Journal of Clinical Epidemiology 165 (2024) 111199

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Open science practices need substantial improvement in prognostic model studies in oncology using machine learning

Gary S. Collins[a,1,*], Rebecca Whittle[a,1], Garrett S. Bullock[b,c], Patricia Logullo[a], Paula Dhiman[a], Jennifer A. de Beyer[a], Richard D. Riley[d], Michael M. Schlussel[a]

[a]*Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, University of Oxford, Oxford, United Kingdom*

[b]*Department of Orthopaedic Surgery, Wake Forest School of Medicine, Winston-Salem, NC, USA*

[c]*Centre for Sport, Exercise and Osteoarthritis Research Versus Arthritis, University of Oxford, Oxford, United Kingdom*

[d]*Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom*

Accepted 20 October 2023; Published online 28 October 2023

Abstract

Objective: To describe the frequency of open science practices in a contemporary sample of studies developing prognostic models using machine learning methods in the field of oncology.

Study design and setting: We conducted a systematic review, searching the MEDLINE database between December 1, 2022, and December 31, 2022, for studies developing a multivariable prognostic model using machine learning methods (as defined by the authors) in oncology. Two authors independently screened records and extracted open science practices.

Results: We identified 46 publications describing the development of a multivariable prognostic model. The adoption of open science principles was poor. Only one study reported availability of a study protocol, and only one study was registered. Funding statements and conflicts of interest statements were common. Thirty-five studies (76%) provided data sharing statements, with 21 (46%) indicating data were available on request to the authors and seven declaring data sharing was not applicable. Two studies (4%) shared data. Only 12 studies (26%) provided code sharing statements, including 2 (4%) that indicated the code was available on request to the authors. Only 11 studies (24%) provided sufficient information to allow their model to be used in practice. The use of reporting guidelines was rare: eight studies (18%) mentioning using a reporting guideline, with 4 (10%) using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis statement, 1 (2%) using Minimum Information About Clinical Artificial Intelligence Modeling and Consolidated Standards Of Reporting Trials-Artificial Intelligence, 1 (2%) using Strengthening The Reporting Of Observational Studies In Epidemiology, 1 (2%) using Standards for Reporting Diagnostic Accuracy Studies, and 1 (2%) using Transparent Reporting of Evaluations with Nonrandomized Designs.

Conclusion: The adoption of open science principles in oncology studies developing prognostic models using machine learning methods is poor. Guidance and an increased awareness of benefits and best practices of open science are needed for prediction research in oncology.

Keywords: Open science; Prognosis; Machine learning; Reporting; Data sharing; Code sharing

Funding statement: G.S.C., P.L., Jd.B., R.W., and M.M.S. are supported by Cancer Research UK (programme grant: C49297/A27294). GSC and RDR are supported by an MRC-NIHR Better Methods Better Research grant (reference: MR/V038168/1). RDR was supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. PD was supported by Cancer Research UK (project grant: PRCPJT-Nov21\100021). 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.

[1] Joint first authors.

* Corresponding author. Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, Centre for Statistics in Medicine, University of Oxford, Oxford OX3 7LD, United Kingdom. Tel.: +44-0-1865-223460; fax: +44-1865-227966.

E-mail address: gary.collins@csm.ox.ac.uk (G.S. Collins).
@GSCollins (G.S. Collins)

https://doi.org/10.1016/j.jclinepi.2023.10.015

What is new?

Key findings

- The adoption of open science principles in machine learning studies in oncology is poor.

What this adds to what was known?

- Study registration and making protocols available is rare for studies developing machine learning prognostic models in oncology.
- Despite having the aim to develop models for individualized estimates of prognosis, very few studies make their model available in a useable format. Reporting guidelines exist for prognostic model studies (such as the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD) Statement) but are infrequently used.

What is the implication and what should change now?

- Open science is linked to transparency, trustworthiness, and research integrity as it promotes the sharing of study materials, ensuring that scientific processes are accessible, reproducible, and accountable.
- Specific areas for improvement include study registration, making study protocols available, sharing data (and making data sharing statements meaningful), sharing of analysis code (including code to implement prognostic models for independent evaluation and implementation), and adherence to reporting guidelines such as the TRIPOD Statement.

1. Introduction

Open science principles advance scientific research by promoting collaboration, transparency, and accessibility [1]. Access to research outputs such as data sets, code, and methodologies, open science principles stimulate innovation, reduce redundancy, and address disparities in scientific information access. Open science is linked to transparency, trustworthiness, and research integrity as it promotes the sharing of study materials, ensuring that scientific processes are accessible, reproducible, and accountable [2].

Open science is more than making study materials such as data and code available. It is about transparency in all aspects of the research process, including (but not limited to) declarations of conflicts of interest, funding statements, study registration, access to protocols, disclosure of authorship contributions, and transparent reporting of study design, conduct, and findings [3]. Transparent reporting means readers should be explicitly told by authors why they carried out a piece of research, how they carried it out, what the findings were, and what those findings mean, along with any limitations of the research. Sharing data can be futile if the steps leading to its curation (e.g., study design) are not sufficiently described. Reporting guidelines provide authors with a set of recommendations for essential information to report and are a key element of open science [4]. They can also be used as a mechanism for promoting open science in published research by recommending that authors describe the availability of their data, code, and protocols.

In oncology, prognostic models are increasingly used to aid health care providers in estimating a patient's future risk of developing cancer, predicting outcomes and prognosis, and guiding treatment decisions [5−10]. There has been an exponential rise in the number of prediction models published each year, particularly with increasing interest in applying artificial intelligence powered by machine learning methods. Concerns have been raised about the lack of transparency and reproducibility in the field of machine learning [11,12] and oncology [13,14].

The aim of this study is to describe open science practices in studies developing a prognostic model using machine learning methods in oncology and to highlight areas where guidance is needed.

2. Methods

The study protocol is available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/6DX9Y), and was posted prior to data collection. Searching MEDLINE via the OVID platform, we identified prognostic model studies in oncology that were added to PubMed within the calendar month of December 2022. The search strategy was adapted from an existing search developed by a senior information specialist [11,15]. It included relevant MeSH subject headings and free-text terms searched in the title, abstract, or keyword fields, covering specific machine learning modeling terms (such as ''classification and regression tree'', ''decision tree'', ''random forest'', ''naïve bayes'', ''neural networks'', ''support vector machine'', ''gradient boosting machine'', and ''K nearest neighbor''), cancer-related terms (such as ''cancer'', ''neoplasm'', or ''tumor''), and prediction terms (such as ''predict'', ''prognosis'', or ''risk''). Modeling, cancer, and prediction search terms were combined to retrieve publications satisfying all three sets of search terms. A date restriction was added to limit the search to studies added to PubMed in December 2022 ensuring a contemporary sample of studies. No other search limits were applied. The full search strategy for the MEDLINE database is provided in Supplementary Table 1. The reporting of the search followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses-S recommendations [16].

2.1. Eligibility criteria

Studies were included if they described the development of a prognostic model in the field of oncology, developed using machine learning methods. Publications were eligible for this review based on the following inclusion and exclusion criteria.

2.1.1. Inclusion criteria

- Models developing
  - o using machine learning methods (as defined by the authors)
  - o using two or more predictors in combination
- English language studies

2.1.2. Exclusion criteria

- External validation studies
- Prognostic marker studies
- Diagnostic prediction models

2.2. Data extraction

Two researchers (G.S.C., and M.M.S.) independently screened the titles and abstracts of the identified publications using Zotero. Two researchers, from a combination of three reviewers (G.S.C., M.M.S., and G.S.B.), independently reviewed the full text of potentially eligible publications and performed a double data extraction of eligible publications. Data were extracted using a standardized data extraction form in Excel (available at https://osf.io/6dx9y/). After the extracted data were reconciled between pairs of reviewers (G.S.C., M.M.S., and G.S.B.), another researcher (R.W.) checked a random sample of 10 studies, with differences reconciled between G.S.C. and R.W. No additional data were sought from the authors of the included studies.

2.3. Data items

The following study characteristics were extracted: first author, title, journal, publisher, 2022 journal impact factor, date manuscript received, accepted, and published, and country of lead author. Basic details on the prognostic model were collected, as follows: type of cancer, predicted outcome, and modeling approach. Data relating to broad open science research and publication practices were collected. Publication practices included authorship contribution statement, author Open Researcher and Contributor ID (ORCID) identifiers, funding details, conflicts of interest statement, whether there was open peer review, whether the study was published open access, and if so the open access licence. Research practices included protocol details, study registration, data sharing statement, code availability, prognostic model availability, prognostic model format, supplementary material details, and whether a reporting guideline was mentioned. Where information was not reported in the paper (e.g., availability of any open peer review reports), we tried to locate this information on the article's page on the journal website. Where a link was reported to data or code we followed the link to confirm their existence.

2.4. Data analysis

Data were summarized using descriptive statistics (frequency, and percentages, median and interquartile range). A narrative synthesis was used to describe the open science practices. The Clopper-Pearson exact method was used to calculate 95% confidence intervals. All analyses were carried out in R (version 4.3.1), using the DescTools package.

3. Results

Our search identified 618 unique publications added to PubMed database between December 01, 2022, and December 31, 2022, and indexed in the MEDLINE database. We excluded 514 publications during abstract screening and 58 publications during full-text screening for not meeting the eligibility criteria. Reasons for exclusion were primarily not predicting a prognostic outcome or using a method not defined as machine learning by the authors (Fig. 1). We extracted data from 46 publications published in 37 journals [17−61] (Supplementary Table 2).

3.1. Study characteristics

Patient groups for whom most prognostic models had been developed were lung (n = 9) [17,18,31,39,41,49,50,61,62], gynecological (n = 6) [21,38,44,46,58,63], and breast (n = 5) [28,34,35,51,56] cancer patients (Table 1). Most studies had a first author from China (n = 20), followed by the United States (n = 7). The median time from submission to publication was 131 days (range 34−817 days; IQR 87, 185 days). Logistic regression (n = 21) [18,22,26−28,30,32,37,38,40−42,44,45,47−49,53,58,61,63] and random forests (n = 21) [20,22,23,27,28,30,32,35,37,38,41,42,44−46,48,51,52,55,57,58] were the most common model-building methods, followed by support vector machines (n = 20) [22,23,27,28,33−35,37,38,41,44,46,48,51−53,57,58,61,63], deep learning (n = 15) [18,19,21,24,27,29,31,32,35,43,54,59,60,61,63], and XGBoost (n = 14) [18,23,27,28,30,32,37,44,46,52,57,58,61]. Twenty-three studies (50%) developed models using two or more model-building approaches, with one study investigating 35 approaches [22].

3.2. Open science: publication practices

Most studies reported an authorship contribution statement (n = 38, 83%). Few studies reported ORCID identifiers for all authors (n = 3, 7%) [17,18,45]. Most

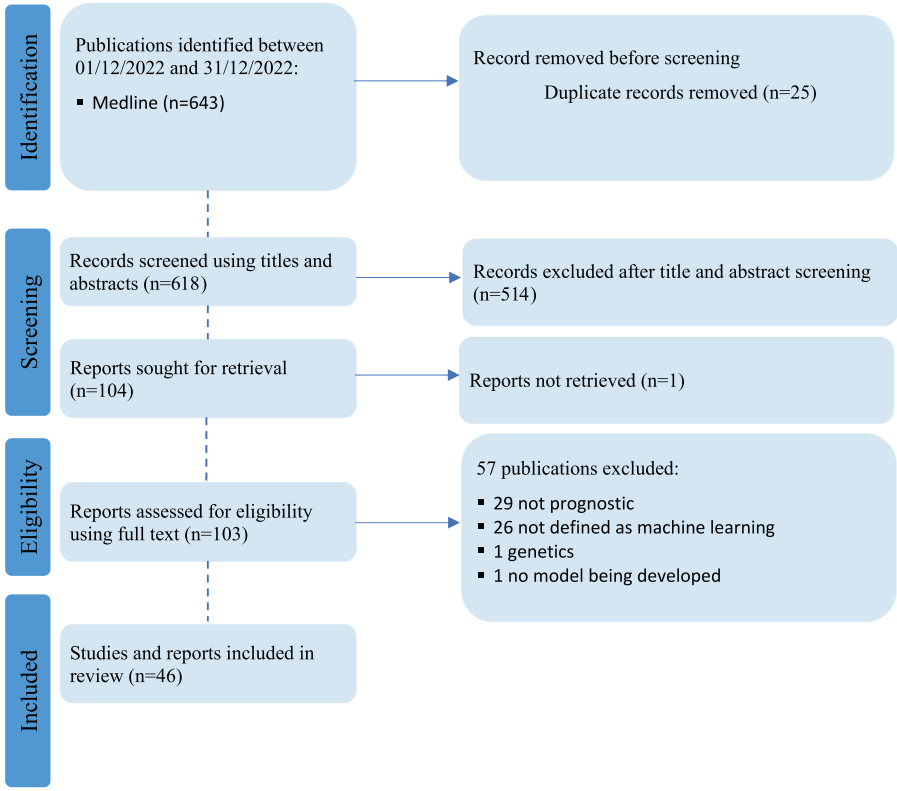

Fig. 1. PRISMA flow diagram of studies included in the systematic review.

reported no ORCID identifiers (n = 27, 59%), while two or more authors were reported in 16 studies (35%) see Table 2. Three studies (7%) had open peer review reports available on the article page on the journal website.

Thirty-three articles (72%) were published open access, primarily with a CC BY 4.0 open access license (n = 27). Funding statements were reported in most studies (n = 37, 80%), even for studies that received no funding (n = 4) [34,36,46,49]. A conflict of interest statement was reported for all but one of the included studies [43] (n = 45, 98%).

3.3. Open science: research practices

Data sharing statements were reported for 35 studies (76%). However, only two studies (4%) [actually] shared their data, in this case by providing a link to GitHub [20] or making it available in the supplementary material [31]. Twenty-one studies (46%) stated that data were available upon request [17,19,22,25,26,30,33,34,37–40,42,47–49,52–54,57,58], and six studies stated that the data were not available [18,23,28,44,55,63] (see Table 3). One study reported ‘‘not applicable’’ under the data sharing section [51]. Two stated that all data generated or analyzed for the study were included within or alongside the published article, yet no data, only results, were included in the article and no download link could be found on the journal website. Two studies linked to the the Surveillance, Epidemiology, and End Results[2] website and one to the Cancer Genome Atlas Program (Supplementary Table 3) [21,46,59].

Code sharing statements were only reported in twelve studies (26%) [18–20,26,27,31,39,40,42,48,52,61]. Of these, eight studies (17%) made their code available on GitHub [18–20,27,39,40,52,61], two stated that code was available upon request [26,42], and two made the code available in the supplementary material or through a provided link [31,48]. Eleven studies (24%) provided enough information for implementing the model in practice (i.e., enabling predictions for new individuals) [18,26,27,31,32,40,43,44,47,48,57]. Of those that did provide enough information, most presented their models as a web calculator (e.g., shiny app) (n = 6) [26,27,32,44,48,57] (although the web address was not accessible for one study [48]) or nomogram (n = 3) [26,43,47].

A protocol was made available for only 1 (2%) study [23]. It was also the only study to provide details of study registration.

[2] NIH National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) programme (seer.cancer.gov).

Table 1. Characteristics of included studies (n = 46)

Characteristic	*n* (%)
Cancer	
Lung	9 (20%)
Gynecological	6 (13%)
Breast	5 (11%)
Liver	4 (9%)
Colon/rectal	4 (9%)
Bladder	2 (4%)
Brain	2 (4%)
Esophageal/gastric	2 (4%)
Other	7 (15%)
Multiple cancers	5 (11%)
Country of first author	
China	20 (43%)
US	7 (15%)
Japan	3 (7%)
South Korea	3 (4%)
Canada	2 (4%)
Iran	2 (4%)
Egypt, France, Germany, India, Ireland, Taiwan, Thailand, UK, China/US (all n = 1)	9 (18%)
Common modeling approaches[a]	
Logistic regression	21 (46%)
Random forest	21 (46%)
Support vector machine	20 (43%)
Deep learning	15 (33%)
XGBoost	14 (30%)
Decision tree	8 (17%)
Naïve Bayes	6 (13%)
k-nearest neighbor	5 (11%)
Cox regression	3 (7%)
Time from submission to publication; median (IQR), range	131 days (IQR 87, 185); Range 34–817 days

[a] Only includes the most commonly reported modeling approaches. Some studies also fit other models.

Reporting guidelines were described as being used in eight (17%) studies [18,19,23,25,40,46,54,56]. The most used reporting guideline was the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD [64]) statement, used by four studies [18,40,46,54]. All other reporting guidelines were each used by one study: the Strengthening the Reporting of Observational Studies in Epidemiology ([65]) statement (n = 1) [25], the Standards for Reporting Diagnostic Accuracy Studies ([66]) checklist (n = 1) [56], and Transparent Reporting of Evaluations with Nonrandomized Designs (TREND [67]) statement (n = 1) [23]. One study [19] used both Minimum Information about CLinical Artificial Intelligence Modelling [68] and CONsolidated Standards Of Reporting Trials-Artificial Intelligence [69]. Of the studies that mentioned using a reporting guideline, only three provided a completed checklist in the supplementary material (two for TRIPOD [40,54]; one for TREND [23]).

Table 2. Summary of studies adhering to open science principles: publication practices (n = 46)

Open science practice	Frequency	% (95 CI)
Authorship contribution statement reported	38	83% (69–92%)
ORCID identifiers reported	19	41% (27–57%)
Yes, all authors	3	7% (1–18%)
Yes, some authors	16	35% (21–50%)
Funding statement reported	37	80% (66–91%)
Conflict of interest statement reported	45	98% (88–100%)
Availability of open-peer review reports	3	7% (1–18%)
Article published open access	33	72% (57–84%)
Open access licence		
CC BY 4.0	27	59% (43–73%)
CC BY-NC 4.0	3	7% (1–18%)
CC BY-NC-ND 4.0	3	7% (1–18%)

Table 3. Summary of studies adhering to open science principles: research practices (n = 46)

Open science practice	Frequency	% (95 CI)
Data sharing statement	35	76% (61–87%)
Available upon request	21	46% (31–61%)
Explicitly not shared	6	13% (5–26%)
Links to a website (e.g., SEER)	3	7% (1–18%)
Reported as available in the article but not	2	4% (0–15%)
Available (in supplementary material)	2	4% (0–15%)
'Not applicable'	1	2% (0–12%)
Code sharing statement	12	26% (14–41%)
GitHub	8	17% (8–31%)
Available upon request	2	4% (0–15%)
Other (e.g., supplementary material)	2	4% (0–15%)
Protocol availability	1	2% (0–12%)
Study registration	1	2% (0–12%)
Reporting guideline used	8	17% (8–31%)
MI-CLAIM and CONSORT-AI	1	2% (0–12%)
STARD	1	2% (0–12%)
STROBE	1	2% (0–12%)
TREND	1	2% (0–12%)
TRIPOD	4	9% (2–21%)

Abbreviations: STARD, standards for reporting diagnostic accuracy studies; STROBE, strengthening the reporting of observational studies in epidemiology; TREND, transparent reporting of evaluations with nonrandomized designs; TRIPOD: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; MI-CLAIM, Minimum Information about CLinical Artificial Intelligence Modelling; CONSORT-AI, CONsolidated Standards Of Reporting Trials-Artificial Intelligence.

4. Discussion

4.1. Summary of findings

In this review, we assessed the frequency of open science practices in studies added to PubMed in December 2022 that describe the development of prognostic models in the field of oncology. In terms of open sciences practices related to research, we found that only one study was reported as being registered; however, we were unable to verify the registration details [23]. Only one had a protocol available [23], two shared data [20,31], ten shared analysis code [18−20,27,31,39,40,48,52,61], and eight mentioned using a reporting guideline [18,19,23,25,40,46,54,56]. Only fifteen studies (33%) met at least 1 of these five open science practices. Only 1 study adhered to three practices [23], which was the highest adherence level seen.

Although most studies reported an authorship contribution statement, which is often a requirement for journal submission, many did not provide ORCID identifiers (www.orcid.org). The ORCID identifier was established over 10 years ago to allow authors to be reliably connected with their work, preventing mistaken identity and improving author recognition. While many major manuscript submission systems now integrate ORCID into the manuscript submission process, it is not usually a requirement for all authors to provide them.

Data sharing is critical in advancing research knowledge [70]. We found that data sharing statements were provided in over two-thirds of the articles reviewed ($n = 35$), yet only two studies made their data directly available. Numerous studies have shown that authors claiming their data are accessible upon (reasonable) request [14,71−76] (in our study $n = 21$) holds little worth. These statements usually play mere lip service to editorial requirements, as data sharing is infrequently practiced [74]. Rowhani-Farid and Barnett [71] found that of those that stated in the *British Medical Journal* that data were available on request only 16% provided the data when requested. Similarly, when Savage and Vickers [72] contacted authors who had published in *PLoS Medicine* or *PLoS Clinical Trials*, only 1 out of ten authors contacted provided an original data set. Data sharing statements are often a requirement of a journal's submission process, but in practice, they are meaningless if inadequately implemented. There are understandable challenges in sharing data and not all data are shareable, such as registry or electronic health records. However, much more can be done to both increase data sharing when possible and provide data sharing statements that are honest and have clear instructions about data access.

Sharing code not only facilitates replication and reproducibility, but also aids understanding of the analysis. Study descriptions in articles are often fraught with ambiguities or missing details. In contrast, the code used to analyze the data will list all the analysis steps, including any seeds, preprocessing, and parameter tuning. Only twelve studies (26%) included a code sharing statement in our study: eight released code on GitHub and two indicated that code was available on request from the authors. We only noted the reported claims of code availability, which were often opaque in description, and confirmed that the code existed. We did not inspect the contents of the code.

The 'product' of a prediction model study is the model that can be used to obtain predictions. These models should be made available for researchers to evaluate in their own data and potentially use in clinical practice [77]. Models can be made available in many formats, such as an equation, web calculator, nomogram, or software object [78]. However, not all formats facilitate an external validation in new data. External validation requires the actual model, for example as an equation or software object, so that it can be applied to make predictions in new individuals. We found that prognostic models in only 11 studies (24%) included in our review could in principle be used to obtain a prediction for an individual. Very few of the developed models were presented in a format or made available (e.g., equation, code) to allow external validation by independent researchers. For example, the regression coefficients for all predictors in a regression model should be reported, along with the intercept value for a logistic regression model or baseline survival at 1 or more time points of interest for a survival model. Twenty-four studies in our review implemented a regression model. Alternatively, code to implement the model should be made available, which is distinct from the analysis code to reproduce the manuscript's findings. However, there is no shareable equation for nonregression models such as a random forest or deep learner. Although studies reporting such models must make the code or a software object available to allow other researchers evaluate the model, this was rarely observed in the studies included in this review.

Sharing data and code to reproduce and understand the methods and study findings only has relevance if the details about the study design and conduct that gave rise to the data are completely and clearly reported, so that readers can understand their provenance and robustness. Not all readers have the technical expertize to read analysis code, and code only lists the steps in the analysis. Reporting guidelines underpin and enhance transparency, replication, and research integrity, fostering good research practice and underpinning all aspects of open science. Complete and transparent reporting of all aspects of the study is therefore critical. The TRIPOD statement published in 2015 provides recommendations for reporting the development and validation of clinical prediction models [64,79]. Only 10% of the publications included in the review reported using the TRIPOD statement, with a further 8% using an alternative reporting guideline. Reporting recommendations are in preparation for models developed using machine learning methods (TRIPOD + AI), which also includes recommendations on reporting open science practices [80,81].

4.1.1. Current literature

The inclusion of data sharing statements does appear to be rising. However, most of these statements are not proving to be valuable and do not lead to the data being available or shared. Hamilton et al. [14] investigated how data and code were made available in cancer research published in 2019. They found that although a fifth of articles declared that data were publicly available, only 16% were available when investigated, and less than 1% complied with the FAIR principles for sharing research [82]. They also found that only 4% of reported codes to be available. The low level of code sharing has been noted elsewhere, including studies of artificial intelligence [83,84] and systematic reviews [76,85] and across different study designs in oncology [14]. Our findings are also consistent with those observed by Walters et al. [13] who concluded open science was absent from a random sample of published oncology studies between 2014 and 2018.

Open-access publishing is on the rise: 67% of the articles reviewed here were open access, whereas Hua et al. [86] found that only 58% of oncology articles published in December 2014 were open access. Similarly, Piwowar et al. [87] found that the prevalence of open access in general increased between 2009 and 2015, with 44.7% of articles published open access in 2015. Although publishing with open access is beneficial for improving the reach of research and for transparency, it often has huge cost implications. These costs are not feasible for some researchers, particularly those conducting unfunded research and those from lower- and middle-income countries. However, the recent increased use of preprint websites could alleviate this problem, ensuring a nonpeer-reviewed version of the manuscript is publicly available [88].

4.2. Strengths and limitations

While we searched MEDLINE, a major database for studies developing clinical prognostic models, we may have missed some eligible publications. We also only examined articles added to PubMed in December 2022, which might not generalize outside this window. However, our aim was to describe a contemporary sample of publications reflecting current practice rather than finding all available existing evidence. Additional studies would unlikely change the conclusions of this review.

4.3. Future research

The importance of a study protocol cannot be stressed enough [89]. A protocol describes the design, conduct, and data analysis but can be used to prompt researchers upfront on how they will embrace open science. For example, data and code sharing, along with any restrictions, disseminating the search, and using reporting guidelines to ensure complete and transparent reporting of all aspects of the study can all be planned before the research has been conducted. Future studies could build on our findings by focusing on research protocols and checking the consistency of what is proposed in the protocols to what was actually reported after study completion.

5. Conclusions

Publications describing the development of prognostic models using machine learning methods in the field of oncology rarely follow open science practices. Guidance and an increased awareness of best practices are needed, outlining the benefits of open science for research transparency and engendering trust in research findings. Focus is needed on providing meaningful data and code sharing statements and adhering to reporting guidelines.

Data sharing statement

The data generated during this study are available at https://osf.io/6dx9y/.

Code sharing statement

The R code used in the analysis reported in this paper is available at https://github.com/gscollins1973.

CRediT authorship contribution statement

Conceptualization; Gary S. Collins. Data curation; Gary S. Collins, Michael Schlussel, Garrett Bullock, Rebecca Whittle. Formal analysis; Gary S. Collins, Rebecca Whittle. Funding acquisition; Gary S. Collins. Investigation; Gary S. Collins, Garrett S. Bullock, Rebecca Whittle, Michael M. Schlussel. Methodology; Gary S. Collins, Patricia Logullo, Paula Dhiman, Jennifer A. de Beyer, Garrett S. Bullock, Rebecca Whittle, Richard D. Riley, Michael M. Schlussel. Project administration; Gary S. Collins. Resources; Gary S. Collins. Software; Gary S. Collins. Supervision; Gary S. Collins. Validation; Not applicable. Visualization; Gary S. Collins. Roles/Writing - original draft; Gary S. Collins, Rebecca Whittle. Writing - review & editing; Gary S. Collins, Patricia Logullo, Paula Dhiman, Jennifer A. de Beyer, Garrett S. Bullock, Rebecca Whittle, Richard D. Riley, Michael M. Schlussel.

Data availability

All data and code have been shared on the Open Science Framework, which has been mentioned in the manuscript.

Declaration of competing interest

All authors declare no conflicts of interest.

Acknowledgments

We would like to thank the two anonymous reviewers whose comments and suggestions helped improve our manuscript and availability of our study materials.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2023.10.015.

References

[1] UNESCO. UNESCO recommendation on open science. UNESCO 2021. Available at https://unesdoc.unesco.org/ark:/48223/pf0000379949. Accessed July 3, 2023.
[2] Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Human Behav 2017;1(1):0021.
[3] Cobey KD, Haustein S, Brehaut J, Dirnagl U, Franzen DL, Hemkens LG, et al. Community consensus on core open science practices to monitor in biomedicine. PLoS Biol 2023;21(1):e3001949.
[4] Azad TD, Ehresman J, Ahmed AK, Staartjes VE, Lubelski D, Stienen MN, et al. Fostering reproducibility and generalizability in machine learning for clinical prediction modeling in spine surgery. Spine J 2021;21(10):1610−6.
[5] Fong Y, Evans J, Brook D, Kenkre J, Jarvis P, Gower-Thomas K. The Nottingham Prognostic Index: five- and ten-year data for all-cause survival within a screened population. Ann R Coll Surg Engl 2015;97(2):137−9.
[6] Kattan MW, Eastham JA, Stapleton AM, Wheeler TM, Scardino PT. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J Natl Cancer Inst 1998;90:766−71.
[7] Corbelli J, Borrero S, Bonnema R, McNamara M, Kraemer K, Rubio D, et al. Use of the gail model and breast cancer preventive therapy among three primary care specialties. J Womens Health 2014;23(9):746−52.
[8] Markaki M, Tsamardinos I, Langhammer A, Lagani V, Hveem K, Røe OD. A validated clinical risk prediction model for lung cancer in smokers of all ages and exposure types: a HUNT study. EBioMedicine 2018;31:36−46.
[9] Lebrett MB, Balata H, Evison M, Colligan D, Duerden R, Elton P, et al. Analysis of lung cancer risk model (PLCOM2012 and LLPv2) performance in a community-based lung cancer screening programme. Thorax 2020;75(8):661−8.
[10] 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(3):e007825.
[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] Navarro CLA. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol 2022;22:12.
[13] Walters C, Harter ZJ, Wayant C, Vo N, Warren M, Chronister J, et al. Do oncology researchers adhere to reproducible and transparent principles? A cross-sectional survey of published oncology literature. BMJ Open 2019;9(12):e033962.
[14] Hamilton DG, Page MJ, Finch S, Everitt S, Fidler F. How often do cancer researchers make their data and code available and what factors are associated with sharing? BMC Med 2022;20(1):438.
[15] 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(1):13.
[16] 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(1):39.
[17] Araki K, Matsumoto N, Togo K, Yonemoto N, Ohki E, Xu L, et al. Developing artificial intelligence models for extracting oncologic outcomes from Japanese electronic health records. Adv Ther 2022;40:934−50.
[18] Chandran U, Reps J, Yang R, Vachani A, Maldonado F, Kalsekar I. Machine learning and real-world data to predict lung cancer risk in routine care. Cancer Epidemiol Biomarkers Prev 2022;32:337−43.
[19] Choo H, Yoo SY, Moon S, Park M, Lee J, Sung KW, et al. Deep-learning-based personalized prediction of absolute neutrophil count recovery and comparison with clinicians for validation. J Biomed Inform 2022;137:104268.
[20] 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(1):20975.
[21] Dong T, Wang L, Li R, Liu Q, Xu Y, Wei Y, et al. Development of a novel deep learning-based prediction model for the prognosis of operable cervical cancer. Comput Math Methods Med 2022;2022:4364663.
[22] El-Falouji AI, Sabri DM, Lotfi NM, Medany DM, Mohamed SA, Alaa-Eldin M, et al. Rapid detection of recurrent non-muscle invasive bladder cancer in urine using ATR-FTIR technology. Molecules 2022;27(24):8890.
[23] Gangil T, Sharan K, Rao BD, Palanisamy K, Chakrabarti B, Kadavigere R. Utility of adding Radiomics to clinical features in predicting the outcomes of radiotherapy for head and neck cancer using machine learning. PLoS One 2022;17:e0277168.
[24] Hao D, Li Q, Feng QX, Qi L, Liu XS, Arefan D, et al. SurvivalCNN: a deep learning-based method for gastric cancer survival prediction using radiological imaging data and clinicopathological variables. Artif Intell Med 2022;134:102424.
[25] Huang Z, Shu Z, Zhu RH, Xin JY, Wu LL, Wang HZ, et al. Deep learning-based radiomics based on contrast-enhanced ultrasound predicts early recurrence and survival outcome in hepatocellular carcinoma. World J Gastrointest Oncol 2022;14(12):2380−92.
[26] Huo Z, Chong F, Yin L, Li N, Zhang M, Guo J, et al. Development and validation of an online dynamic nomogram system for predicting cancer cachexia among inpatients: a real-world cohort study in China. Support Care Cancer 2022;31(1):72.
[27] 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.
[28] Jiang C, Xiu Y, Zhang S, Yu X, Qiao K, Huang Y. A novel inflammatory and nutritional prognostic scoring system for nonpathological complete response breast cancer patients undergoing neoadjuvant chemotherapy. Dis Markers 2022;2022:8044550.
[29] Kapoor R, Sleeman W 4th, Palta J, Weiss E. 3D deep convolution neural network for radiation pneumonitis prediction following stereotactic body radiotherapy. J Appl Clin Med Phys 2022;24:e13875.
[30] Ko BS, Jeon S, Son D, Choi SH, Shin TG, Jo YH, et al. Machine learning model development and validation for predicting outcome in stage 4 solid cancer patients with septic shock visiting the emergency department: a multi-center, prospective cohort study. J Clin Med 2022;11(23):7230.
[31] Li S, Li W, Ma T, Fu S, Gao X, Qin N, et al. Assessing the efficacy of immunotherapy in lung squamous carcinoma using artificial intelligence neural network. Front Immunol 2022;13:1024707.
[32] Li W, Dong Y, Liu W, Tang Z, Sun C, Lowe S, et al. A deep belief network-based clinical decision system for patients with osteosarcoma. Front Immunol 2022;13:1003347.
[33] Liu Z, Wang Y, Shen F, Zhang Z, Gong J, Fu C, et al. Radiomics based on readout-segmented echo-planar imaging (RS-EPI)

diffusion-weighted imaging (DWI) for prognostic risk stratification of patients with rectal cancer: a two-centre, machine learning study using the framework of predictive, preventive, and personalized medicine. EPMA J 2022;13(4):633−47.
[34] McAnena P, Moloney BM, Browne R, O'Halloran N, Walsh L, Walsh S, et al. A radiomic model to classify response to neoadjuvant chemotherapy in breast cancer. BMC med imaging 2022;22(1):225.
[35] Mortazavi SAR, Tahmasebi S, Parsaei H, Taleie A, Faraz M, Rezaianzadeh A, et al. Machine learning models for predicting breast cancer risk in women exposed to blue light from digital screens. J Biomed Phys Eng 2022;12(6):637−44.
[36] Nakanishi R, Morooka K, Omori K, Toyota S, Tanaka Y, Hasuda H, et al. Artificial intelligence-based prediction of recurrence after curative resection for colorectal cancer from digital pathological images. Ann Surg Oncol 2022;30:3506−14.
[37] Nan Y, Zhao S, Zhang X, Xiao Z, Guo R. Machine learning-based predictive and risk analysis using real-world data with blood biomarkers for hepatitis B patients in the malignant progression of hepatocellular carcinoma. Front Immunol 2022;13:1031400.
[38] Ou Z, Mao W, Tan L, Yang Y, Liu S, Zhang Y, et al. Prediction of postoperative pathologic risk factors in cervical cancer patients treated with radical hysterectomy by machine learning. Curr Oncol 2022;29(12):9613−29.
[39] Pan X, Lin H, Han C, Feng Z, Wang Y, Lin J, et al. Computerized tumor-infiltrating lymphocytes density score predicts survival of patients with resectable lung adenocarcinoma. iScience 2022;25(12):105605.
[40] Parikh RB, Hasler JS, Zhang Y, Liu M, Chivers C, Ferrell W, et al. Development of machine learning algorithms incorporating electronic health record data, patient-reported outcomes, or both to predict mortality for outpatients with cancer. JCO Clin Cancer Inform 2022;6:e2200073.
[41] Park SB, Kim KU, 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(2):161−8.
[42] 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(1):202.
[43] Qin X, Zhu J, Tu Z, Ma Q, Tang J, Zhang C. Contrast-Enhanced ultrasound with deep learning with attention mechanisms for predicting microvascular invasion in single hepatocellular carcinoma. Acad Radiol 2022;30:S73−80.
[44] Shazly SA, Coronado PJ, Yilmaz E, Melekoglu R, Sahin H, Giannella L, et al. Endometrial Cancer Individualised Scoring System (ECISS): a machine learning-based prediction model of endometrial cancer prognosis. Int J Gynaecol Obstet 2022;161:760−8.
[45] Siddiqui NS, Klein A, Godara A, Buchsbaum RJ, Hughes MC. Predicting in-hospital mortality after acute myeloid leukemia therapy: through supervised machine learning algorithms. JCO Clin Cancer Inform 2022;6:e2200044.
[46] 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 Inform Decis Mak 2022;22(1):345.
[47] Sun S, Wang J, Yang B, Wang Y, Yao W, Yue P, et al. A nomogram for evaluation and analysis of difficulty in retroperitoneal laparoscopic adrenalectomy: a single-center study with prospective validation using LASSO-logistic regression. Front Endocrinol 2022;13:1004112.
[48] 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.
[49] 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 2022;14(23):5931.
[50] Terada Y, Isaka M, Kawata T, Mizuno K, Muramatsu K, Katsumata S, et al. The efficacy of a machine learning algorithm for assessing tumour components as a prognostic marker of surgically resected stage IA lung adenocarcinoma. Jpn J Clin Oncol 2022;53:161−7.
[51] 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 2022;14(24):6261.
[52] Wang D, Lee SH, Geng H, Zhong H, Plastaras J, Wojcieszynski A, et al. Interpretable machine learning for predicting pathologic complete response in patients treated with chemoradiation therapy for rectal adenocarcinoma. Front artif intell 2022;5:1059033.
[53] Wang H, Qin J, Chen X, Zhang T, Zhang L, Ding H, et al. Contrast-enhanced computed tomography radiomics in predicting primary site response to neoadjuvant chemotherapy in high-risk neuroblastoma. Abdom Radiol 2023;48:976−86.
[54] 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(22):1194.
[55] Wiltgen T, Fleischmann DF, Kaiser L, Holzgreve A, Corradini S, Landry G, et al. 18F-FET PET radiomics-based survival prediction in glioblastoma patients receiving radio(chemo)therapy. Radiat oncol 2022;17(1):198.
[56] Wu G, Cheligeer C, Brisson AM, Quan ML, Cheung WY, Brenner D, et al. A new method of identifying pathologic complete response after neoadjuvant chemotherapy for breast cancer patients using a population-based electronic medical record system. Ann Surg Oncol 2022;30:2095−103.
[57] 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.
[58] 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 2022;12:6867−76.
[59] Zheng Q, Yang R, Ni X, Yang S, Jiao P, Wu J, et al. Quantitative assessment of tumor-infiltrating lymphocytes using machine learning predicts survival in muscle-invasive bladder cancer. J Clin Med 2022;11(23).
[60] Zhong S, Ren JX, Yu ZP, Peng YD, Yu CW, Deng D, et al. Predicting glioblastoma molecular subtypes and prognosis with a multimodal model integrating convolutional neural network, radiomics, and semantics. J Neurosurg 2022;139:305−14.
[61] 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 2022;33:3918−30.
[62] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science. arXiv 2022. Available at http://arxiv.org/abs/2207.07048. Accessed December 29, 2022.
[63] Sidey-Gibbons CJ, Sun C, Schneider A, Lu SC, 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(1):21269.
[64] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55−63.
[65] Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806−8.
[66] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of

diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1−12.
[67] Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004;94:361−6.
[68] Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26: 1320−4.
[69] Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370: m3164.
[70] Naudet F, Siebert M, Pellen C, Gaba J, Axfors C, Cristea I, et al. Medical journal requirements for clinical trial data sharing: ripe for improvement. PLoS Med 2021;18(10):e1003844.
[71] Rowhani-Farid A, Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. BMJ Open 2016; 6(10):e011784.
[72] Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. PLoS One 2009;4:e7078.
[73] Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in *the BMJ* and *PLOS Medicine*. BMJ 2018; 390:k400.
[74] Gabelica M, Bojčić R, Puljak L. Many researchers were not compliant with their published data sharing statement: mixed-methods study. J Clin Epidemiol 2022;150:33−41.
[75] Tedersoo L, Küngas R, Oras E, Köster K, Eenmaa H, Leijen Ä, et al. Data sharing practices and data availability upon request differ across scientific disciplines. Sci Data 2021;8(1):192.
[76] Hamilton DG, Hong K, Fraser H, Rowhani-Farid A, Fidler F, Page MJ. Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. BMJ 2023;382:e075767.
[77] Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inf Assoc 2019;26:1651−4.
[78] Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. BMJ 2019; 365:l737.
[79] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1−73.
[80] Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11(7):e048008.
[81] Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393(10181):1577−9.
[82] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3(1):160018.
[83] Gundersen OE, Kjensmo S. State of the art: reproducibility in artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence, 32; 2018:1644−51.
[84] Venkatesh K, Santomartino SM, Sulam J, Yi PH. Code and data sharing practices in the radiology artificial intelligence literature: a meta-research study. Radiol Artif Intell 2022;4(5):e220081.
[85] Page MJ, Nguyen PY, Hamilton DG, Haddaway NR, Kanukula R, Moher D, et al. Data and code availability statements in systematic reviews of interventions were often missing or inaccurate: a content analysis. J Clin Epidemiol 2022;147:1−10.
[86] Hua F, Sun H, Walsh T, Glenny AM, Worthington H. Open access to journal articles in oncology: current situation and citation impact. Ann Oncol 2017;28:2612−7.
[87] Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, et al. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ 2018;6:e4375.
[88] Krumholz HM, Bloom T, Sever R, Rawlinson C, Inglis JR, Ross JS. Submissions and downloads of preprints in the first year of medRxiv. JAMA 2020;324:1903−5.
[89] Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KG, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med 2014;11: e1001671.