Systematic review of machine-learning models in orthopaedic trauma an overview and quality assessment of 45 studies

On behalf of the Machine Learning Consortium

DOI
10.1302/2633-1462.51.BJO-2023-0095.R1

Publication date
2024

Document Version
Final published version

Published in
Bone and Joint Open

Citation (APA)
On behalf of the Machine Learning Consortium (2024). Systematic review of machine-learning models in orthopaedic trauma an overview and quality assessment of 45 studies. *Bone and Joint Open*, *5*(1), 9-19. https://doi.org/10.1302/2633-1462.51.BJO-2023-0095.R1

Systematic review of machine-learning models in orthopaedic trauma

an overview and quality assessment of 45 studies

Cite this article:
Bone Jt Open 2024;5(1):
9–19.

DOI: 10.1302/2633-1462.
51.BJO-2023-0095.R1

Correspondence should be sent to H. Dijkstra h.b.dijkstra@umcg.nl

H. Dijkstra,[1,2] A. van de Kuit,[1] T. de Groot,[1,3] O. Canta,[1] O. Q. Groot,[4] J. H.F. Oosterhoff,[5] J. N. Doornberg,[1,6] On behalf of the Machine Learning Consortium

[1]Department of Orthopaedic Surgery, University Medical Centre Groningen, Groningen, Netherlands
[2]University Center for Geriatric Medicine, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
[3]Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[4]Department of Orthopaedic Surgery, University Medical Centre Utrecht, University of Utrecht, Utrecht, Netherlands
[5]Department of Engineering Systems & Services, Faculty Technology Policy and Management, Delft University of Technology, Delft, Netherlands
[6]Department of Orthopaedic Trauma Surgery, Flinders Medical Center, Flinders University, Adelaide, Australia

Aims

Machine-learning (ML) prediction models in orthopaedic trauma hold great promise in assisting clinicians in various tasks, such as personalized risk stratification. However, an overview of current applications and critical appraisal to peer-reviewed guidelines is lacking. The objectives of this study are to 1) provide an overview of current ML prediction models in orthopaedic trauma; 2) evaluate the completeness of reporting following the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement; and 3) assess the risk of bias following the Prediction model Risk Of Bias Assessment Tool (PROBAST) tool.

Methods

A systematic search screening 3,252 studies identified 45 ML-based prediction models in orthopaedic trauma up to January 2023. The TRIPOD statement assessed transparent reporting and the PROBAST tool the risk of bias.

Results

A total of 40 studies reported on training and internal validation; four studies performed both development and external validation, and one study performed only external validation. The most commonly reported outcomes were mortality (33%, 15/45) and length of hospital stay (9%, 4/45), and the majority of prediction models were developed in the hip fracture population (60%, 27/45). The overall median completeness for the TRIPOD statement was 62% (interquartile range 30 to 81%). The overall risk of bias in the PROBAST tool was low in 24% (11/45), high in 69% (31/45), and unclear in 7% (3/45) of the studies. High risk of bias was mainly due to analysis domain concerns including small datasets with low number of outcomes, complete-case analysis in case of missing data, and no reporting of performance measures.

Conclusion

The results of this study showed that despite a myriad of potential clinically useful applications, a substantial part of ML studies in orthopaedic trauma lack transparent reporting, and are at high risk

Bone & Joint Open

of bias. These problems must be resolved by following established guidelines to instil confidence in ML models among patients and clinicians. Otherwise, there will remain a sizeable gap between the development of ML prediction models and their clinical application in our day-to-day orthopaedic trauma practice.

Take home message

- Useful applications of machine-learning prediction models in orthopaedic trauma exist, but a substantial proportion lack external validation and transparent reporting, and are at high risk of bias.

Introduction

Machine learning (ML) has shown great potential in aiding clinicians with different tasks in orthopaedic trauma.[1,2] Specific applications of artificial intelligence (AI) and ML have emerged, such as risk stratification methods serving as decision support tools for prediction of a diagnostic (e.g. suspected fracture)[3-5] or prognostic outcome of interest (e.g. postoperative delirium, mortality estimation, infection, or risk of revision surgery);[6-9] the latter is often referred to as prognostic modelling. ML-driven probability calculators and prediction models have the great – theoretical – potential to assess individual patients' risk stratification in order to support shared decision-making, and drive personalize care.

However, integration of these decision support tools in clinical practice remains challenging. There is a substantial gap between the myriad of published articles on ML models in orthopaedic trauma, and the models that reach patients and surgeons in our day-to-day practice. Before these decision support tools can be implemented in clinical practice, external validation and prospective testing should be performed.[10] External validation refers to evaluating the model's predictive performance on an independent dataset that was not used during model development. Generalizability of a prediction model cannot be assessed after a single external validation study, but it should be examined after thorough independent external validation for each population if the population differs considerably in setting, in patient demographics or outcome incidence. As a first step towards clinical implementation, future studies reporting on the development of decision support tools should be of sufficient quality in terms of transparency and completeness of reporting by adhering to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement.[11] This statement consists of recommendations for developing, validating, and/or updating a prediction model whether for diagnostic or prognostic purposes. To assess the risk of bias of studies developing decision support tools, the Prediction model Risk Of Bias Assessment Tool (PROBAST) was developed.[12] This tool enables structured bias assessment in four domains: participants, predictors, outcome, and analysis in prediction model studies.

Recent systematic reviews suggested that studies reporting on ML in the field of elective orthopaedic surgery are at high risk of bias, lack transparency and complete reporting, and often are not externally validated.[2,9,13] Lans et al[14] recently showed that reporting relevant information is also limited in diagnostic ML imaging studies in the field of orthopaedics. To our knowledge, no studies report on the application of ML prediction models focusing on structured data in the field of orthopaedic trauma. The goal of this study is to identify details that researchers commonly fail to provide in order to improve the quality of ML models papers in the field of orthopaedic trauma surgery. By doing this, we attempt to improve our understanding of the gap between ML models in orthopaedic trauma scientific domain, and our day-to-day orthopaedic trauma practice, in order to eventually close the gap.

Therefore, in this systematic review, we will: 1) provide an overview of current ML models in orthopaedic trauma; 2) evaluate the completeness of reporting following the TRIPOD statement; and 3) assess the risk of bias following the PROBAST tool.

Methods

Systematic literature search

The study was performed in accordance to the PRISMA guidelines,[15] and the checklist was added (Supplementary Table i). The study was not registered. A systematic search of the available literature was performed in PubMed, EMBASE, and Cochrane Library up to 10 January 2023 (Supplementary Table ii). Two domains of medical subject headings (MeSH) terms and keywords were combined with "AND" and within the two domains the terms were combined with "OR". The first domain included words related to ML, and the second domain to orthopaedic trauma surgery. Terms were restricted to MeSH, title, abstract, and keywords. Titles and abstracts were independently screened by four reviewers (HD, AK, TG, OC). Subsequently, full-text articles were then independently assessed for eligibility by the same four reviewers. Discrepancies between the four reviewers were assessed by two orthopaedic trauma research fellows (OQG, JHFO).

Eligibility criteria

Inclusion and exclusion criteria are shown in Table I.

Data extraction

Four reviewers extracted the following data of the included studies (manually): year of publication, use of national or registry database (yes/no), study type (development, validation or both in one study), external validation study (yes/no), study goal (i.e. diagnostic, prognostic, or classification), injury type, number of patients with the outcome (if applicable), predicted outcome, type of algorithm used (e.g. support vector machine), performance measures (e.g. area under the receiver operating characteristic (ROC) curve (area under the curve (AUC), discrimination), calibration metrics (calibration reflects the agreement between the observed

Table I. Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
1. ML based probability calculators or prediction models in orthopaedic trauma outcome studies (i.e. diagnostic and prognostic modelling studies).	1. Non-English studies
2. Orthopaedic trauma studies were defined as studies investigating primary injuries to the (musculo-) skeletal system (including bones and/or tendons and/or joints and/or muscles and/or soft-tissue) with the following injury types: fractures, ruptures, dislocations, and sprains.	2. Non-relevant study types such as animal studies, letters to the editors, case-reports, and reviews.
	3. Non-ML techniques. Advanced logistic regression models such as penalized logistic regression (LASSO, ridge, or elastic-net), boosted logistic regression and bagged logistic regression[16] were considered as ML.
	4. Studies reporting on ML models for diagnostic imaging in orthopaedic surgery.
	5. Studies reporting on ML models for NLP.
	6. Oral, maxillofacial, and ophthalmological studies.

LASSO, least absolute shrinkage and selection operator; ML, machine learning; NLP, natural language processing.

outcome and the predicted probability), Brier score (a composite of discrimination and calibration), decision-curve analysis, precision-recall,[16,17] availability of a digital application (yes/no), TRIPOD items, and PROBAST domains. Two reviewers (HD, OC) rated adherence to the TRIPOD statement, and two other reviewers (AK, TG) rated the PROBAST tool domains, under the direct supervision of two experienced reviewers (OQG, JHFO).

TRIPOD statement

The TRIPOD statement consists of 22 main items. Overall, 13 of the 22 items had no subitems, seven items had two subitems, and two items had three subitems. Six (sub)items refer to model updating or external model validation and were therefore only extracted in studies performing external validation.[11] Some items could be scored with "referenced" (e.g. item 6a). Referenced was considered "completed" and included when calculating the completeness of reporting. Each item may consist of multiple elements. Both elements must be scored with "yes" for the item to be scored "completed". If a study reported on multiple prediction models (e.g. prediction model for one-year and five-year survival), we extracted only data on the best performing model.

PROBAST tool

PROBAST assesses the risk of bias in prediction model studies.[12] This tool consists of 20 signalling questions across four domains: participant selection (one), predictors (two), outcome (three), and analysis (four). Each domain is rated as having a 'low', 'high', or 'unclear' risk of bias. 'Unclear' indicates that the reported information is insufficient: no reliable judgment on low or high risk of bias can be made. Participant selection (domain one) covers potential sources of bias in the origin of data and criteria for participant selection to assess whether patient inclusion has been performed adequately. Predictors (domain two) should include a list of all considered predictors, a clear definition and timing of measurement. An outcome (domain three) should include clear definitions and timing of measurements, and a description of the time interval between predictor assessment and outcome determination. Finally, analysis (domain four) covers potential sources of bias related to inappropriate analysis methods, or omission of key performance measures, such as discrimination and calibration.

The ratings of the four domains resulted in an overall judgement of the risk of bias. In accordance with the PROBAST checklist, low overall risk of bias was assigned to a study when each domain scored low.[12] High overall risk of bias was assigned when at least one domain was judged to be high risk of bias. Unclear overall risk of bias was noted if at least one domain was judged unclear and all other domains low. The four domains and the overall judgment were reported.

Statistical analysis

The degree of completeness of reporting to TRIPOD statement and PROBAST domains were calculated and presented with percentages using medians with interquartile ranges (IQRs). Additionally, graph bars visualized the completeness of each guideline. We used Excel version 16 (Microsoft, USA) to extract and record data, and SPSS version 26 (IBM, USA) for statistical analysis.

Results

Study selection

In total, 3,951 studies were identified and 3,252 unique studies remained after duplicate removal. Title/abstract screening resulted in 147 potentially relevant studies. Of those, 45 met the inclusion criteria (Figure 1).[5,7,8,18–58] Studies were excluded because they did not meet the inclusion criteria (e.g. elective surgery, prediction of fractures, no application of artificial intelligence, imaging study) or full text was unavailable.

Overview of ML models

Most studies (93%; 42/45) were prognostic, developmental studies (Table II, Supplementary Table iii). Electronic health records (EHRs) (38%, 17/45) and national databases or registries (29%; 13/45) were often used to develop prediction models. Most studies originated from the USA or China (Figure 2). The most frequently reported ML algorithms were support vector machine (38%; 17/45), random forest (36%; 16/45), and the artificial neural networks (31%; 14/45). The median number of patients included was 875 (IQR 259 to 6,975). The most commonly reported outcome domains were mortality (33%; 15/45), and length of hospital stay (9%; 4/45). Most studies focused on injuries of the lower limb (91%; 41/45). Four studies included both development and external validation of

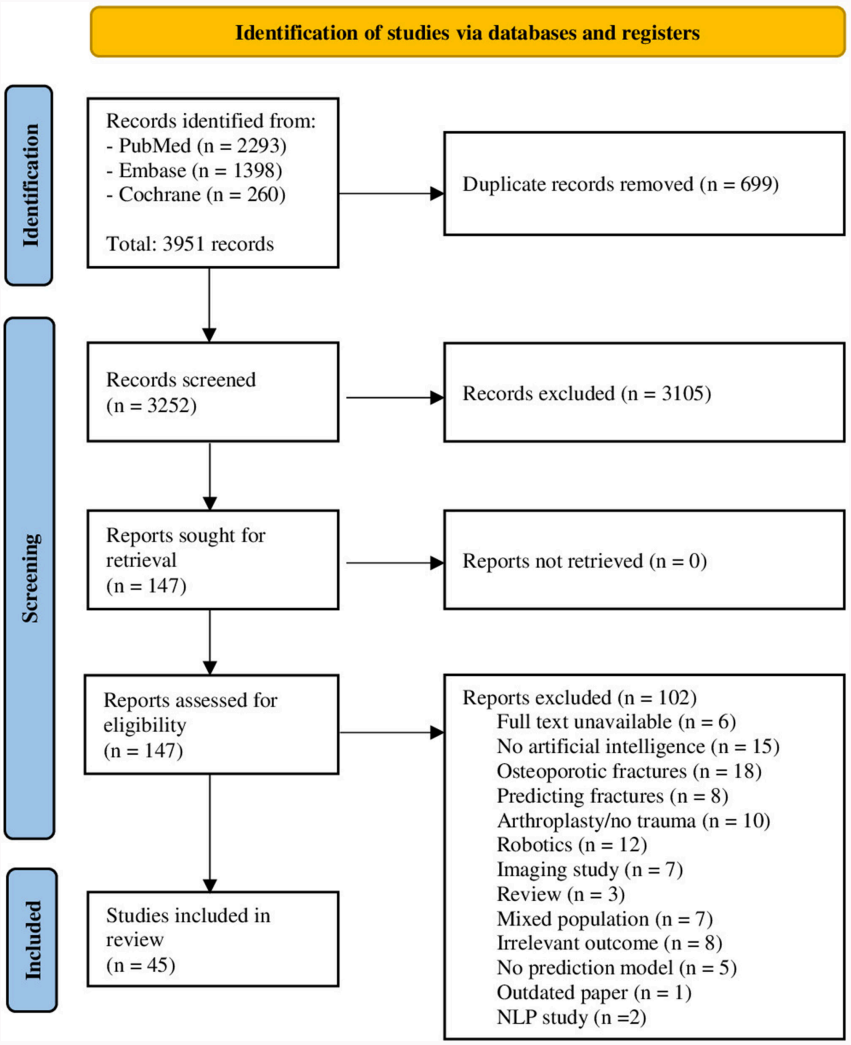

Fig. 1
PRISMA flowchart of study inclusions and exclusions. NLP, natural language processing.

the ML models in the same study (9%; 4/45). One performed single external validation.

TRIPOD statement

The overall median completeness for the TRIPOD statement was 62% (IQR 30 to 81) (Figure 3, Supplementary Table iv). Method items adhered to a median completeness of 62% (IQR 40 to 82) results items to 40% (IQR 22 to 52) and discussion items to 96% (IQR 87 to 99) (Figure 3). Six items were reported in > 90% and nine items in < 25% of studies (Table III). No study reported risk group creation. Of the items referring to model updating or external model validation, the completeness was low, with items 10c (how predictors were handled in validation) and 17 (reporting any results from model updating) in none of the studies.

PROBAST tool

The overall risk of bias was low in 24% (11/45), high in 69% (31/45), and unclear in 7% (3/45) (Figure 4, Supplementary Table v). The high risk of bias was mainly due to the analysis domain (64%; 29/45). The risk of bias was predominantly low in the other three domains (participant selection, predictors, and outcome). In the predictors domain, studies lacked reporting information about when the model was intended to be used. In the outcomes domain, studies lacked information about the time interval between predictor assessments and the outcome determination. Studies were mainly rated as high in the analysis domain due to small datasets with low numbers of events, complete-case analysis in case of missing data, and no information about complexities in the data (e.g. competing risk analysis). The AUC was often the only reported performance measure, and was presented in 96% (43/45) of the studies. Calibration metrics were reported in only in 42% (19/45) of the studies and the Brier score in only in 24% (11/45).

Discussion

In this systematic review, we provided an overview and assessed the transparency and quality of reporting of papers reporting on decision support tools using ML in the field of orthopaedic trauma surgery. We aimed to identify details that researchers commonly fail to provide, which provides insight to future researchers developing new ML decision support tools in the field of orthopaedic trauma surgery. Reporting of the abstract, and results such as performance measures, had the worst adherence. In order for orthopaedic trauma practice

Table II. Characteristics of included studies (n = 45).

Variables	Value
Median sample size (IQR)	875 (259 to 6,975)
Number of publications per journal, n (%)	
1	21 (70.0)
2	6 (20.0)
3	1 (3.3)
4	1 (3.3)
5	1 (3.3)
Year of publication, n (%)	
2022	18 (40.0)
2021	15 (33.3)
2020	5 (11.1)
2014 to 2019	3 (6.7)
< 2014	4 (8.9)
Type of database, n (%)	
Electronic health record	17 (37.8)
National/registry*	13 (28.9)
Other	15 (33.3)
Type of paper, n (%)	
Development	40 (88.9)
External validation	1 (3.3)
Development and external validation	4 (8.9)
Injury type, n (%)	
Hip fracture	27 (60.0)
ACL rupture	5 (11.1)
Tibia fracture	3 (6.7)
Pelvic fracture	2 (4.4)
Other	8 (17.8)
Predicted outcome, n (%)	
Mortality	15 (33.3)
Length of hospital stay	4 (8.9)
Need for amputation	2 (4.4)
Delirium	3 (6.7)
Osteonecrosis	2 (4.4)
Other†	22 (48.9)
Type of algorithm used, n (%)	
SVM	17 (12.5)
RF	16 (11.8)
ANN	14 (10.3)
XGB	11 (8.1)
PLR	6 (4.4)
Other‡	72 (52.9)
Digital application made available, n (%)	14 (31.1)

*Includes databases such as American College of Surgeons National Surgical Quality Improvement Program ACS NSQIP) and other specific registries such as National Trauma Registries.
†Other outcomes include (for each one): need for amputation, sarcopenia, posterior malleolar fracture, bladder rupture, need for blood transfusion, risk for infection, unplannend subsequent surgery, postoperative living setting, rehabilitation outcome metrics such as relative gain in function, prolonged opioid use, 'true' scaphoid fracture, adverse events on radiological follow-up, cerebral infarction, early acute kidney injury, Medicare inpatient payments, overnight hospital stay, and refracture.
‡Other algorithms included: logistic regression classifiers, convolutional neural networks, logistic regression lasso, naive Bayers classifiers, boosted decision tree, Bayes point machine, nearest neighbour classifiers, ensemble classifiers, AdaBoost, CatBoost, ExtraTrees, voting ensemble, multilayer perception, principal component regression, and linear discriminant classifier.
ACL, anterior cruciate ligament; ANN, artificial neural network; IQR, interquartile range; PLR, penalized logistic regression; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting.

to benefit from our rapid improvement in understanding of ML applications, future studies should adhere to recognized guidelines such as the TRIPOD when developing and reporting on ML-based decision support tools. This ensures development of reliable decision support tools that can guide medical decision-making for both patients and clinicians.

Limitations

This review has several limitations. First, some studies meeting the selection criteria may have been missed. However, considering the number of the screened and included studies, adding potentially missed studies would most likely not change our main message. Second, the TRIPOD guidelines were employed as a benchmark. The relative importance of each item and what composes an acceptable score can be debated. Nonetheless, the TRIPOD statement is now a widely recognized methodological standard for reporting on prediction models, and many journals now demand using this standard. Third, when scoring the completeness, strict adherence on all elements of each item was implemented. For example, item 2 "abstract" consists of 12 elements, which all have to be fulfilled in order for item 2 to be marked as "completely reported". Authors as well as reviewers might have good reasons to exclude certain elements. For example, authors may not report the "potential clinical use of the model" (item 20) if they believe their prediction model is not (yet) fit for clinical use. Moreover, some authors may be restricted by the maximum word count per section and have therefore not reported certain elements. This limits the authors in reporting according to the TRIPOD statement.

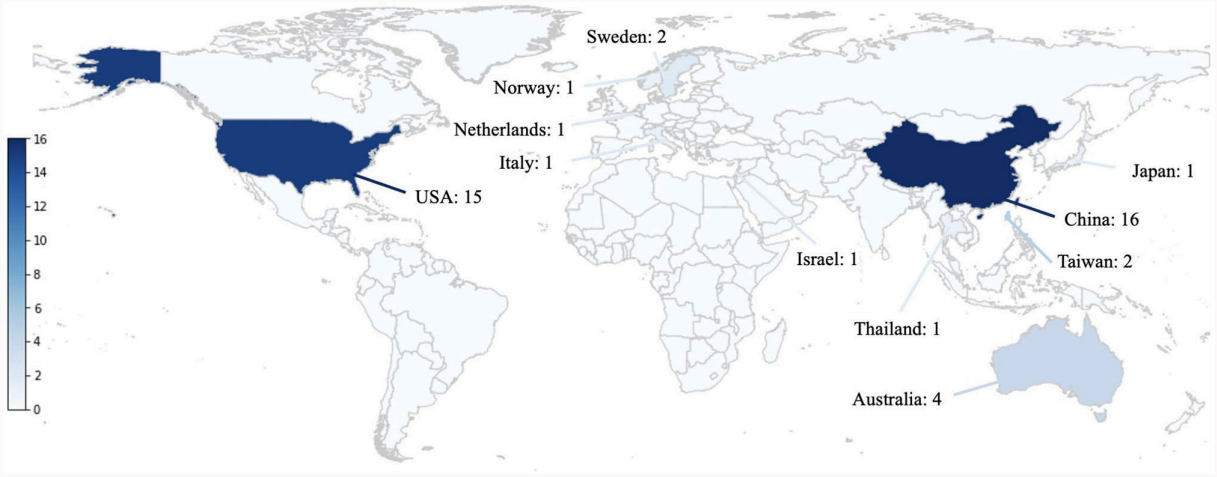

Fig. 2
World map showing the distribution of the included studies consisting of 44 development studies and one external validation study (from the USA). Of the developmental studies, four also performed external validation (three from the USA and one from Israel).

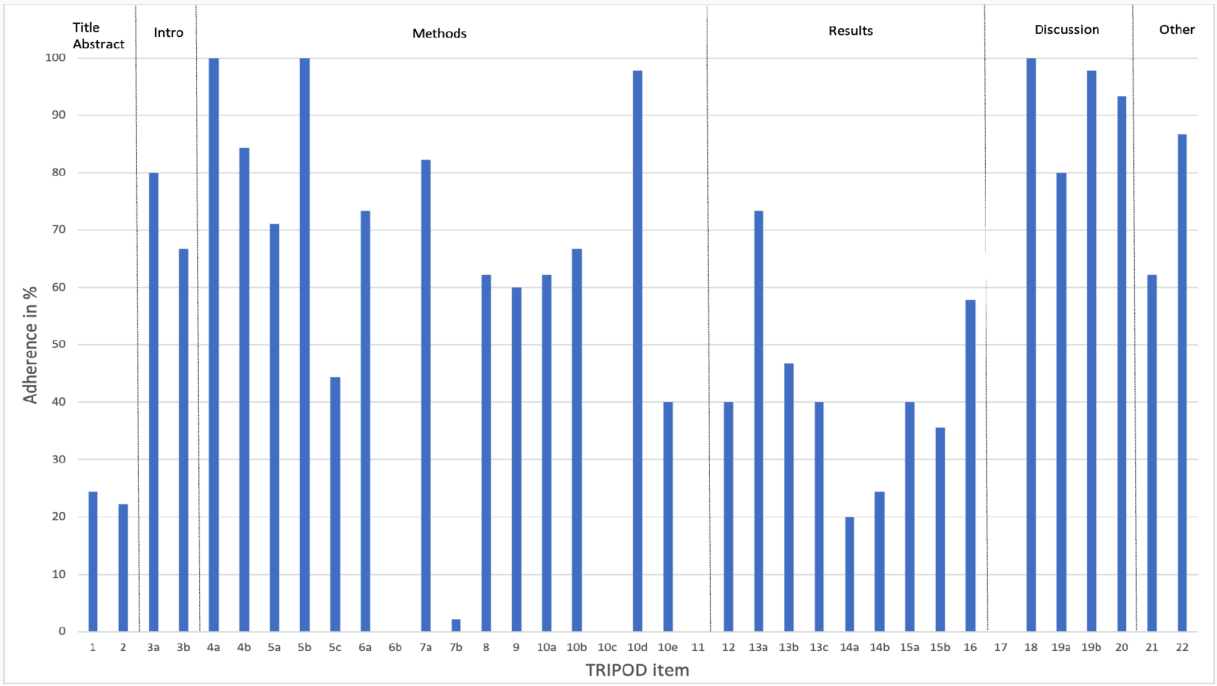

Fig. 3
Overall adherence for each TRIPOD item. Median completeness for the TRIPOD statement was 62% (interquartile range (IQR) 30 to 81). Method items adhered to a median completeness of 62% (IQR 40 to 82), results items to 40% (IQR 22 to 52), and discussion items to 96% (IQR 87 to 99).

Fourth, neither the TRIPOD nor PROBAST statements were designed for evaluating studies reporting on ML prediction models. Therefore, several key reporting items specific for ML studies might still lack critical appraisal. Despite these limitations, this review provides the first comprehensive overview of completeness of transparent reporting for a ML prediction model in the field of orthopaedic trauma surgery.

TRIPOD

This study shows a moderate overall median completeness of 62%, which is comparable to other systematic reviews in the same field.[9,13] The majority of the studies (91%; 41/45) were published after 2015, the publication year of the TRIPOD statement. Slow implementation of guidelines is nothing new,[9] but improvement in completeness of reporting could have been expected by now. The TRIPOD authors announced the development of a ML-specific extension of the guideline, the TRIPOD-AI checklist,[59,60] which at the time of writing is not yet published. This could possibly lead to a higher completeness, since some items of the statement do not apply to ML studies. Instead, these items apply better to studies using logistic regression techniques (i.e. presentation of regression coefficients, as stated in item 15a). In addition, improved adherence may occur if journals demand the extended TRIPOD-AI checklist for ML prediction model studies in the future. The lowest completeness in reporting was found to be

Table III. Sorted by completeness of above 90% reporting and under 25% of individual TRIPOD items.

TRIPOD item	TRIPOD % (n/total)
Complete reporting > 90%	
4a - Describe the study design or source of data (e.g. randomized trial, cohort, or registry data), separately for the validation dataset.	100 (45/45)
5b - Describe eligibility criteria for participants.	100 (45/45)
10d - Specify all measures used to assess model performance and, if relevant, to compare multiple models.	98 (44/45)
18 - Discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data).	100 (45/45)
19b - Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	98 (44/45)
20 - Discuss the potential clinical use of the model and implications for future research.	93 (42/45)
Complete reporting < 25%	
1 - Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	24 (11/45)
2 - Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	22 (10/45)
6b - Report any actions to blind assessment of the outcome to be predicted.	0 (0/45)
7b - Report any actions to blind assessment of predictors for the outcome and other predictors.	2 (1/45)
10c - For validation, describe how the predictions were calculated.	0 (0/5)*
11 - Provide details on how risk groups were created, if done.	0 (0/45)
14a - Specify the number of participants and outcome events in each analysis.	20 (9/45)
14b - If done, report the unadjusted association between each candidate predictor and outcome.	24 (11/45)
17 - If done, report the results from any model updating (i.e. model specification, model performance).	0 (0/5)*

*External validation was performed in five studies, resulting in a total of five for items 10c and 17, which pertain validation.

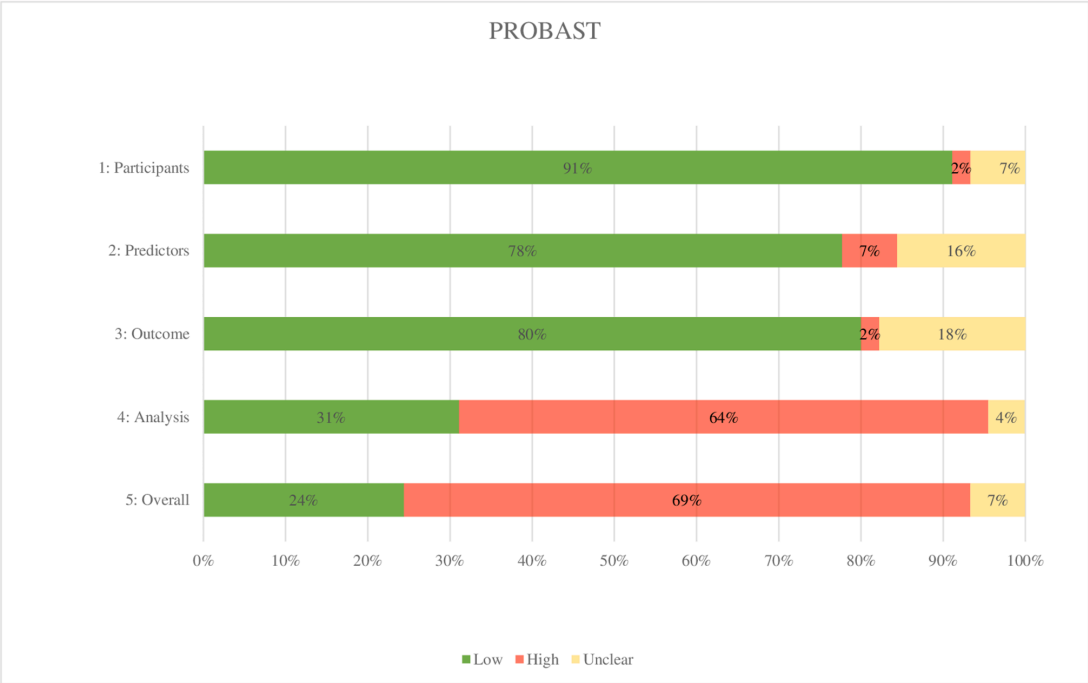

Fig. 4
Risk of bias in the PROBAST tool (n = 45).

in the results section, identifying need for better reporting of this section.

Only 11% (5/45) of all included studies reported external validation. It underlines the need for future external validation studies, since external validation is an essential step before clinical implementation of a model.[61] International collaborations and standardization of international registries may allow for universally externally valid ML decision support tools. This would be the next step for moving these tools from a single country task to a coordinated global effort.[62]

PROBAST

Most of the included studies (69%; 31/45) were at high risk of bias for several recurring reasons. First, missing data were often not handled appropriately, leading to biased predictor-outcome associations and biased model performance, as the sample is not a representative group of the study population.[12,63,64] Many studies chose to remove patients with missing data from the analysis (complete case analysis). A more appropriate method for handling missing data is multiple imputation.[12] In addition, variables with excessive missing data (often described as > 20%) are excluded in most studies. This exclusion may result in selection bias, as variables with a strong predictive accuracy may be missed. Therefore, the use of prospective, complete datasets is strongly recommended.[65]

Second, complexities in data (e.g. censoring and competing risks) are not accounted for appropriately. Time-to-event analysis (such as Cox regression or Fine-and-Gray analyses) should be used for predicting long-term outcomes, to account for loss to follow-up due to mortality (competing risk). An example for this is the study in which osteonecrosis of the femoral head is predicted in patients after a hip fracture, where death in the elderly may occur before the diagnosis osteonecrosis is established.[51] Absolute risks predictions will be overestimated and biased because patients with the competing event are then censored.[66] Current studies are evaluating the use of ML compared to traditional time-to-event analyses in orthopaedic populations, but the benefit remains limited.[67,68] To date, no studies have evaluated the benefit specific for orthopaedic trauma cohorts.[67] A separate competing risk analysis could be performed to show differences in hazard ratios for both outcomes (event and competing event).[7]

Third, performance measures were often incompletely reported. Most studies described the area under the ROC (a discrimination measure), without reporting calibration measures. Both model calibration and discrimination metrics should be assessed. If not, the study is at risk of bias because the accurate individual probabilities are not completely known.[12] Researchers should therefore at least report calibration and discrimination metrics.

Recommendations for future research

Our findings lead to several careful recommendations for researchers developing and validating ML prediction models in the field of orthopaedic trauma. First, we recommend adhering to the TRIPOD and PROBAST guidelines. Currently, specific AI extensions for each guideline are being developed because there are clear differences between ML models and traditional prediction models in model development, validation, and updating.[60] Until that time, we recommend using the guidelines as they are, without the AI extension.

Moreover, future studies should provide an online probability calculator or prediction tool: only 29% (13/45) studies provided an online, open-access digital application. The primary goal of ML prediction models is adoption in clinical practice. Before clinical implementation, external validation and prospective testing are required to test the accuracy and generalizability to other study populations. Providing access to a developed model facilitates sharing and collaborating, hopefully resulting in increased external validation. Innovations in ML, such as federated learning (a way to develop or validate an AI model without sharing data) might also contribute to this, as exchanging data remains difficult. The World Health Organization has recently developed a guideline to improve sharing possibilities.[69]

Additionally, more (international) collaborations between clinicians and data scientists are needed. Surgeons are well positioned to help integrate AI into modern practice;[70] to optimize the quality of ML papers, surgeons should partner with data scientists to capture data across phases of care, and to provide clinical context.

In summary, a great many ML prediction models exist in orthopaedic trauma surgery. Unfortunately, a substantial number of these models lack transparent reporting and are at high risk of bias, and only 5/45 models are externally validated, a required step before moving to implementation. To minimize the risk of bias and improve completeness in reporting in these studies, we recommend that future studies aiming to develop ML prediction models in orthopaedic trauma adhere to recognized guidelines, and provide proper validation opportunities by providing open-access models. This will ensure reliable and accurate prediction tools that can supplement the shared decision-making process in orthopaedic trauma practice.

Supplementary material

PRISMA checklist, literature search, basic information of the included studies, completed TRIPOD checklists, and completed PROBAST checklists.

References

1. **Oosterhoff JHF**, **Doornberg JN**, **Machine Learning Consortium**. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. *EFORT Open Rev*. 2020;5(10):593–603.
2. **Ogink PT**, **Groot OQ**, **Karhade AV**, **et al**. Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review. *Acta Orthop*. 2021;92(5):526–531.
3. **Adams M**, **Chen W**, **Holcdorf D**, **McCusker MW**, **Howe PD**, **Gaillard F**. Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. *J Med Imaging Radiat Oncol*. 2019;63(1):27–32.
4. **Beyaz S**, **Açıcı K**, **Sümer E**. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. *Jt Dis Relat Surg*. 2020;31(2):175–183.
5. **Bulstra AEJ**, **Machine Learning Consortium**. A machine learning algorithm to estimate the probability of a true scaphoid fracture after wrist trauma. *J Hand Surg Am*. 2022;47(8):709–718.
6. **Oosterhoff JHF**, **Gravesteijn BY**, **Karhade AV**, **et al**. Feasibility of machine learning and logistic regression algorithms to predict outcome

in orthopaedic trauma surgery. *J Bone Joint Surg Am*. 2022;104-A(6):544–551.
7. **van de Kuit A**, **Oosterhoff JHF**, **Dijkstra H**, **et al**. Patients with femoral neck fractures are at risk for conversion to arthroplasty after internal fixation: A machine-learning algorithm. *Clin Orthop Relat Res*. 2022;480(12):2350–2360.
8. **Oosterhoff JHF**, **Karhade AV**, **Oberai T**, **Franco-Garcia E**, **Doornberg JN**, **Schwab JH**. Prediction of postoperative delirium in geriatric hip fracture patients: A clinical prediction model using machine learning algorithms. *Geriatr Orthop Surg Rehabil*. 2021;12:21514593211062277.
9. **Groot OQ**, **Ogink PT**, **Lans A**, **et al**. Machine learning prediction models in orthopedic surgery: A systematic review in transparent reporting. *J Orthop Res*. 2022;40(2):475–483.
10. **Steyerberg EW**. Applications of prediction models. In: *Clinical Prediction Models*. New York, New York, USA: Springer, 2008. https://link.springer.com/chapter/10.1007/978-0-387-77244-8_2 (date last accessed December 2023).18
11. **Collins GS**, **Reitsma JB**, **Altman DG**, **Moons KGM**. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
12. **Moons KGM**, **Wolff RF**, **Riley RD**, **et al**. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1–W33.
13. **Groot OQ**, **Bindels BJJ**, **Ogink PT**, **et al**. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop*. 2021;92(4):385–393.
14. **Lans A**, **Pierik RJB**, **Bales JR**, **et al**. Quality assessment of machine learning models for diagnostic imaging in orthopaedics: A systematic review. *Artif Intell Med*. 2022;132:102396.
15. **Moher D**, **Liberati A**, **Tetzlaff J**, **Altman DG**, **Altman D**, **Antes G**. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLOS Med*. 2009;6(7):e1000097.
16. **Steyerberg EW**, **Vergouwe Y**. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925–1931.
17. **Cox DR**. Two further applications of a model for binary regression. *Biometrika*. 1958;45(3/4):562.
18. **Anderson AB**, **Grazal CF**, **Balazs GC**, **Potter BK**, **Dickens JF**, **Forsberg JA**. Can predictive modeling tools identify patients at high risk of prolonged opioid use after ACL reconstruction? *Clin Orthop Relat Res*. 2020;478(7):0–1618.
19. **Bevevino AJ**, **Dickens JF**, **Potter BK**, **Dworak T**, **Gordon W**, **Forsberg JA**. A model to predict limb salvage in severe combat-related open calcaneus fractures. *Clin Orthop Relat Res*. 2014;472(10):3002–3009.
20. **Bolourani S**, **Thompson D**, **Siskind S**, **Kalyon BD**, **Patel VM**, **Mussa FF**. Cleaning up the MESS: Can machine learning be used to predict lower extremity amputation after trauma-associated arterial injury? *J Am Coll Surg*. 2021;232(1):102–113.
21. **Cao Y**, **Forssten MP**, **Mohammad Ismail A**, **et al**. Predictive values of preoperative characteristics for 30-day mortality in traumatic hip fracture patients. *J Pers Med*. 2021;11(5):353.
22. **Cary MP**, **Zhuang F**, **Draelos RL**, **et al**. Machine learning algorithms to predict mortality and allocate palliative care for older patients with hip fracture. *J Am Med Dir Assoc*. 2021;22(2):291–296.
23. **Chen CY**, **Chen YF**, **Chen HY**, **Hung CT**, **Shi HY**. Artificial neural network and Cox regression models for predicting mortality after hip fracture surgery: A population-based comparison. *Medicina (Kaunas)*. 2020;56(5):243.
24. **Cui S**, **Zhao L**, **Wang Y**, **et al**. Using naive Bayes classifier to predict osteonecrosis of the femoral head with cannulated screw fixation. *Injury*. 2018;49(10):1865–1870.
25. **DeBaun MR**, **Chavez G**, **Fithian A**, **et al**. Artificial neural networks predict 30-day mortality after hip fracture: Insights from machine learning. *J Am Acad Orthop Surg*. 2021;29(22):977–983.
26. **Dong S**, **Li Z**, **Tang ZR**, **Zheng Y**, **Yang H**, **Zeng Q**. Predictors of adverse events after percutaneous pedicle screws fixation in patients with single-segment thoracolumbar burst fractures. *BMC Musculoskelet Disord*. 2022;23(1):168.
27. **Forssten MP**, **Bass GA**, **Ismail AM**, **Mohseni S**, **Cao Y**. Predicting 1-year mortality after hip fracture surgery: An evaluation of multiple machine learning approaches. *J Pers Med*. 2021;11(8):727.
28. **Harris AHS**, **Trickey AW**, **Eddington HS**, **et al**. A tool to estimate risk of 30-day mortality and complications after hip fracture surgery: Accurate enough for some but not all purposes? A study from the ACS-NSQIP Database. *Clin Orthop Relat Res*. 2022;480(12):2335–2346.
29. **Hendrickx LAM**, **Sobol GL**, **Langerhuizen DWG**, **et al**. A machine learning algorithm to predict the probability of (occult) posterior malleolar fractures associated with tibial shaft fractures to guide "malleolus first" fixation. *J Orthop Trauma*. 2020;34(3):131–138.
30. **Hertz AM**, **Hertz NM**, **Johnsen NV**. Identifying bladder rupture following traumatic pelvic fracture: A machine learning approach. *Injury*. 2020;51(2):334–339.
31. **Huang CB**, **Tan K**, **Wu ZY**, **Yang L**. Application of machine learning model to predict lacunar cerebral infarction in elderly patients with femoral neck fracture before surgery. *BMC Geriatr*. 2022;22(1):912.
32. **Huang X**, **Wang Y**, **Chen B**, **et al**. Ability of a machine learning algorithm to predict the need for perioperative red blood cells transfusion in pelvic fracture patients: A multicenter cohort study in China. *Front Med (Lausanne)*. 2021;8:694733.
33. **Karnuta JM**, **Navarro SM**, **Haeberle HS**, **Billow DG**, **Krebs VE**, **Ramkumar PN**. Bundled care for hip fractures: A machine-learning approach to an untenable patient-specific payment model. *J Orthop Trauma*. 2019;33(7):324–330.
34. **Kitcharanant N**, **Chotiyarnwong P**, **Tanphiriyakun T**, **et al**. Development and internal validation of a machine-learning-developed model for predicting 1-year mortality after fragility hip fracture. *BMC Geriatr*. 2022;22(1):451.
35. **Liu J**, **Xu L**, **Zhu E**, **Han C**, **Ai Z**. Prediction of acute kidney injury in patients with femoral neck fracture utilizing machine learning. *Front Surg*. 2022;9:928750.
36. **Lei M**, **Han Z**, **Wang S**, **et al**. A machine learning-based prediction model for in-hospital mortality among critically ill patients with hip fracture: An internal and external validated study. *Injury*. 2023;54(2):636–644.
37. **Lin CC**, **Ou YK**, **Chen SH**, **Liu YC**, **Lin J**. Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture. *Injury*. 2010;41(8):869–873.
38. **Lu Y**, **Jurgensmeier K**, **Till SE**, **et al**. Early ACLR and risk and timing of secondary meniscal injury compared with delayed ACLR or nonoperative treatment: A time-to-event analysis using machine learning. *Am J Sports Med*. 2022;50(13):3544–3556.
39. **Lu Y**, **Forlenza E**, **Cohn MR**, **et al**. Machine learning can reliably identify patients at risk of overnight hospital admission following anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc*. 2021;29(9):2958–2966.
40. **Martin RK**, **Wastvedt S**, **Pareek A**, **et al**. Predicting anterior cruciate ligament reconstruction revision: A machine learning analysis utilizing the Norwegian Knee Ligament Register. *J Bone Joint Surg Am*. 2022;104-A(2):145–153.
41. **Merrill RK**, **Ferrandino RM**, **Hoffman R**, **Shaffer GW**, **Ndu A**. Machine learning accurately predicts short-term outcomes following open reduction and internal fixation of ankle fractures. *J Foot Ankle Surg*. 2019;58(3):410–416.
42. **Machine Learning Consortium, on behalf of the SPRINT and FLOW Investigators**. A machine learning algorithm to identify patients with tibial shaft fractures at risk for infection after operative treatment. *J Bone Joint Surg Am*. 2021;103-A(6):532–540.
43. **Machine Learning Consortium on behalf of the SPRINT Investigators**. A machine learning algorithm to identify patients at risk of unplanned subsequent surgery after intramedullary nailing for tibial shaft fractures. *J Orthop Trauma*. 2021;35(10):e381–e388.
44. **Oosterhoff JHF**, **Savelberg ABMC**, **Karhade AV**, **et al**. Development and internal validation of a clinical prediction model using machine learning algorithms for 90 day and 2 year mortality in femoral neck fracture patients aged 65 years or above. *Eur J Trauma Emerg Surg*. 2022;48(6):4669–4682.
45. **Oosterhoff JHF**, **Oberai T**, **Karhade AV**, **et al**. Does the SORG Orthopaedic Research Group hip fracture delirium algorithm perform well on an independent intercontinental cohort of patients with hip fractures who are 60 years or older? *Clin Orthop Relat Res*. 2022;480(11):2205–2213.
46. **Ottenbacher KJ**, **Linn RT**, **Smith PM**, **Illig SB**, **Mancuso M**, **Granger CV**. Comparison of logistic regression and neural network analysis

applied to predicting living setting after hip fracture. *Ann Epidemiol*. 2004;14(8):551–559.
47. **Ricciardi C**, **Ponsiglione AM**, **Scala A**, **et al**. Machine learning and regression analysis to model the length of hospital stay in patients with femur fracture. *Bioengineering (Basel)*. 2022;9(4):172.
48. **Shi L**, **Wang XC**, **Wang YS**. Artificial neural network models for predicting 1-year mortality in elderly patients with intertrochanteric fractures in China. *Braz J Med Biol Res*. 2013;46(11):993–999.
49. **Shimizu H**, **Enda K**, **Shimizu T**, **et al**. Machine learning algorithms: Prediction and feature selection for clinical refracture after surgically treated fragility fracture. *J Clin Med*. 2022;11(7):2021.
50. **Shtar G**, **Rokach L**, **Shapira B**, **Nissan R**, **Hershkovitz A**. Using machine learning to predict rehabilitation outcomes in postacute hip fracture patients. *Arch Phys Med Rehabil*. 2021;102(3):386–394.
51. **Wang H**, **Wu W**, **Han C**, **et al**. Prediction model of osteonecrosis of the femoral head after femoral neck fracture: Machine learning-based development and validation study. *JMIR Med Inform*. 2021;9(11):e30079.
52. **Xing F**, **Luo R**, **Liu M**, **Zhou Z**, **Xiang Z**, **Duan X**. A new random forest algorithm-based prediction model of post-operative mortality in geriatric patients with hip fractures. *Front Med (Lausanne)*. 2022;9:829977.
53. **Yang B**, **Gao L**, **Wang X**, **et al**. Application of supervised machine learning algorithms to predict the risk of hidden blood loss during the perioperative period in thoracolumbar burst fracture patients complicated with neurological compromise. *Front Public Health*. 2022;10:969919.
54. **Ye Z**, **Zhang T**, **Wu C**, **et al**. Predicting the objective and subjective clinical outcomes of anterior cruciate ligament reconstruction: A machine learning analysis of 432 patients. *Am J Sports Med*. 2022;50(14):3786–3795.
55. **Zhang Y**, **Huang L**, **Liu Y**, **Chen Q**, **Li X**, **Hu J**. Prediction of mortality at one year after surgery for pertrochanteric fracture in the elderly via a Bayesian belief network. *Injury*. 2020;51(2):407–413.
56. **Zhao H**, **You J**, **Peng Y**, **Feng Y**. Machine learning algorithm using electronic chart-derived data to predict delirium after elderly hip fracture surgeries: A retrospective case-control study. *Front Surg*. 2021;8:634629.
57. **Zheng X**, **Xiao C**, **Xie Z**, **Liu L**, **Chen Y**. Prediction models for prognosis of femoral neck-fracture patients 6 months after total hip arthroplasty. *Int J Gen Med*. 2022;15:4339–4356.
58. **Zhong H**, **Wang B**, **Wang D**, **et al**. The application of machine learning algorithms in predicting the length of stay following femoral neck fracture. *Int J Med Inform*. 2021;155:104572.
59. **Collins GS**, **Moons KGM**. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):1577–1579.
60. **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(7):e048008.
61. **van de Sande D**, **Van Genderen ME**, **Smit JM**, **et al**. Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health Care Inform*. 2022;29(1):e100495.
62. **Johansen A**, **Golding D**, **Brent L**, **et al**. Using national hip fracture registries and audit databases to develop an international perspective. *Injury*. 2017;48(10):2174–2179.
63. **Janssen KJM**, **Donders ART**, **Harrell FE Jr**, **et al**. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol*. 2010;63(7):721–727.
64. **Little RJA**, **Rubin DB**. *Statistical Analysis with Missing Data*. Wiley-Interscience, 2002.
65. **Paxton C**, **Niculescu-Mizil A**, **Saria S**. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA Annu Symp Proc*. 2013;2013:1109–1115.
66. **Wolbers M**, **Koller MT**, **Witteman JCM**, **Steyerberg EW**. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*. 2009;20(4):555–561.
67. **Aram P**, **Trela-Larsen L**, **Sayers A**, **et al**. Estimating an individual's probability of revision surgery after knee replacement: A comparison of modeling approaches using a national data set. *Am J Epidemiol*. 2018;187(10):2252–2262.
68. **Martin RK**, **Wastvedt S**, **Lange J**, **Pareek A**, **Wolfson J**, **Lund B**. Limited clinical utility of a machine learning revision prediction model based on a national hip arthroscopy registry. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(6):2079–2089.
69. **World Health Organization**. Sharing and reuse of health-related data for research purposes: WHO policy and implementation guidance. 2022. https://www.who.int/publications/i/item/9789240044968 (date last accessed 18 December 2023).
70. **Hashimoto DA**, **Rosman G**, **Rus D**, **Meireles OR**. Artificial intelligence in surgery: Promises and perils. *Ann Surg*. 2018;268(1):70–76.

Author information

H. Dijkstra, MD, PhD Candidate,
Department of Orthopaedic Surgery, University Medical Centre Groningen, Groningen, Netherlands;
University Center for Geriatric Medicine, University of Groningen, University Medical Center Groningen, Groningen, Netherlands.

A. van de Kuit, MD, Resident Orthopaedic Surgeon, Researcher
O. Canta, BSc, Medical Student, PhD Candidate
Department of Orthopaedic Surgery, University Medical Centre Groningen, Groningen, Netherlands.

T. de Groot, BSc, Medical Student, PhD Candidate, Department of Orthopaedic Surgery, University Medical Centre Groningen, Groningen, Netherlands; Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.

O. Q. Groot, MD, PhD, Resident in Orthopaedic Surgery, Department of Orthopaedic Surgery, University Medical Centre Utrecht, University of Utrecht, Utrecht, Netherlands.

J. H.F. Oosterhoff, MD, PhD, Assistant Professor, Department of Engineering Systems & Services, Faculty Technology Policy and Management, Delft University of Technology, Delft, Netherlands.

J. N. Doornberg, MD, PhD, Professor, Orthopaedic Trauma Surgeon, Department of Orthopaedic Surgery, University Medical Centre Groningen, Groningen, Netherlands; Department of Orthopaedic Trauma Surgery, Flinders Medical Center, Flinders University, Adelaide, Australia.

Author contributions

H. Dijkstra: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – original draft.
A. van de Kuit: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – original draft.
T. de Groot: Methodology, Formal analysis, Visualization, Writing – review & editing.
O. Canta: Methodology, Formal analysis, Writing – review & editing.
O. Q. Groot: Conceptualization, Methodology, Supervision, Writing – review & editing.
J. H. F. Oosterhoff: Conceptualization, Methodology, Supervision, Writing – review & editing.
J. N. Doornberg: Conceptualization, Supervision, Writing – review & editing.

H. Dijkstra and A. van de Kuit contributed equally to this work.

Funding statement

The authors received no financial or material support for the research, authorship, and/or publication of this article.

ICMJE COI statement

Each author certifies that he or she has no commercial associations (e.g., consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

Data sharing

The datasets generated and analyzed in the current study are not publicly available due to data protection regulations. Access to data is limited to the researchers who have obtained permission for data processing. Further inquiries can be made to the corresponding author.

Acknowledgements

Machine Learnig Consortium: Michel van den Bekerom, Santiago Lozano Calderon, Joost Colaris, Kaj ten Duis, Soheil Ashkani Esfahani, Chris DiGiovanni, Max Gordon, Daniel Guss, Frank IJpma, Ruurd Jaarsma, Michiel Janssen, Prakah Jayakumar, Gino M. Kerkhoffs, Ross Leighton, Barbara van Munster, Rudolf Poolman, David Ring, Emil Schemtisch, Vincent Stirler, Paul Tornetta, Mathieu Wijffels.

Open access funding

The authors have no support or funding to report for the open access fee of the current study.