Assessment of bias in scoring of AI-based radiotherapy segmentation and planning studies using modified TRIPOD and PROBAST guidelines as an example

Hurkmans, Coen ; Bibault, Jean-Emmanuel ; Clementel, Enrico ; Dhont, Jennifer ; van Elmpt, Wouter ; Kantidakis, Georgios ; Andratschke, Nicolaus

DOI: https://doi.org/10.1016/j.radonc.2024.110196

Radiotherapy and Oncology 194 (2024) 110196

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Assessment of bias in scoring of AI-based radiotherapy segmentation and planning studies using modified TRIPOD and PROBAST guidelines as an example

Coen Hurkmans [a,b,*], Jean-Emmanuel Bibault [c], Enrico Clementel [d], Jennifer Dhont [e,f], Wouter van Elmpt [g], Georgios Kantidakis [d], Nicolaus Andratschke [h]

[a] Dept. of Radiation Oncology, Catharina Hospital Eindhoven, the Netherlands
[b] Dept. of Electrical Engineering, Technical University Eindhoven, the Netherlands
[c] Dept. of Radiation Oncology, Hôpital Européen Georges Pompidou, Université Paris Cité, Paris, France
[d] European Organisation for the Research and Treatment of Cancer (EORTC), Brussels, Belgium
[e] Université libre de Bruxelles (ULB), Hôpital Universitaire de Bruxelles (H.U.B), Institut Jules Bordet, Department of Medical Physics, Brussels, Belgium
[f] Université Libre De Bruxelles (ULB), Radiophysics and MRI Physics Laboratory, Brussels, Belgium
[g] Department of Radiation Oncology (MAASTRO), GROW – School for Oncology and Reproduction, Maastricht University Medical Center+, Maastricht, the Netherlands
[h] Dept. of Radiation Oncology, University Hospital of Zurich, The University of Zurich, Zurich, Switzerland

ARTICLE INFO

Keywords:
Bias
Transparency
Artificial intelligence
Checklists
Guidelines
Radiation therapy
Inter-observer variation
Distinctiveness
Oncology

ABSTRACT

Background and purpose: Studies investigating the application of Artificial Intelligence (AI) in the field of radiotherapy exhibit substantial variations in terms of quality. The goal of this study was to assess the amount of transparency and bias in scoring articles with a specific focus on AI based segmentation and treatment planning, using modified PROBAST and TRIPOD checklists, in order to provide recommendations for future guideline developers and reviewers.

Materials and methods: The TRIPOD and PROBAST checklist items were discussed and modified using a Delphi process. After consensus was reached, 2 groups of 3 co-authors scored 2 articles to evaluate usability and further optimize the adapted checklists. Finally, 10 articles were scored by all co-authors. Fleiss' kappa was calculated to assess the reliability of agreement between observers.

Results: Three of the 37 TRIPOD items and 5 of the 32 PROBAST items were deemed irrelevant. General terminology in the items (e.g., multivariable prediction model, predictors) was modified to align with AI-specific terms. After the first scoring round, further improvements of the items were formulated, e.g., by preventing the use of sub-questions or subjective words and adding clarifications on how to score an item. Using the final consensus list to score the 10 articles, only 2 out of the 61 items resulted in a statistically significant kappa of 0.4 or more demonstrating substantial agreement. For 41 items no statistically significant kappa was obtained indicating that the level of agreement among multiple observers is due to chance alone.

Conclusion: Our study showed low reliability scores with the adapted TRIPOD and PROBAST checklists. Although such checklists have shown great value during development and reporting, this raises concerns about the applicability of such checklists to objectively score scientific articles for AI applications. When developing or revising guidelines, it is essential to consider their applicability to score articles without introducing bias.

Introduction

There is a rapid increase of scientific papers on the development and use of artificial intelligence (AI) in radiation therapy [3,13,17,24]. However, with the exception of automatic segmentation for organs at risk and to some extent automated treatment planning, clinical implementation of AI models for e.g. decision support systems is low. This is partly due to the lack of large curated datasets for model building, trust and reliable human-level interpretation of these models, and consistent reproducibility of these methods for routine clinical use [7]. To be able

* Corresponding author at: Dept. Of Radiation Oncology, Catharina Hospital Eindhoven, 5632 EJ, Michelangelolaan 2, Eindhoven, The Netherlands.
E-mail address: coen.hurkmans@cze.nl (C. Hurkmans).

https://doi.org/10.1016/j.radonc.2024.110196
Received 27 November 2023; Received in revised form 29 January 2024; Accepted 26 February 2024
Available online 2 March 2024

to enhance trust and to assess the usability of a model, a structured reporting methodology on the development, testing and potential use of such model should be present which is comprehensive, unambiguous, with a focus on transparency and prevention of bias.

For prognostic and diagnostic models, such a guideline has already been developed: The Transparent Reporting of a multivariable prediction model for Individual Prognosis (TRIPOD) [11]. The TRIPOD has following sections: Title and Abstract, Introduction, Methods, Results and Discussion and encompassing a total of 34 items. The items further specify what should be reported. For example, in methods, information should be provided regarding the source of data, participants, outcome, predictors, sample size, missing data, statistical analysis methods, risk groups and development versus validation data. To further assess risk of bias, the Prediction model Risk Of Bias ASsessment Tool (PROBAST) was developed a few years later, wherein some authors contributed to both studies [31]. The PROBAST tools considers four main domains: Participants, Predictors, Outcome, and Analysis and encompasses 26 items. These items for example include questions about the appropriateness of data collection and processing or ask to describe whether pre-defined standards for collection, processing or outcome definition were followed.

Both studies have been highly cited [11] $>$ 700 times per 1–1-2023 and [31] $>$ 350 times.

The combination of both guidelines in particular provides a robust and transparent assessment on the reporting of prediction models and aims specifically to enable transparent and unbiased assessment of reports. There are, however, to the best of our knowledge, no studies that actually evaluate the consistency with which such checklists can be used to score scientific articles on transparency and bias for AI based applications. Evidently, a large inter-observer variation in the use of such checklists to score articles is highly undesirable, as it might implicate unnecessary variation or ambiguity in the assessment of bias.

Other checklists have since also been published, such as CLAIM [26], MI-CLAIM [28] SPIRIT-AI, CONSORT-AI [12,25] and journal specific ML article submission checklists [15]. These are not specifically aimed at scoring bias and are less frequently used, while SPIRIT-AI and CONSORT-AI are more focused on clinical trials. In addition, TRIPOD has already been successfully used for assessment of prognostic models in the field of radiation oncology [29]. We therefore believe that the TRIPOD and PROBAST guidelines might be best suited to score articles in a transparent way without bias introduction.

Thus, It would be of value if the TRIPOD and PROBAST guidelines could be employed for easy assessment of other predictive models, and in particular those that are AI-based, within the field of radiation oncology. However, it is clear that these guidelines cannot be applied directly without any modification or clarification. E.g., the term multivariate prediction model is not directly clear outside the context of individual prognosis prediction. As segmentation and treatment planning are the main radiotherapy fields for which AI is now used, we focused on these domains.

The goal of this study was therefore two-fold: to modify the PROBAST and TRIPOD checklists to be suitable for assessment of bias and transparency of AI-based radiotherapy studies and to assess the amount of bias in scoring articles and the causes for such assessment bias in order to give recommendations for future guideline developers and reviewers.

Materials and methods

The TRIPOD and PROBAST articles and the corresponding checklists were studied by all co-authors. A Delphi process was used to modify the checklist items to adapt them for scoring scientific papers on AI-based segmentation and treatment planning. First, checklist items that might be omitted if not pertaining to this context, or added, if deemed important were discussed. Secondly, each item was discussed in detail to determine if any adaptation was needed to increase clarity in this context. Before each discussion session, the group voted on the content of suggested item changes and the relative implementation. All PROBAST items could be scored as low/high or unclear risk of bias as in the original tool. All TRIPOD items were scored as "yes" (requirement fulfilled) or "no". Only item 6b "Report any actions to blind assessment of the outcome to be predicted" could also be scored as not applicable. For example, if only a quantitative assessment of a dose distribution is conducted, without a qualitative scoring by a radiation oncologist, this item would not apply to that specific study.

After reaching consensus on all checklist items, a pilot study with the adapted checklists was performed where 6 co-authors scored four articles (2 groups of 3 co-authors scoring the same 2 articles) [1,5,23,27]. Each author replied to the checklist items independently. Replies were subsequently collected and analyzed. Results from this analysis were discussed to identify potential improvements of the checklist items.

Answers to items were scored as in agreement if all reviewers answered the same for a TRIPOD item. Otherwise, they were scored as not in agreement. For the PROBAST items, a score of some agreement was also possible if at least one observer answered "unclear" and the rest answered unanimously "yes" or unanimously "no".

After the pilot study, the final list of checklist items was used to score 10 articles which were a sample of the current literature based on various domain specific journals in the field and a mix of topics [2,4,6,9,16,21,22,30,32,33]. Seven co-authors conducted the scoring to assess the consistency of evaluations between observers.

Fleiss' kappas were calculated for the scores of the 10 articles to assess the reliability of agreement between the observers. This measure corrects for the rate of observers that agree on the answer to a question by chance. kappa $= -1$ indicates observed dis-agreement, kappa0 indicates that agreement was no better than chance, kappa $= 1$ indicates perfect agreement between observers. For values below 0, the agreement is less than the agreement expected by chance, while for values above 0 they are more than expected by chance. Although there is no general consensus on this, a kappa between 0.41 and 0.60 is said to indicate a moderate (fair to good) agreement and a kappa between 0.61 and 0.80 indicates a substantial agreement [18]. The null hypothesis assumes that the level of agreement among multiple observers is due to chance alone. On the other hand, the alternative hypothesis suggests that the level of agreement among the observers is not solely due to random variation (chance alone).Hence, rejecting the null hypothesis indicating a meaningful degree of agreement that cannot be explained by random variation.

The kappa assesses the reliability of agreement. Yet, for kappa to have statistical relevance for inference, they must exhibit statistical significance themselves. Thus, p-values for the calculation of kappa were calculated. Kappa was considered significant if $p < 0.05$.

Descriptive statistics were also calculated to provide more insight into the inter-observer agreement and distinctiveness of each checklist item. For any given item, the number of articles scored identically by at least 6 of the 7 observers was calculated to quantify inter-observer agreement. If less than 6 observers scored a question for a specific article identically, it indicated a relatively low inter-observer agreement. This accounts for the fact that with a yes/no answer scored by 7 observers necessarily, at least four observers will have to provide an identical answer.

To quantify the distinctiveness of a checklist item, the number of observers that did not provide the same reply to the item for all 10 papers was calculated. If several observers provided the same reply on a checklist item for all 10 articles, it indicates that there is not enough diversity pertained to that item within the articles scored, i.e., the information is always provided (or not) in the articles. This makes such an item not very distinctive in scoring or ranking articles.

Results

A complete list of the TRIPOD and PROBAST checklist items and the adaptations made based on the Delphi process and pilot study can be

found in Supplement A.

There were a few general modifications implemented to the TRIPOD and PROBAST terminology. The term "(multivariable) prediction model" was replaced by "AI model" and the term "predictors" by "input parameters (of the AI model)". "Outcome" was replaced by "AI model output".

Original checklist items 10c: "For validation, describe how the predictions were calculated.", 11:" Provide details on how risk groups were created, if done." and 14b: "If done, report the unadjusted association between each candidate predictor and outcome.", were deemed not applicable in this context and were removed.

Some PROBAST items were adapted, like item 2a:" List and describe predictors included in the final model, e.g. definition and timing of assessment: ", which was adapted to: "List and describe AI model input parameters included in the final model, e.g., definition and timing of assessment, imaging modalities used for planning, and the respective parameters like CT slice thickness or dose voxel size etc." This would give the scorer a better idea of which input parameters should be given in the article in order to score a lower risk of bias.

Original checklist items 2b:"Concern that the definition, assessment or timing of predictors in the model do not match the review question", 3.6:"Was the time interval between predictor assessment and outcome determination appropriate?", 3b: "At what time point was the outcome determined: If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome", 4.5: "Was selection of predictors based on univariable analysis avoided?" and 4.9:" Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?", were considered not applicable and were removed.

The results of the pilot study (scoring of 2 articles per group of 3 observers) is given in Fig. 1 for all four papers combined.

More specific adaptations to checklist items were implemented based on the pilot study.

Ways to obtain a better consensus were formulated:

1. Checklist items composed of sub-questions/summations (e.g., "e.g., objectives, sample size, input parameters, statistical analysis, study design and conclusions") could be further clarified or clarified how this should be scored if this is partly answered.
2. Items with subjective words like "clearly", "appropriate" or "explain" can be problematic as leaving more space to subjective interpretation and may even drift over time if the field gets more mature. It was however decided not to replace them as otherwise these items would start to substantially deviate from the original items.
3. Some items did not appear to have a high level of discriminative ability between studies. It was decided not to delete these items and to perform the final analysis including them in order to test this effect in a larger set of articles.
4. The different background of the observers could lead to a different interpretation of a checklist item. Although this was noticed, no strategy on how to prevent this could be identified.
5. Some extra guidance could be given for a number of items to the reviewers to improve the consistency of item evaluation.

Ad 1:

For example, for TRIPOD item 2, only the most important examples given needed to be present (summary of objectives, study design, setting, participants, sample size, input parameters [anatomical and/or dosimetric features], model output, statistical analysis, results, and conclusions).

Ad 5:

For example, the distinction between development and validation of a model was further clarified for TRIPOD item 1. Also, for TRIPOD item 8: "Explain how the study size was arrived at", it was discussed that no guidelines have been published on this yet. As long as some reason is given, this could be scored as "yes".

The results for the final scoring of 10 articles by all observers is given next. To clarify the analysis, a more detailed presentation of the results for a specific item (4b according to TRIPOD) is provided initially. The detailed results for item 4b: "Describe how the model was developed (for example in regards to modelling technique (type of model) and input parameter selection)", are given in Table 1. The inter-observer agreement is high for 8 out of the 10 articles. Furthermore, all observers have limited variation in their replies across the 10 articles scored. Thus, it seems this checklist item can consistently differentiate papers from each other. The corresponding kappa was 0.58 ($p < 0.001$), the highest value observed.

The results for all adapted checklist items can be found in Fig. 2.

In the following paragraph, the statistically significant kappa results are discussed ($p < 0.05$).

A statistically significant kappa was found for only 19 out of 61 items. Out of these 19 items, only 2 had a kappa of 0.4 or more, indicating at least a moderate degree of agreement (item 4b and item 22).. These items could be considered as the best items to judge transparency and risk of bias of an article. The remaining 17 kappas ranged between −0.16 to 0 and 0 to 0.26. For the negative values, this means a less than moderate degree of disagreement. For the positive values, this means a less than moderate degree of agreement.. Kappa reduces when an observer always provides the same reply for each article or when larger

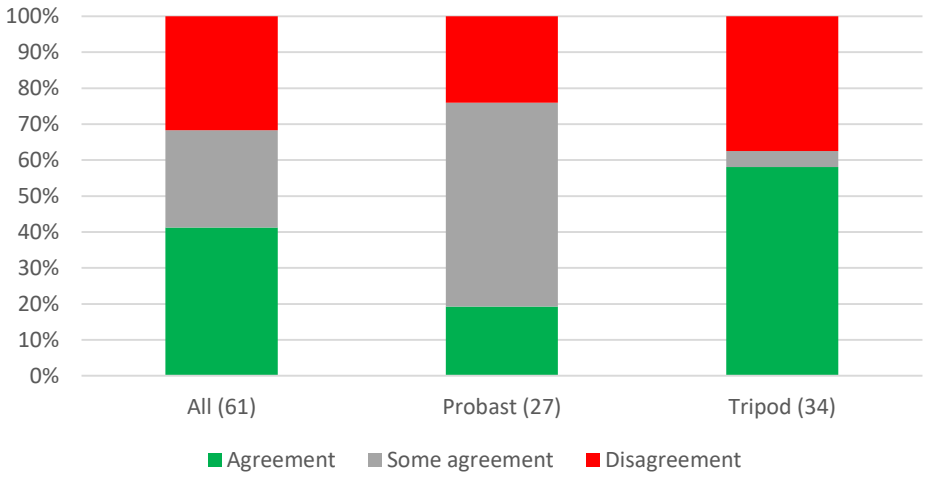

Fig. 1. Percentage of agreement in answers to the adapted PROBAST and TRIPOD items combined for the 4 papers scored in the pilot study by 3 observers for each article.

Table 1
Answers on checklist item 4b "Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up" in the final scoring of 10 articles by the 7 co-authors. Articles with low inter-observer agreement are marked grey. The distinctiveness of the item was good as no observer gave the same answer for every article.

	Observer						
Article	a	b	c	d	e	f	g
1	No	No	No	No	No	No	No
2	Yes	Yes	Yes	Yes	Yes	Yes	Yes
3	No	Yes	No	No	No	No	No
4	Yes	Yes	Yes	Yes	Yes	Yes	Yes
5	Yes	Yes	Yes	Yes	Yes	Yes	Yes
6	No	Yes	Yes	Yes	No	Yes	Yes
7	No	Yes	Yes	No	No	No	No
8	No	No	Yes	No	No	No	No
9	No	No	Yes	No	No	No	No
10	No	No	Yes	No	No	No	No

inter-observers variations exist. For 3 out of these 17 items the distinctiveness was not (very) low and inter-observer variation not large (items 7a, 10b and 18). The corresponding kappas were 0.15 ($p = 0.03$), 0.15 ($p = 0.03$) and 0.26 ($p < 0.01$), respectively. This seems to indicate that these items at least have the potential to contribute to identification of transparency and bias in an article.

In the following paragraph, the not statistically significant kappa results are discussed ($p > 0.05$).

Interestingly, for 42 checklist items no significant kappa could be calculated. For only 2 out of these 42 items the distinctiveness was not very low and inter-observer variation not large (items 10a and 13c). With kappas of 0.11 ($p = 0.13$) and 0.02 ($p = 0.76$) it seems to indicate that the potential of these items to contribute to the identification of transparency and bias in an article is lower than for items 7a, 10b and 18 mentioned above).

Discussion

The TRIPOD and PROBAST checklists were adapted to be used for scoring articles on AI based segmentation and planning in radiotherapy using a Delphi process. Results based on the scoring of 61 checklist items for 10 articles by 7 observers showed very low reliability of agreement between evaluators. Items 4b "Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up" and

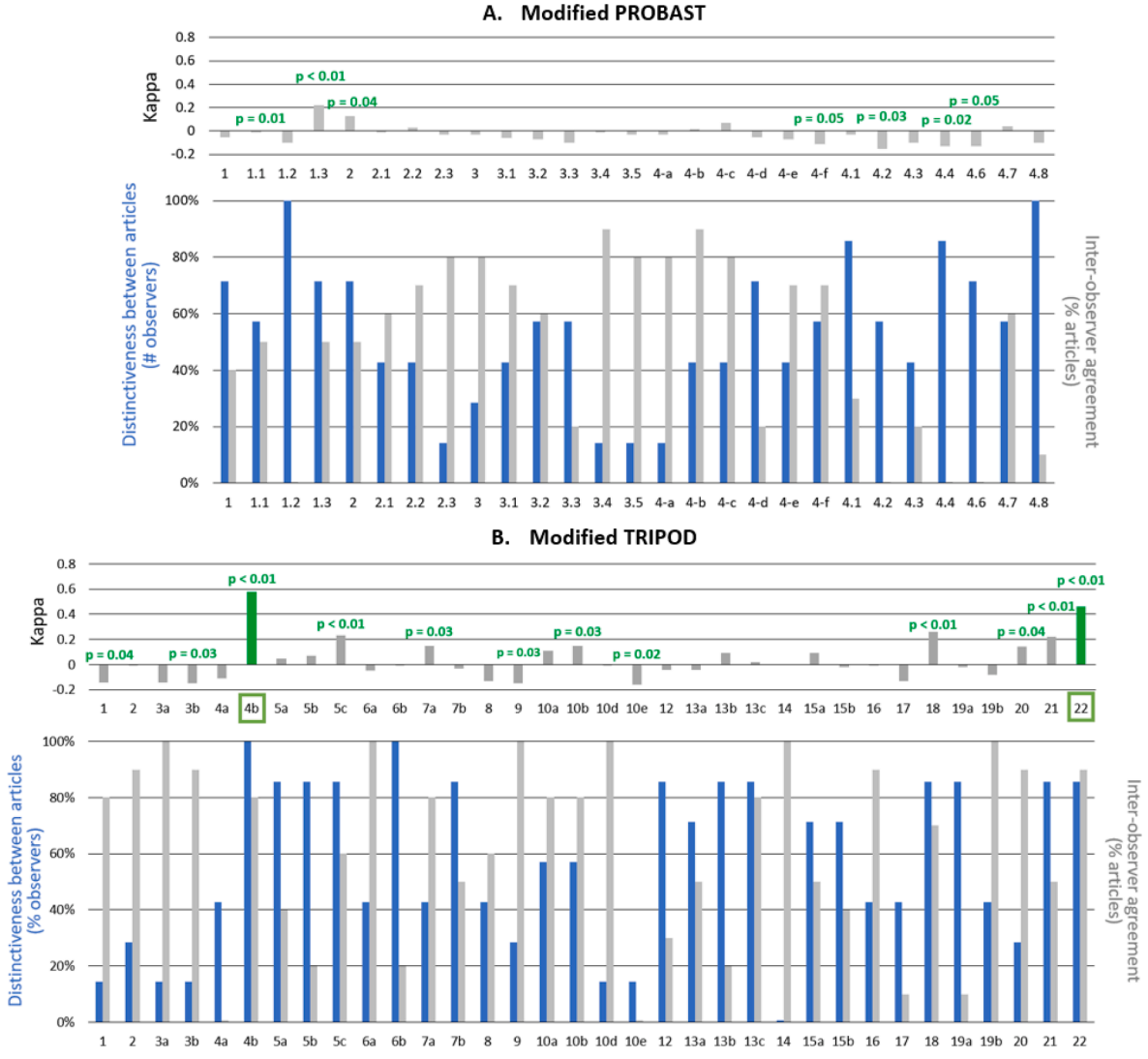

Fig. 2. Results for all checklist items separately. Kappas which could be determined with a p value of 0.05 or less are indicated with their p-value. Kappas of >0.4 with p value <0.05 are indicated with green bars. Distinctiveness of an item (% observers with not the same answer for all articles) is given in blue bars. Inter-observer agreement (% articles scored identical by 6 or 7 observers) is given in grey bars. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

22" Give the source of funding and the role of the funders for the present study." were the only 2 items with kappas above 0.4. To score the latter item as sufficient it was further clarified to accept it if the funding source was given only; info on whether they had any influence on the study (role) was not mandatory. We believe these high values are due to the fact that these are well-defined questions with a straightforward way for scoring, leading to high interobserver agreement, combined with enough variation in the articles regarding these items resulting in a high distinctiveness between articles. Although not AI or radiotherapy specific, the way these items are formulated together with the guidance on how to score them, might be good examples on how future items could be developed (e.g., for TRIPOD-AI or PROBAST-AI).

Scoring systems are often used to evaluate scientific papers, but a statistical analysis of the distinctiveness or inter-observer agreement of such scoring systems is often not performed. In an interesting article of the influence of Likert-type scoring on editor's decisions to accept or reject articles, the statistical analysis showed considerable heterogeneity in scores leading to rejection or acceptance. It was noted that both the details in the article and the review comments were important [8]. Additional studies involving multiple observers are needed to assess the efficacy of scoring systems, identify the most effective ones, and explore opportunities for improvement.

The TRIPOD checklist has been employed many times in the past and can be readily used for classification of prediction models as type 1a, development, up to type 4, external validation [29]. However, to the best of our knowledge no reliability of agreement studies have been performed in scoring articles using the complete checklist.

A guideline and checklist specific to radiotherapy planning studies has been proposed previously by Hansen et al. [19]. The overall aim of their framework is to improve the scientific quality of treatment planning studies and papers, but the authors pointed out it might also be used by reviewers and journal editors to support the evaluation of the reporting in scientific manuscripts of planning studies. Also for this checklist, which they called Radiotherapy Treatment planning study Guidelines (RATING), reliability of agreement studies have not been performed and it was commented shortly after publication that there will naturally be a variance in the RATING scores achieved for any specific study or paper due to their subjective evaluation and the fact that all the questions were constrained to binary responses [20].

Similarly, the Medical Physics journal published guidelines for authors wanting to publish on AI [15]. This concise guideline is very helpful although not validated yet.

There is always a balance between the level of details and length of guidelines. While too lengthy guidelines may not be well read or less general applicable, the omission of sufficient details or use of subjective general terms might hamper the objective interpretation of a guideline and the consistent scoring of articles based on such guidelines. For example, the question if "relevant patient demographics" [15] are given might for the same data given lead to different answers from different observers, while "is the age distribution and Body Mass Index given by means of a median and interquartile range" would be scored identical by different observers. The word "relevant" is subjective. We recommend future guideline authors to refrain as much as possible from using subjective wordings in their checklists. Also, composite questions are often used, like "is the included number of patients stated, explained and justified?". It is then often not explained how this item should be scored if 1 or 2 out of the 3 sub-questions is answered [19]. Furthermore, many checklist state one should "describe" certain topics or that "details" should be given or choices "explained" e.g., [25,28]. We observed that the use of such words sparked discussions. One could deem that such an item is sufficiently adhered to if some description/detail/explanation is given, but one could also be more rigorous and require a much more elaborate description/detail/explanation. With this observation, we recommend such words should be avoided as much as possible in future guideline. Also, it might be worthwhile to revise some of the current guidelines to address the points raised here.

Another initiative to provide information of the whole creation pipeline of AI solutions, of the datasets used to develop AI, along with their biases, is the creation of Machine Learning Canvas, Datasets for Datasheets, and Model cards. Application examples of this methodology to radiotherapy have been given in Biase et al. [14]. A new initiative to make the TRIPOD and PROBAST scoring systems applicable for AI studies is currently running [10]. We eagerly anticipate the results and foresee it will be another step in the continuous effort to score and reduce bias and to improve transparency in AI based studies. Our results may potentially be of interest for these authors too.

Our study has several limitations. First, although the observers all have ample experience in writing and evaluating radiotherapy planning and segmentation studies, they are not experts in bias estimations and were not involved in the TRIPOD and PROBAST development itself. However, the observers do belong to the categories of readers of such articles, which we believe are mostly medical physicists, radiation oncologists or researchers in the field of radiotherapy treatment planning. Secondly, we selected only 10 articles and these were scored by only 7 observers. Judging from the fact that the kappas could often not be predicted ($p > 0.05$) with sufficient accuracy, a larger number of articles or observers might have improved these statistics. In fact, almost all items have a binary outcome, and the kappa analysis is particularly sensitive to prevalence of outcomes.

On the other hand, if the variation in bias and transparency in articles in this domain is inherently low, a larger number of articles may still lead to similar results. We would like to stress that the fact that low kappa values were found for most items, does not mean the information requested in the items is not relevant to be reported in the articles.

Conclusions

There are several frameworks developed aiming at standardized and transparent development and reporting of prediction studies, some specifically developed for AI and/or radiation therapy. Although they all have specific merits, the reliability of agreement using the corresponding checklists to score scientific papers is rarely investigated. Our study showed low reliability scores of the TRIPOD and PROBAST checklists adapted for use on AI papers for segmentation and treatment planning in radiotherapy. Suggestions to improve new guidelines are presented.

CRediT authorship contribution statement

Coen Hurkmans: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jean-Emmanuel Bibault:** Methodology, Validation, Writing – review & editing. **Enrico Clementel:** Conceptualization, Methodology, Validation, Writing – review & editing. **Jennifer Dhont:** Methodology, Validation, Writing – review & editing. **Wouter van Elmpt:** Methodology, Validation, Writing – review & editing. **Georgios Kantidakis:** Methodology, Validation, Writing – review & editing. **Nicolaus Andratschke:** Methodology, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the invaluable help by Marjan Sharabiani throughout the project which led to this article.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.radonc.2024.110196.

References

[1] Ahn SH, Kim E, Kim C, et al. Deep learning method for prediction of patient-specific dose distribution in breast cancer. Radiat Oncol 2021;16:154-01864. https://doi.org/10.1186/s13014-021-01864-9.

[2] Ahn SH, Yeo AU, Kim KH, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. Radiat Oncol 2019;14:213–1392. https://doi.org/10.1186/s13014-019-1392-z.

[3] Almberg SS, Lervôg C, Frengen J, et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. Radiother Oncol 2022;173:62–8. https://doi.org/10.1016/j.radonc.2022.05.018.

[4] Ambroa EM, Pérez-Alija J, Gallego P. Convolutional neural network and transfer learning for dose volume histogram prediction for prostate cancer radiotherapy. Med Dosim 2021;46:335–41. https://doi.org/10.1016/j.meddos.2021.03.005.

[5] Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. Knowledge-based automated planning with three-dimensional generative adversarial networks. Med Phys 2020;47:297–306. https://doi.org/10.1002/mp.13896.

[6] Bai P, Weng X, Quan K, et al. A knowledge-based intensity-modulated radiation therapy treatment planning technique for locally advanced nasopharyngeal carcinoma radiotherapy. Radiat Oncol 2020;15:188–01626. https://doi.org/10.1186/s13014-020-01626-z.

[7] Balagurunathan Y, Mitchell R, El Naqa I. Requirements and reliability of AI in the medical context. Phys Med 2021;83:72–8. https://doi.org/10.1016/j.ejmp.2021.02.024.

[8] Callaham M, John LK. What does it take to change an editor's mind? identifying minimally important difference thresholds for peer reviewer rating scores of scientific articles. Ann Emerg Med 2018;72:314–8. https://doi.org/10.1016/j.annemergmed.2017.12.004.

[9] Campbell WG, Miften M, Olsen L, et al. Neural network dose models for knowledge-based planning in pancreatic SBRT. Med Phys 2017;44:6148–58. https://doi.org/10.1002/mp.12621.

[10] Collins GS, Dhiman P, Andaur Navarro CL, 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:e048008. https://doi.org/10.1136/bmjopen-2020-048008.

[11] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). Ann Intern Med 2015;19:735–6. https://doi.org/10.7326/M14-0698.

[12] Cruz RS, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–63. https://doi.org/10.1038/s41591-020-1037-7.

[13] de Hond YJM, Kerckhaert CEM, van Eijnatten MAJM, et al. Anatomical evaluation of deep-learning synthetic computed tomography images generated from male pelvis cone-beam computed tomography. Phys Imaging Radiat Oncol 2023;25:100416. https://doi.org/10.1016/j.phro.2023.100416.

[14] de BA, Sourlos N, van Ooijen PMA. Standardization of artificial intelligence development in radiotherapy. Semin Radiat Oncol 2022;32:415–20. https://doi.org/10.1016/j.semradonc.2022.06.010.

[15] El Naqa I, Boone JM, Benedict SH, et al. AI in medical physics: guidelines for publication. Med Phys 2021;48:4711–4. https://doi.org/10.1002/mp.15170.

[16] Gronberg MP, Beadle BM, Garden AS, et al. Deep learning-based dose prediction for automated, individualized quality assurance of head and neck radiation therapy plans. Pract Radiat Oncol 2023;13:e282–91. https://doi.org/10.1016/j.prro.2022.12.003.

[17] Guerreiro F, Seravalli E, Janssens GO, et al. Deep learning prediction of proton and photon dose distributions for paediatric abdominal tumours. Radiother Oncol 2021;156:36–42. https://doi.org/10.1016/j.radonc.2020.11.026.

[18] Gwet KL. Handbook of inter-rater reliability. 2014.

[19] Hansen CR, Crijns W, Hussein M, et al. Radiotherapy Treatment plannINg study Guidelines (RATING): a framework for setting up and reporting on scientific treatment planning studies. Radiother Oncol 2020;153:67–78. https://doi.org/10.1016/j.radonc.2020.09.033.

[20] Hansen CR, Crijns W, Hussein M, et al. Response to the letter to the editor "application of the RATING score: in regards to Hansen et al.". Radiother Oncol 2021;158:311. https://doi.org/10.1016/j.radonc.2021.01.012.

[21] Hedden N, Xu H. Radiation therapy dose prediction for left-sided breast cancers using two-dimensional and three-dimensional deep learning models. Phys Med 2021;83:101–7. https://doi.org/10.1016/j.ejmp.2021.02.021.

[22] Huang M, Feng C, Sun D, Cui M, Zhao D. Segmentation of clinical target volume from CT images for cervical cancer using deep learning. Technol Cancer Res Treat 2023;22:15330338221139164. https://doi.org/10.1177/15330338221139164.

[23] Jiao SX, Wang ML, Chen LX, Liu XW. Evaluation of dose-volume histogram prediction for organ-at risk and planning target volume based on machine learning. Sci Rep 2021;11:3117–82749. https://doi.org/10.1038/s41598-021-82749-5.

[24] Lenkowicz J, Votta C, Nardini M, et al. A deep learning approach to generate synthetic CT in low field MR-guided radiotherapy for lung cases. Radiother Oncol 2022;176:31–8. https://doi.org/10.1016/j.radonc.2022.08.028.

[25] Liu X, Cruz RS, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020;2:e537–48. https://doi.org/10.1016/S2589-7500(20)30218-1.

[26] Mongan J, Moy L, Kahn Jr CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2:e200029.

[27] Nguyen D, Jia X, Sher D, et al. 3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture. Phys Med Biol 2019;64:065020. https://doi.org/10.1088/1361-6560/ab039b.

[28] Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320–4. https://doi.org/10.1038/s41591-020-1041-y.

[29] Sharabiani M, Clementel E, Andratschke N, Hurkmans C. Generalizability assessment of head and neck cancer NTCP models based on the TRIPOD criteria. Radiother Oncol 2020;146:143–50. https://doi.org/10.1016/j.radonc.2020.02.013.

[30] van de SD, Sharabiani M, Bluemink H, et al. Artificial intelligence based treatment planning of radiotherapy for locally advanced breast cancer. Phys Imaging Radiat Oncol 2021;20:111–6. https://doi.org/10.1016/j.phro.2021.11.007.

[31] Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51–8. https://doi.org/10.7326/M18-1377.

[32] Yang J, Zhao Y, Zhang F, Liao M, Yang X. Deep learning architecture with transformer and semantic field alignment for voxel-level dose prediction on brain tumors. Med Phys 2023;50:1149–61. https://doi.org/10.1002/mp.16122.

[33] Yu J, Goh Y, Song KJ, et al. Feasibility of automated planning for whole-brain radiation therapy using deep learning. J Appl Clin Med Phys 2021;22:184–90. https://doi.org/10.1002/acm2.13130.