and external validity of studies by identifying potential biases in participant selection, predictor and outcome measurement, analytical strategies, and data presentation (10).

The response options in PROBAST were: low RoB/low concern regarding applicability (+), high RoB/high concern regarding applicability (-), and unclear RoB/unclear concern regarding applicability (?).

2.5.3 AI performance metrics

The study presented the main metrics used to evaluate the performance of each AI model included in the analyzed articles across different phases (i.e., training, testing, validation). These metrics included sensitivity (also known as recall or true positive rate), specificity (true negative rate), overall accuracy (probability of correct classification), positive predictive value (precision), F1 score, and the area under the ROC curve (AUC).

2.6 Statistical analysis

All statistical analyses were performed using R. Our study employed a descriptive approach based on percentages and absolute frequencies of CREMLS, TRIPOD+AI, and PROBAST scores. No inferential approaches focused on hypothesis testing were used.

3 Results

3.1 Study selection

Our study searched PubMed for ML models related to diagnosis, treatment, and prognosis that met the inclusion criteria. To standardize selection, we included only the first 15

articles for each category: prognosis, diagnosis, and treatment. The screening process, which involved title and abstract review followed by full-text assessment, is shown in [Figure 1](#).

3.2 Characteristics of the included studies

The first authors of the included studies were affiliated with institutions from 16 different countries. China had the highest number of publications ($n=27/45$; 60%), followed by Egypt ($n=3/45$; 6.7%). The most frequently studied cancer types were breast cancer ($n=7/45$; 15.6%), lung cancer ($n=7/45$; 15.6%), and liver cancer ($n=5/45$; 11.1%). More than half of the studies did not specify the source of funding ($n=25/45$; 55.6%). The individual characteristics of each study are summarized in [Supplementary Table 1](#). The list of included articles is presented in [Supplementary Table 2](#).

3.3 Report quality by CREMLS

Our study found that, in the study design section, all evaluated studies adequately reported the medical/clinical task of interest, research question, overall study design, and intended use of the ML model. However, only 51% of the studies provided information on whether existing model performance benchmarks for this task were considered.

In the data section, all studies reported information on methods of data collection and data characteristics. However, a large proportion did not report details on sample size calculation (98%), known quality issues with the data (69%), or bias introduced due to the data collection method used (62%).

In the methodology section, all studies reported the rationale for selecting the ML algorithm, and 98% described the method used to evaluate model performance during training. However, no studies reported strategies for handling outliers, and the majority did not provide information on strategies for model pre-training (92%) or data augmentation (79%).

In the evaluation section, nearly all studies reported some type of performance metrics used to evaluate the model (98%) and the results of internal validation (98%). However, no studies provided information on characteristics relevant for detecting data shift and drift, and 93% did not report or discuss the cost or consequences of errors.

In the explainability and transparency section, 82% of the studies reported the most important features and their relationship to the outcomes. Compliance with different criteria for each individual study is detailed in [Figure 2](#).

3.4 Report quality by TRIPOD+AI

The included studies demonstrate high compliance with key reporting standards for predictive modeling. All studies (100%) clearly identified their nature as predictive modeling investigations, specified their objectives, and provided detailed descriptions of data sources and participant eligibility ([Figure 3](#)). Also, they precisely

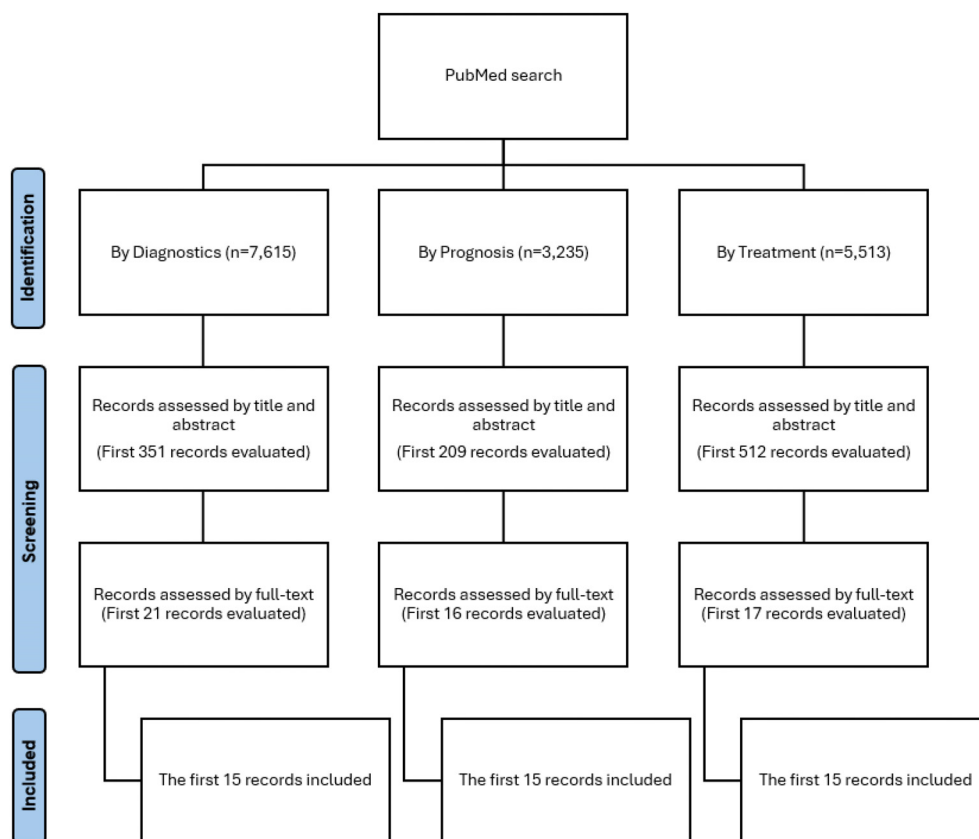


FIGURE 1
Flowchart.

defined predictors and outcomes, justifying their selection and evaluation methodology. Data preparation and model validation were also well-documented, with comprehensive descriptions of analytical methods, including model type, hyperparameter tuning, and performance metrics. Furthermore, all studies addressed the interpretation of results and discussed limitations, biases, and generalizability. Ethical considerations were reported in most cases, with 96% of studies indicating approval from an ethics committee or an equivalent review process. The flow of participants and differences between development and evaluation datasets were also consistently documented.

Despite strong adherence to several TRIPOD+AI criteria, significant reporting gaps were identified. 98% of studies did not report potential health inequalities among sociodemographic groups or describe the demographic characteristics and qualifications of outcome and predictor assessors. Additionally, 98% of studies failed to indicate where the study protocol could be accessed, and 96% did not provide information on study registration or describe methods to address model equity or justify their approach. The availability of analytical code was also limited, with 93% of studies not providing details on access. A critical issue was the complete absence of patient and public involvement in study development (100%). Furthermore, 93% of studies did not compare development and evaluation data for key predictors, while 96% did not report heterogeneity in model performance. Lastly, 98% of studies did not present results on model

updating, and 91% did not specify the level of user interaction or expertise required for model implementation.

3.5 Risk of bias by PROBAST

The methodological quality assessment using PROBAST revealed that 89% of the included studies exhibited a low overall risk of bias, suggesting adequate internal validity in most evaluated studies (see Table 1). Additionally, all studies demonstrated a low risk of bias in terms of applicability, indicating that the assessed predictive models are potentially transferable to real-world clinical settings.

At the domain level, the risk of bias in the participant and outcome domains was minimal, with 98% of the studies showing a low risk of bias in these areas. Regarding applicability, 100% of the studies reported low concern for the applicability of participants, while 98% indicated low concern for the applicability of outcomes.

However, the predictor domain showed an unclear risk of bias in 29% of the studies, both in terms of bias risk assessment and applicability. Furthermore, in the analysis domain, only 76% of the studies exhibited a low risk of bias, suggesting that a significant proportion may have employed methodological practices that compromise the validity of their findings. Issues such as handling of missing data, predictor selection, and correction for overfitting require further attention to enhance the robustness of predictive models in future studies.

TABLE 1 Risk of Bias by PROBAST.

First author (year)	ROB				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Dong (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Liang & Luo (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Zhang et al. (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Noman (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Zhou (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Liu (2025)	Low	Low	Low	Unclear	Low	Low	Low	Low	Low
Franco-Moreno (2025)	Low	Unclear	Low	Unclear	Low	Unclear	Low	Unclear	Low
Li (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Hamano (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Yin (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Xiao (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Liu (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Su (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Alshwayyat (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Liu (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Maleki (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Park (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
EL kati (2024)	Unclear	Low	Low	Unclear	Low	Low	Low	Unclear	Low
Wu (2025)	Low	Low	Low	High	Low	Low	Low	High	Low
Shan (2025)	Low	Low	Low	High	Low	Low	Low	High	Low
Ni (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Rafiepoor (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Shehta (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Chiu (2025)	Low	Low	Low	Unclear	Low	Low	Low	Low	Low
Wang (2025)	Low	Unclear	Low	High	Low	Unclear	Low	High	Low
Natha (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low

(Continued)

TABLE 1 Continued

First author (year)	ROB				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Gui (2024)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Şahin (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Chen (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Zhong (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Pan (2025)	Low	Low	Low	Unclear	Low	Low	Low	Low	Low
Wang (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Wang (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Nashat (2025)	Low	Low	Low	Unclear	Low	Low	Low	Low	Low
Zhou (2024)	Low	Low	Low	Unclear	Low	Low	Low	Low	Low
Torok (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Bozcuk (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Chen (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Pan & Wang (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Chufal (2025)	Low	Low	Low	Unclear	Low	Low	Low	Low	Low
Miyamoto (2025)	Low	Unclear	Low	Low	Low	Unclear	Low	Low	Low
Grosu (2025)	Low	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Huang (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Zhang (2025)	Low	Low	Low	Low	Low	Low	Low	Low	Low
Ramasamy (2024)	Low	Low	Low	Low	Low	Low	Low	Low	Low

ROB, risk of bias; Low, indicates low ROB/low concern regarding applicability; High, indicates high ROB/high concern regarding applicability; Unclear, indicates unclear ROB/unclear concern regarding applicability.

Section	Item	Yes	No
Study Details	1.1. The medical/clinical use of interest	100%	0%
	1.2. The study design	100%	0%
	1.3. Current medical/clinical practice	100%	0%
	1.4. The known predictors and confounders to avoid in target prediction / diagnosis	100%	0%
	1.5. The target clinical outcome	100%	0%
	1.6. The target population	100%	0%
	1.7. The target patient population	100%	0%
	1.8. The intended use of the ML model	100%	0%
	1.9. The model's performance	100%	0%
	1.10. The model's performance	100%	0%
	1.11. The model's performance	100%	0%
	1.12. The model's performance	100%	0%
	1.13. The model's performance	100%	0%
	1.14. The model's performance	100%	0%
Data	2.1. Inclusion / exclusion criteria for the patient cohort	100%	0%
	2.2. Methods of data collection	100%	0%
	2.3. Methods of data collection	100%	0%
	2.4. Data transparency	100%	0%
	2.5. Methods of data transformation and preprocessing applied	100%	0%
	2.6. Known quality issues with the data	100%	0%
	2.7. Data availability	100%	0%
	2.8. Data availability	100%	0%
	2.9. Data availability	100%	0%
	2.10. Data availability	100%	0%
	2.11. Data availability	100%	0%
	2.12. Data availability	100%	0%
	2.13. Data availability	100%	0%
	2.14. Data availability	100%	0%
Methodology	3.1. Strategies for handling missing data	100%	0%
	3.2. Strategies for handling missing data	100%	0%
	3.3. Strategies for handling missing data	100%	0%
	3.4. Strategies for handling missing data	100%	0%
	3.5. Strategies for handling missing data	100%	0%
	3.6. Strategies for handling missing data	100%	0%
	3.7. Strategies for handling missing data	100%	0%
	3.8. Strategies for handling missing data	100%	0%
	3.9. Strategies for handling missing data	100%	0%
	3.10. Strategies for handling missing data	100%	0%
	3.11. Strategies for handling missing data	100%	0%
	3.12. Strategies for handling missing data	100%	0%
	3.13. Strategies for handling missing data	100%	0%
	3.14. Strategies for handling missing data	100%	0%
Evaluation	4.1. The cost or consequence of errors	100%	0%
	4.2. The cost or consequence of errors	100%	0%
	4.3. The cost or consequence of errors	100%	0%
	4.4. The cost or consequence of errors	100%	0%
	4.5. The cost or consequence of errors	100%	0%
	4.6. The cost or consequence of errors	100%	0%
	4.7. The cost or consequence of errors	100%	0%
	4.8. The cost or consequence of errors	100%	0%
	4.9. The cost or consequence of errors	100%	0%
	4.10. The cost or consequence of errors	100%	0%
	4.11. The cost or consequence of errors	100%	0%
	4.12. The cost or consequence of errors	100%	0%
	4.13. The cost or consequence of errors	100%	0%
	4.14. The cost or consequence of errors	100%	0%
Explanatory and Transparency	5.1. The most important features and their relation to the outcomes	100%	0%
	5.2. The most important features and their relation to the outcomes	100%	0%
	5.3. The most important features and their relation to the outcomes	100%	0%
	5.4. The most important features and their relation to the outcomes	100%	0%
	5.5. The most important features and their relation to the outcomes	100%	0%
	5.6. The most important features and their relation to the outcomes	100%	0%
	5.7. The most important features and their relation to the outcomes	100%	0%
	5.8. The most important features and their relation to the outcomes	100%	0%
	5.9. The most important features and their relation to the outcomes	100%	0%
	5.10. The most important features and their relation to the outcomes	100%	0%
	5.11. The most important features and their relation to the outcomes	100%	0%
	5.12. The most important features and their relation to the outcomes	100%	0%
	5.13. The most important features and their relation to the outcomes	100%	0%
	5.14. The most important features and their relation to the outcomes	100%	0%
Study Details	1.1. The medical/clinical use of interest	100%	0%
	1.2. The study design	100%	0%
	1.3. Current medical/clinical practice	100%	0%
	1.4. The known predictors and confounders to avoid in target prediction / diagnosis	100%	0%
	1.5. The target clinical outcome	100%	0%
	1.6. The target population	100%	0%
	1.7. The target patient population	100%	0%
	1.8. The intended use of the ML model	100%	0%
	1.9. The model's performance	100%	0%
	1.10. The model's performance	100%	0%
	1.11. The model's performance	100%	0%
	1.12. The model's performance	100%	0%
	1.13. The model's performance	100%	0%
	1.14. The model's performance	100%	0%
Data	2.1. Inclusion / exclusion criteria for the patient cohort	100%	0%
	2.2. Methods of data collection	100%	0%
	2.3. Methods of data collection	100%	0%
	2.4. Data transparency	100%	0%
	2.5. Methods of data transformation and preprocessing applied	100%	0%
	2.6. Known quality issues with the data	100%	0%
	2.7. Data availability	100%	0%
	2.8. Data availability	100%	0%
	2.9. Data availability	100%	0%
	2.10. Data availability	100%	0%
	2.11. Data availability	100%	0%
	2.12. Data availability	100%	0%
	2.13. Data availability	100%	0%
	2.14. Data availability	100%	0%
Methodology	3.1. Strategies for handling missing data	100%	0%
	3.2. Strategies for handling missing data	100%	0%
	3.3. Strategies for handling missing data	100%	0%
	3.4. Strategies for handling missing data	100%	0%
	3.5. Strategies for handling missing data	100%	0%
	3.6. Strategies for handling missing data	100%	0%
	3.7. Strategies for handling missing data	100%	0%
	3.8. Strategies for handling missing data	100%	0%
	3.9. Strategies for handling missing data	100%	0%
	3.10. Strategies for handling missing data	100%	0%
	3.11. Strategies for handling missing data	100%	0%
	3.12. Strategies for handling missing data	100%	0%
	3.13. Strategies for handling missing data	100%	0%
	3.14. Strategies for handling missing data	100%	0%
Evaluation	4.1. The cost or consequence of errors	100%	0%
	4.2. The cost or consequence of errors	100%	0%
	4.3. The cost or consequence of errors	100%	0%
	4.4. The cost or consequence of errors	100%	0%
	4.5. The cost or consequence of errors	100%	0%
	4.6. The cost or consequence of errors	100%	0%
	4.7. The cost or consequence of errors	100%	0%
	4.8. The cost or consequence of errors	100%	0%
	4.9. The cost or consequence of errors	100%	0%
	4.10. The cost or consequence of errors	100%	0%
	4.11. The cost or consequence of errors	100%	0%
	4.12. The cost or consequence of errors	100%	0%
	4.13. The cost or consequence of errors	100%	0%
	4.14. The cost or consequence of errors	100%	0%
Explanatory and Transparency	5.1. The most important features and their relation to the outcomes	100%	0%
	5.2. The most important features and their relation to the outcomes	100%	0%
	5.3. The most important features and their relation to the outcomes	100%	0%
	5.4. The most important features and their relation to the outcomes	100%	0%
	5.5. The most important features and their relation to the outcomes	100%	0%
	5.6. The most important features and their relation to the outcomes	100%	0%
	5.7. The most important features and their relation to the outcomes	100%	0%
	5.8. The most important features and their relation to the outcomes	100%	0%
	5.9. The most important features and their relation to the outcomes	100%	0%
	5.10. The most important features and their relation to the outcomes	100%	0%
	5.11. The most important features and their relation to the outcomes	100%	0%
	5.12. The most important features and their relation to the outcomes	100%	0%
	5.13. The most important features and their relation to the outcomes	100%	0%
	5.14. The most important features and their relation to the outcomes	100%	0%

FIGURE 2
CREMLS checklist. +, Yes; -, No/No reported/The report is not clear; NA, Not applicable.

3.6 AI performance metrics

In terms of the AI model family used, 73.3% of the studies employed Supervised ML models (n = 33), while 24.4% used Deep Learning models (n = 11). Regarding the specific AI models identified as the best-performing, Random Forest (RF) and XGBoost were the most frequently reported, each used in 17.8% of the studies (n = 8). The performance metrics for each study are presented in Table 2.

4 Discussion

Our study included 45 studies that employed ML models for diagnostic, prognostic, and treatment outcomes in cancer patients. The findings indicate several deficiencies in reporting quality, as assessed by CREMLS and TRIPOD+AI. These deficiencies primarily relate to sample size calculation, reporting on data quality, strategies for handling outliers, documentation of ML model predictors, access to training or validation data, and reporting on model performance heterogeneity.

These reporting limitations align with the PROBAST risk of bias assessment, which identified unclear risk of bias in the predictors and analysis domains in approximately one in four included studies. Therefore, it is plausible to conclude that reporting quality is most limited in sections related to methodology, analysis, and elements that support study reproducibility (e.g., code and datasets). These limitations may introduce a higher risk of bias.

Our study aligns with previous research identifying significant issues in the quality of reporting in studies employing AI techniques to develop clinical prediction models, including the presence of “spin” practices and poor reporting standards (12). Moreover, another systematic review concluded that most studies are at high risk of bias, citing issues such as small sample sizes, inadequate handling of missing data, and lack of external validation (13). These problems are not limited to predictive models for cancer but also extend to other disciplines within clinical diagnostics, for instance, cardiology or mental health (14, 15).

The potential of AI models to predict cancer diagnoses, prognosis, and treatment are substantial, representing an opportunity to enhance healthcare access (16). While some studies have demonstrated that the diagnostic accuracy of AI models for detecting neoplasms is remarkably high, particularly in cases such as upper gastrointestinal tract cancers (17), the methodological quality of these studies remains low, with a high risk of selection bias. Additionally, other review studies evaluating the use of AI for predicting treatment and prognostic outcomes in cancer patients have also identified reporting issues and highly restricted access to data necessary for study replication (18, 19). Regarding performance metrics, the evidence highlights the need to standardize the reporting of metrics in studies due to variability in the parameters used and the inconsistent reporting of performance outcomes (20, 21).

Currently, several initiatives aim to promote the responsible use of AI models in oncology patients, such as the American Society of Clinical Oncology’s transparency principles for AI. These principles seek to

TABLE 2 AI performance metrics by study.

First author (year)	Type	Type of data	Type of AI tested or developed	AI models used	Best model	AI metrics
Dong (2025)	Prognosis	Cervical screening data, including hrHPV full genotyping, cytology results, and gynecological examination findings.	Supervised Machine Learning	XGBoost, Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB)	XGBoost	AUROC: 0.989 (maximum), 0.781 (minimum); Brier Score: 0.118 (maximum), 0.01 (minimum)
Liang & Luo (2025)	Prognosis	Structured data	Supervised Machine Learning	XGBoost, MLP, KNN, Random Forest (RF), Logistic Regression, AJCC Staging	XGBoost	AUC (Training: 0.95, Validation: 0.78)
Zhang et al. (2025)	Prognosis	Structured data (clinical, lab results)	Supervised Machine Learning	Logistic Regression, KNN, RF, SVM, XGBoost, LightGBM	Logistic Regression	Accuracy: 0.765 (LR), Precision: 0.750 (LR), Recall: 0.710 (LR), F1-Score: 0.695 (LR), AUC: 0.812 (LR)
Noman (2025)	Prognosis	Clinical and pathological features (tumor grade, ER status, HER2, lymph node involvement, tumor size, survival status).	Supervised Machine Learning	LightGBM, XGBoost, Random Forest, Support Vector Machine (SVM), Neural Networks (NN), K-Nearest Neighbors (KNN)	LightGBM (for recurrence), XGBoost (for recurrence type differentiation)	C-index: 0.837 (survival analysis), AUC: 92% (LightGBM for recurrence), Accuracy: 86% (XGBoost for recurrence type classification)
Zhou (2025)	Prognosis	Clinical, demographic, psychological, and health-related data.	Supervised Machine Learning	Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosted Trees (GBDT), Multi-Layer Perceptron (MLP)	Random Forest (RF)	AUC: 0.925, Accuracy: 84.7%, Sensitivity: 90.5%, Specificity: 80.9%, Brier Score: 0.107
Liu (2025)	Prognosis	Pathomic features extracted from whole slide imaging (WSI) histopathology.	Deep Learning, Weakly Supervised Learning	DenseNet121, ResNet50, Inception_v3, XGBoost, Random Forest, LightGBM, Cox Regression	XGBoost (for metastasis prediction)	AUC: 0.981 (training), 0.804 (testing); C-index: 0.892 (training), 0.710 (testing)
Franco-Moreno (2025)	Prognosis	Clinical and laboratory data (D-dimer, hemoglobin, cholesterol, platelets, leukocyte count, etc.)	Supervised Machine Learning	CatBoost, XGBoost, LightGBM	CatBoost	AUC-ROC: 0.86 (95% CI: 0.83–0.87), Sensitivity: 62%, Specificity: 94%, PPV: 75%, NPV: 93%
Li (2025)	Prognosis	Clinical, pathological, and molecular data (tumor stage, lymph nodes, neural invasion, Ki67, biomarkers)	Supervised Machine Learning	Random Survival Forest (RSF), Extreme Gradient Boosting (XGBoost), Decision Survival Tree (DST)	Random Survival Forest (RSF)	C-index: 0.968 (training), 0.936 (validation); AUC: 0.968 (training), 0.936 (validation); Brier Score: 0.067
Hamano (2025)	Prognosis	Clinical and laboratory data (heart rate, respiratory rate, leukocyte count, albumin, C-reactive protein, etc.)	Supervised Machine Learning	Fractional Polynomial (FP) Regression, Kernel Fisher Discriminant Analysis (KFDA), Kernel Support Vector Machine (KSVM), XGBoost	Kernel Support Vector Machine (KSVM)	AUC-ROC: 0.834
Yin (2025)	Prognosis	Computed tomography (CT) images and clinical data (tumor characteristics, biomarkers).	Deep Learning	ResNet50 + Multilayer Perceptron (MLP)	ResNet50 + MLP	AUC-ROC: 0.96 (training), 0.87 (internal testing), 0.85 (external validation)
Xiao (2025)	Prognosis	Clinical and pathological data (tumor markers, blood tests, histopathological features).	Supervised Machine Learning	Cox proportional hazards model, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost)	RFE-XGBoost	AUC-ROC: 0.89 (95% CI: 0.82–0.94), Accuracy: 0.83 (95% CI: 0.76–0.90), F1 score: 0.81 (95% CI: 0.72–0.88), Brier Score: 0.13 (95% CI: 0.09–0.17)

(Continued)

TABLE 2 Continued

First author (year)	Type	Type of data	Type of AI tested or developed	AI models used	Best model	AI metrics
Liu (2025)	Prognosis	Clinical and laboratory data (blood tests, tumor characteristics, surgical variables).	Supervised Machine Learning	Extreme Gradient Boosting (XGBoost), Random Forest (RF), Support Vector Machine (SVM), k-nearest neighbor (KNN)	XGBoost	AUC-ROC: 0.939 (training), 0.876 (validation), Accuracy: 0.8908
Su (2025)	Prognosis	CT-based radiomics features and genomic data (gene expression analysis from TCGA).	Supervised Machine Learning	Cox Regression, LASSO, Random Survival Forest (RSF)	Radiomics-based Cox Regression Model	AUC-ROC: 0.899 (1-year), 0.906 (3-year), 0.869 (5-year) in the test set; C-index: 0.819 (training), 0.892 (validation), 0.851 (test)
Alshwayyat (2025)	Prognosis	Clinical and pathological data (tumor characteristics, treatment modalities, ER/PR status).	Supervised Machine Learning	Random Forest (RF), Gradient Boosting Classifier (GBC), Logistic Regression (LR), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP)	Random Forest (RF)	AUC-ROC: 0.743, Accuracy: 70.78%, Sensitivity: 94.52%
Liu (2025)	Prognosis	Clinical and demographic data (tumor characteristics, metastasis, lymph node involvement, treatment history).	Supervised Machine Learning	XGBoost, SHapley Additive exPlanations (SHAP) tool for model interpretation.	XGBoost	AUC-ROC: 0.813 (1-year), 0.738 (3-year), 0.733 (5-year) in training; 0.781, 0.785, 0.775 in validation.
Maleki (2025)	Diagnosis	Clinical and laboratory data (paraneoplastic autoantibody panels from serum and CSF)	Supervised Machine Learning	Naive Bayes	Naive Bayes	AUC-ROC: 0.9795, Sensitivity: 85.71%, Specificity: 100%, Accuracy: 97.14%, Brier Score: 0.04
Park (2025)	Diagnosis	Electronic medical records (EMR) including medications, laboratory data, and clinical procedures.	Supervised Machine Learning	Random Forest, Lasso Logistic Regression, AdaBoost, Decision Tree, Naïve Bayes, Multilayer Perceptron	Random Forest	AUROC: 0.88 (training), 0.88 (validation); Sensitivity: 62%; Specificity: 94%
EL kati (2024)	Diagnosis	Clinical and imaging data from WBCD and WDBC.	Deep Learning	Deep Neural Network (DNN) with three hidden layers, optimized with Gradient Multi-Verse Optimizer (GMVO).	GMVO-optimized DNN	Accuracy: 93.5% (WBCD), 96.73% (WDBC); Precision: 88.06% (WBCD), 93.38% (WDBC); Specificity: 93.06% (WBCD), 95.83% (WDBC); Sensitivity: 95.64% (WBCD), 98.25% (WDBC).
Wu (2025)	Diagnosis	MRI-based radiomics and deep learning features.	Deep Learning	Deep Transfer Learning (DTL) models: vgg19, GoogLeNet, Inception_v3, Vision Transformer (ViT).	vgg19-combined fusion model	AUC-ROC: 0.990 (internal test), 0.988 (external test); Accuracy: 0.935 (internal test), 0.875 (external test); F1-score: 0.937 (internal test), 0.885 (external test).
Shan (2025)	Diagnosis	Multiphase contrast CT images (arterial, venous, and delayed phases).	Deep Learning	3D-ResUNet (coarse-to-fine segmentation network)	3D-ResUNet	Dice score: 0.8819; Precision: 0.905 (tumors >20 mm); Recall: 0.9728 (tumors >20 mm); F1-score: 0.9377 (tumors >20 mm)
Ni (2025)	Diagnosis	LDCT images, liquid biopsy (CACs), clinical variables (age, extra-thoracic cancer history, gender).	Supervised Machine Learning	Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Light Gradient Boosting (LGB), Least Absolute Shrinkage and Selection Operator (LASSO).	Random Forest (RF)	AUC-ROC: 0.99 (validation cohort); Sensitivity: 92%; Specificity: 97%.

(Continued)

TABLE 2 Continued

First author (year)	Type	Type of data	Type of AI tested or developed	AI models used	Best model	AI metrics
Rafiepoor (2025)	Diagnosis	MiRNA expression data from GEO datasets.	Supervised Machine Learning	Support Vector Machine (SVM), Boosted Trees (BT), Bootstrap Forest (BF), K-Nearest Neighbors (KNN), Neural Networks (NeB), LASSO Regression, Nominal Logistic, Naïve Bayes (NB).	Support Vector Machine (SVM)	AUC-ROC: 0.94, Sensitivity: 91%
Shehta (2025)	Diagnosis	Medical imaging (microscopic blood smear images)	Deep Learning	ResNetRS50, RegNetX016, AlexNet, Convnext, EfficientNet, Inception_V3, Xception, VGG19	ResNetRS50	Accuracy: 97%, Recall: 99%, F1-score: 98%
Chiu (2025)	Diagnosis	Dermoscopic images	Deep Learning	Swin Transformer, Vision Transformer (ViT), EfficientNetB5, EfficientNetV2B2, ResNet50, VGG16	Ensemble model (Swin Transformer + Vision Transformer + EfficientNetB5)	CSMUH dataset: Accuracy 97.31% (test), 99.77% (training). ISIC dataset: Accuracy 85.38% (test), 95.86% (training). False negatives reduced from 124 to 45 (ISIC dataset) and eliminated (CSMUH dataset).
Wang (2025)	Diagnosis	NBI endoscopic images and clinical data.	Supervised Machine Learning	Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT)	Random Forest (RF)	Accuracy: 96%, Precision: 90%, Recall: 100%, AUC: 0.97
Natha (2025)	Diagnosis	Dermoscopic images	Supervised Machine Learning	Random Forest (RF), Multi-layer Perceptron Neural Network (MLPN), Support Vector Machine (SVM)	Max Voting ensemble method	Accuracy: 94.70%, Precision: 93.23%, Recall: 92.54%, F1-score: 94.20%
Gui (2024)	Diagnosis	DCE-MRI images	Deep Learning	Faster R-CNN, Mask R-CNN, YOLOv9	Improved Faster R-CNN (with ROI aligning, FPN, and additional convolutional layers)	mAP: 0.752, Sensitivity: 0.950, False positive rate: 0.133
Şahin (2025)	Diagnosis	CT images	Deep Learning	You Only Look Once (YOLOv8), YOLOv5, YOLOv7	YOLOv8	Accuracy: 95.35%, Precision: 0.972, Recall: 0.919, Sensitivity: 94.74%, Specificity: 95.83%
Chen (2025)	Diagnosis	RNA-seq gene expression data from granulosa cells.	Supervised Machine Learning	LASSO Regression, Support Vector Machine Recursive Feature Elimination (SVM-RFE), XGBoost	XGBoost	AUC-ROC: 0.875 (validation cohort); AUC-ROC: 0.795 (SVM)
Zhong (2025)	Diagnosis	Label-free multiphoton imaging (MPM) data, histopathology features, clinical information.	Supervised Machine Learning	Decision Tree (DT), Multi-Layer Perceptron (MLP), Random Forest (RF), Support Vector Machine (SVM)	Multi-Layer Perceptron (MLP) - Two-stage MINT model	AUC-ROC: 0.92 (stage 1), 1.00 (stage 2); Accuracy: 89.36% (external validation)
Pan (2025)	Treatment	MRI-based radiomics and clinical data (tumor morphology, enhancement	Deep Learning	Logistic Regression, Support Vector Machine (SVM), Bayes, K-Nearest	Bayes Model integrating MRI	AUC-ROC: 0.829, Sensitivity: 76.9%, Specificity: 83.3%

(Continued)

TABLE 2 Continued

First author (year)	Type	Type of data	Type of AI tested or developed	AI models used	Best model	AI metrics
		characteristics, and diffusion-weighted imaging).		Neighbors (KNN), Decision Tree, Random Forest	radiomics and clinical data	
Wang (2025)	Treatment	Clinical, serological, and ultrasound imaging data.	Supervised Machine Learning	Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), XGBoost, LightGBM, Logistic Regression, K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Gradient Boosting Tree (GBT), Backpropagation Neural Network (BPNN), LASSO, Stepwise Regression.	SVM + RF	AUC-ROC: 0.972 (training), 0.922 (internal validation), 0.78 (external validation); Sensitivity: 85.52%, Specificity: 97.73%.
Wang (2025)	Treatment	Transcriptomic data (RNA expression) and qPCR analysis.	Supervised Machine Learning	Random-effects meta-analysis, Forward-search optimization	Meta-analysis optimized gene signature (LAMC2, TSPAN1, MYO1E, MYOF, SULF1)	AUC-ROC: 0.99 (training), 0.89 (external validation), 0.83 (peripheral blood validation)
Nashat (2025)	Treatment	Contrast-enhanced computed tomography (CECT) imaging and clinical features.	Supervised Machine Learning	Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor, Logistic Regression, Multilayer Perceptron, Random Forest	Support Vector Machine (SVM)	Accuracy: 95.24% (volumetric response), 84.13% (histologic response); Sensitivity: 95.65% (volumetric), 89.47% (histologic); Specificity: 94.12% (volumetric), 88% (histologic)
Zhou (2024)	Treatment	Blood biomarkers (C-reactive protein, neutrophil count, lactate dehydrogenase, alanine transaminase).	Supervised Machine Learning	Support Vector Machine (SVM), XGBoost, Random Forest (RF), Decision Tree (DT), Gradient-Boosted Machine (GBM), Generalized Linear Models (GLM), Least Absolute Shrinkage and Selection Operator (LASSO)	Support Vector Machine (SVM)	AUC-ROC: 0.908 (training), 0.666 (validation cohort 1), 0.776 (validation cohort 2)
Torok (2025)	Treatment	Serum N-glycome analysis using capillary electrophoresis with laser-induced fluorescence detection (CGE-LIF).	Supervised Machine Learning	Quadratic Discriminant Analysis (QDA), Random Forest, XGBoost, Neural Network, Support Vector Classifier (SVC)	Quadratic Discriminant Analysis (QDA)	AUC-ROC: 0.829 (regression), 0.8295 (progression), 0.841 (stationary)
Bozcuk (2025)	Treatment	Clinical and mutational data (age, ECOG score, mutation type, metastases, smoking status, blood biomarkers)	Reinforcement Learning	Deep Q-Network (DQN), Extra Trees Classifier (ETC)	Deep Q-Network (DQN)	AUC-ROC: 0.80 (DQN), 0.73 (ETC)
Chen (2025)	Treatment	Multi-omics data (scrNA-seq, bulk transcriptomics, GWAS)	Supervised Machine Learning	Random Survival Forest (RSF), Bayesian Deconvolution, Gene Set Variation Analysis (GSVA)	Multi-omics-driven Machine Learning Signature (MOMLS)	AUC-ROC: 0.875 (validation cohort); Kaplan-Meier survival analysis showed significant prognostic differences.

(Continued)

TABLE 2 Continued

First author (year)	Type	Type of data	Type of AI tested or developed	AI models used	Best model	AI metrics
Pan & Wang (2025)	Treatment	Clinical data (immune cell profiling, HBV DNA levels, antiviral treatment status).	Supervised Machine Learning	LASSO Regression, Random Forest (RSF), XGBoost	Random Forest (RSF)	AUC-ROC: 0.961 (RSF), 0.903 (XGBoost), 0.864 (LASSO)
Chufal (2025)	Treatment	Clinical data and DIBH waveform parameters (breath-hold amplitude, duration, consistency).	Supervised Machine Learning	Gradient Boosting, LightGBM, XGBoost, CatBoost, Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Machines, Naive Bayes	Uncalibrated Gradient Boosting Ensemble	AUC-ROC: 0.803, Recall: 0.526
Miyamoto (2025)	Treatment	CT radiomics features (texture analysis, intensity statistics, shape features).	Supervised Machine Learning	Random Forest (RF), Boruta Algorithm	Random Forest (RF)	AUC-ROC: 0.94 (training), 0.87 (validation)
Grosu (2025)	Treatment	CT colonography images, radiomics features.	Supervised Machine Learning	Random Forest (RF)	Random Forest (RF)	Accuracy: 84% (AI-assisted), Sensitivity: 85%, Specificity: 82%, Fleiss' kappa: 0.92
Huang (2025)	Treatment	Clinical and radiotherapy data (fraction size, cumulative dose, interruptions), pathological risk factors.	Supervised Machine Learning	eXtreme Gradient Boosting (XGBoost), Cox Regression	XGBoost	Accuracy: 57.89%, Sensitivity: 57.14%, Positive Predictive Value: 44.44%, AUC-ROC: 0.58
Zhang (2025)	Treatment	Clinical and demographic data (tumor stage, histological grade, treatment history).	Deep Learning	Balanced Individual Treatment Effect for Survival data (BITES), Cox Mixtures with Heterogeneous Effects (CMHE), DeepSurv, Cox Proportional Hazards (CPH), Random Survival Forest (RSF)	Balanced Individual Treatment Effect for Survival data (BITES)	IPTW-adjusted HR: 0.84 (CRT vs. surgery + RT/CRT); 0.77 (surgery + CRT vs. surgery + RT)
Ramasamy (2024)	Treatment	CT radiomics features, clinical data (performance status, treatment site, smoking history, SBRT dose)	Supervised Machine Learning	Multi-Layer Perceptron (MLP), Support Vector Classification (SVC), Random Forest, Adaptive Boosting (AdaBoost)	Multi-Layer Perceptron (MLP) with ADASYN sampling	AUC-ROC: 0.94 (combined model), 0.95 (clinical model for early-stage patients)

		Yes	No
Ramasamy (2024)		100%	0%
Zhang (2025)		100%	0%
Huang (2025)		100%	0%
Gren (2025)		100%	0%
Myamoto (2025)		2%	98%
Choi H (2025)		100%	0%
Pan & Wang (2025)		100%	0%
Chen (2025)		100%	0%
Bosnak (2025)		100%	0%
Terek (2025)		100%	0%
Zhou (2024)		100%	0%
Nasirak (2025)		100%	0%
Wang (2025)		100%	0%
Wang (2025)		100%	0%
Pan (2025)		100%	0%
Zhong (2025)		100%	0%
Chen (2025)		100%	0%
Sahin (2025)		100%	0%
Gui (2024)		100%	0%
Natha (2025)		100%	0%
Wang (2025)		100%	0%
Chu (2025)		100%	0%
Shenka (2025)		100%	0%
Rahipoor (2025)		100%	0%
Ni (2025)		100%	0%
Shan (2025)		100%	0%
Wu (2025)		100%	0%
El Kai (2024)		100%	0%
Park (2025)		100%	0%
Majaki (2025)		100%	0%
Lu (2025)		100%	0%
Ashwaryak (2025)		100%	0%
Su (2025)		100%	0%
Lu (2025)		100%	0%
Xiao (2025)		100%	0%
Vin (2025)		100%	0%
Herrano (2025)		100%	0%
Li (2025)		100%	0%
Franco-Moreno (2025)		100%	0%
Lu (2025)		100%	0%
Zhou (2025)		100%	0%
Noman (2025)		100%	0%
Zhang et al. (2025)		100%	0%
Liang & Luo (2025)		100%	0%
Dong (2025)		100%	0%
Topic	Item		
Title	1	100%	0%
Abstract	2	100%	0%
Background	3a	100%	0%
Background	3b	100%	0%
Background	3c	2%	98%
Objectives	4	100%	0%
Data	5a	100%	0%
Data	5b	96%	4%
Participants	6a	100%	0%
Participants	6b	100%	0%
Participants	6c	32%	68%
Data preparation	7	100%	0%
Outcome	8a	100%	0%
Outcome	8b	2%	98%
Outcome	8c	2%	98%
Predictors	9a	100%	0%
Predictors	9b	100%	0%
Predictors	9c	2%	98%
Sample size	10	96%	4%
Missing data	11	87%	13%
Analytical methods	12a	100%	0%
Analytical methods	12b	100%	0%
Analytical methods	12c	100%	0%
Analytical methods	12d	29%	71%
Analytical methods	12e	100%	0%
Analytical methods	12f	9%	91%
Analytical methods	12g	11%	89%
Class imbalance	13	67%	33%
Fairness	14	4%	96%
Model output	15	100%	0%
Training versus evaluation	16	100%	0%
Ethical approval	17	96%	4%
Funding	18a	91%	9%
Conflicts of interest	18b	96%	4%
Protocol	18c	2%	98%
Registration	18d	4%	96%
Data sharing	18e	18%	82%
Code sharing	18f	7%	93%
Patient & Public Involvement	19	0%	100%
Participants	20a	100%	0%
Participants	20b	98%	2%
Participants	20c	7%	93%
Model development	21	100%	0%
Model specification	22	53%	47%
Model performance	23a	100%	0%
Model performance	23b	4%	96%
Model updating	24	2%	98%
Interpretation	25	100%	0%
Limitations	26	100%	0%
Usability of the model	27a	11%	89%
Usability of the model	27b	9%	91%
Usability of the model	27c	31%	69%

FIGURE 3
TRIPOD+AI checklist. +, Yes; -, No/No reported/The report is not clear; NA, Not applicable. 1) Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted. 2) See TRIPOD+AI for Abstracts checklist. 3a) Explain the healthcare context, rationale for developing or evaluating the prediction model, including references to existing models. 3b) Describe the target population and intended purpose of the prediction model, including intended users. 3c) Describe any known health inequalities between sociodemographic groups. 4) Specify the study objectives, including whether it describes development or validation of a prediction model. 5a) Describe sources of data separately for development and evaluation datasets, rationale for using these data. 5b) Specify dates of collected participant data, including start and end of participant accrual. 6a) Specify key elements of the study setting including number and location of centers. 6b) Describe the eligibility criteria for study participants. 6c) Give details of any treatments received and how they were handled during model development or evaluation. 7) Describe any data pre-processing and quality checking across sociodemographic groups. 8a) Clearly define the outcome being predicted, time horizon, and assessment method. 8b) Describe qualifications and demographic characteristics of the outcome

FIGURE 3 (Continued)

assessors. 8c) Report any actions to blind assessment of the outcome. 9a) Describe the choice of initial predictors and any pre-selection before model building. 9b) Clearly define all predictors, including how and when measured. 9c) Describe qualifications and demographic characteristics of predictor assessors. 10) Explain how study size was determined, including sample size calculation. 11) Describe how missing data were handled and reasons for omitting data. 12a) Describe how data were used for development and evaluation of model performance. 12b) Describe how predictors were handled in the analyses. 12c) Specify model type, rationale, steps for building, hyperparameter tuning, and validation. 12d) Describe how heterogeneity in estimates and model performance was handled. 12e) Specify all measures and plots used to evaluate model performance. 12f) Describe any model updating arising from evaluation. 12g) Describe how model predictions were calculated. 13) Describe methods for handling class imbalance and recalibrating predictions. 14) Describe approaches used to address model fairness and their rationale. 15) Specify the output of the prediction model and rationale for classification thresholds. 16) Identify differences between development and evaluation data. 17) Name the ethics committee that approved the study and describe consent procedures. 18a) Give the source of funding and role of the funders. 18b) Declare any conflicts of interest and financial disclosures. 18c) Indicate where the study protocol can be accessed or state if unavailable. 18d) Provide registration information or state if not registered. 18e) Provide details of data availability. 18f) Provide details of analytical code availability. 19) Provide details of patient and public involvement or state no involvement. 20a) Describe the flow of participants through the study, including follow-up time. 20b) Report characteristics overall and by data source, including key demographics. 20c) Compare development data with evaluation data for key predictors. 21) Specify the number of participants and outcome events in each analysis. 22) Provide full details of the prediction model, including accessibility. 23a) Report model performance estimates with confidence intervals. 23b) Report results of heterogeneity in model performance. 24) Report results from model updating, including updated model. 25) Provide overall interpretation of results, including fairness. 26) Discuss study limitations, biases, and generalizability. 27a) Describe handling of poor-quality or unavailable input data. 27b) Specify required user interaction and expertise. 27c) Discuss future research steps and generalizability.

enhance transparency in AI models within oncology, potentially improving their real-world application and supporting clinical decision-making (22). However, a recent review published in a journal of this scientific society found that adherence to these principles remains limited, highlighting the need to promote their implementation (23). Nevertheless, this issue does not appear to be exclusive to oncology. A systematic review identified risks of bias and reporting quality issues in ML models across various healthcare contexts, not just oncology (13). Therefore, poor reporting quality seems to be a broader issue affecting health-related disciplines in general.

Incorporating cost-related information into AI model evaluations is essential for their real-world applicability. Understanding both direct and indirect costs would facilitate the integration of these models into clinical practice by providing a clearer assessment of the financial requirements for infrastructure, staff training, and implementation logistics (24). One of the key reporting limitations identified in the included studies was the lack of information on costs and the consequences of classification errors. This may be because most studies focus primarily on accuracy and effectiveness while overlooking the associated costs, despite evidence suggesting that AI can reduce the costs of cancer diagnosis and treatment (25). Additionally, assessing the economic impact of classification errors could help balance model performance with cost-effectiveness, ensuring that AI-driven tools provide both clinical and economic value. Future studies should integrate cost analyses alongside model performance metrics to enhance the feasibility and adoption of AI in oncology.

The analysis of AI model data must be transparent, as improper handling of missing data, outliers, participant imbalance, or dimensionality reduction can introduce biases into the model. For instance, internal biases and errors during model training may lead to misclassifications, potentially resulting in biased clinical decision-making if the model is implemented in real-world settings (26). Therefore, it is essential for AI models to ensure transparency throughout this process by sharing data, code, and any other materials that enable replication (22).

The main strength of our study lies in the detailed assessment of reporting quality using the CREMLS checklist and TRIPOD+AI checklist, and the risk of bias evaluation with PROBAST. However, an important limitation of our study is that it is not a systematic review encompassing all available evidence from multiple databases, resulting in limited representativeness.

5 Conclusion

Our study, using CREMLS and TRIPOD+AI, identified that AI models in oncology exhibit reporting limitations, particularly in the methodology sections related to predictors and the analysis plan. These findings align with an unclear risk of bias in one out of every four included studies, as indicated by PROBAST in the Predictors and Analysis domains. Additionally, our study outlines the specific areas where reporting is deficient, providing researchers with guidance to improve reporting quality in these sections and, consequently, reduce the risk of bias in their studies.

Author contributions

AS: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Supervision, Validation, Writing – review & editing. DV-Z: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. CMR-R: Data curation, Investigation, Writing – review & editing. SE-A: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – review & editing. JF: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study has been in part supported by the contract HT9425-24-1-0264-P00001 from the Congressionally Directed Medical Research Program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fonc.2025.1555247/full#supplementary-material>

SUPPLEMENTARY MATERIAL 1
Search strategy.

SUPPLEMENTARY TABLE 1
Characteristics of the included studies.

SUPPLEMENTARY TABLE 2
List of records included.

References

- Crosby D, Bhatia S, Brindle KM, Coussens LM, Dive C, Emberton M, et al. Early detection of cancer. *Science*. (2022) 375:eaay9040. doi: 10.1126/science.aay9040
- Maiter A, Salehi M, Swift AJ, Alabed S. How should studies using AI be reported? lessons from a systematic review in cardiac MRI. *Front Radiol*. (2023) 3:1112841. doi: 10.3389/fradi.2023.1112841
- Maulana FI, Adi PDP, Lestari D, Ariyanto SV, Purnomo A. (2023). The scientific progress and prospects of artificial intelligence for cancer detection: A bibliometric analysis, in: *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, . pp. 638–42. doi: 10.1109/ISITIA59021.2023.10221162
- Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. *NPJ Digit Med*. (2022) 5:11. doi: 10.1038/s41746-021-00544-y
- Gurumurthy G, Gurumurthy J, Gurumurthy S. Machine learning in paediatric haematological Malignancies: a systematic review of prognosis, toxicity and treatment response models. *Pediatr Res*. (2024). doi: 10.1038/s41390-024-03494-9
- Moharrami M, Azimian Zavareh P, Watson E, Singhal S, Johnson AEW, Hosni A, et al. Prognosing post-treatment outcomes of head and neck cancer using structured data and machine learning: A systematic review. *PLoS One*. (2024) 19:e0307531. doi: 10.1371/journal.pone.0307531
- El Emam K, Leung TI, Malin B, Klement W, Eysenbach G. Consolidated reporting guidelines for prognostic and diagnostic machine learning models (CREMLS). *J Med Internet Res*. (2024) 26:e52508. doi: 10.2196/52508
- Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. (2024) 385:e078378. doi: 10.1136/bmj-2023-078378
- Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med*. (2010) 8:24. doi: 10.1186/1741-7015-8-24
- Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. (2019) 170:51–8. doi: 10.7326/M18-1376
- Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Systematic Rev*. (2016) 5. doi: 10.1186/s13643-016-0384-4
- Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, 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. doi: 10.1016/j.jclinepi.2023.03.024
- Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*. (2021) 375:n2281. doi: 10.1136/bmj.n2281
- Cai Y, Cai Y-Q, Tang L-Y, Wang Y-H, Gong M, Jing T-C, et al. Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: a systematic review. *BMC Med*. (2024) 22:56. doi: 10.1186/s12916-024-03273-7
- Chen Z, Liu X, Yang Q, Wang Y-J, Miao K, Gong Z, et al. Evaluation of risk of bias in neuroimaging-based artificial intelligence models for psychiatric diagnosis: A systematic review. *JAMA Netw Open*. (2023) 6:e231671. doi: 10.1001/jamanetworkopen.2023.1671
- Kapoor DU, Saini PK, Sharma N, Singh A, Prajapati BG, Elossaily GM, et al. AI illuminates paths in oral cancer: transformative insights, diagnostic precision, and personalized strategies. *EXCLI J*. (2024) 23:1091–116. doi: 10.17179/excli2024-7253
- Arribas J, Antonelli G, Frazzoni L, Fuccio L, Ebigbo A, van der Sommen F, et al. Standalone performance of artificial intelligence for upper GI neoplasia: a meta-analysis. *Gut*. (2020) 70:1458–68. doi: 10.1136/gutjnl-2020-321922
- Corti C, Cobanaj M, Marian F, Dee EC, Lloyd MR, Marcu S, et al. Artificial intelligence for prediction of treatment outcomes in breast cancer: Systematic review of design, reporting standards, and bias. *Cancer Treat Rev*. (2022) 108:102410. doi: 10.1016/j.ctrv.2022.102410
- 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. doi: 10.1016/j.jclinepi.2021.06.024
- Kanan M, Alharbi H, Alotaibi N, Almasuood L, Aljoaid S, Alharbi T, et al. AI-driven models for diagnosing and predicting outcomes in lung cancer: A systematic review and meta-analysis. *Cancers (Basel)*. (2024) 16:674. doi: 10.3390/cancers16030674
- Kumar Y, Gupta S, Singla R, Hu Y-C. A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. *Arch Comput Methods Eng*. (2022) 29:2043–70. doi: 10.1007/s11831-021-09648-w
- American Society of Clinical Oncology. Principles for the responsible use of artificial intelligence in oncology (2024). Available online at: <https://society.asco.org/sites/new-www.asco.org/files/ASCO-AI-Principles-2024.pdf> (Accessed October 4, 2024).
- Smiley A, Reategui-Rivera CM, Villarreal-Zegarra D, Escobar-Agreda S, Finkelstein J. Exploring artificial intelligence biases in predictive models for cancer diagnosis. *Cancers (Basel)*. (2025) 17:407. doi: 10.3390/cancers17030407
- Wolff J, Pauling J, Keck A, Baumbach J. The economic impact of artificial intelligence in health care: systematic review. *J Med Internet Res*. (2020) 22:e16866. doi: 10.2196/16866
- Kacew AJ, Strohbehn GW, Saulsberry L, Laiteerapong N, Cipriani NA, Kather JN, et al. Artificial intelligence can cut costs while maintaining accuracy in colorectal cancer genotyping. *Front Oncol*. (2021) 11:630953. doi: 10.3389/fonc.2021.630953
- Corti C, Cobanaj M, Dee EC, Criscitiello C, Tolaney SM, Celi LA, et al. Artificial intelligence in cancer research and precision medicine: Applications, limitations and priorities to drive transformation in the delivery of equitable and unbiased care. *Cancer Treat Rev*. (2023) 112:102498. doi: 10.1016/j.ctrv.2022.102498