PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Developing a machine learning algorithm to predict the probability of aseptic loosening of the glenoid component after anatomic total shoulder arthroplasty: protocol for a retrospective, multicentre study
AUTHORS	Macken, Arno; Macken, Loïc; Oosterhoff, Jacobien; Boileau, Pascal; Athwal, George; Doornberg, Job; Lafosse, Laurent; Lafosse, Thibault; van den Bekerom, Michel; Buijze, Geert Alexander

VERSION 1 – REVIEW

REVIEWER	Matsunaga, Fabio Unifesp EPM, Orthopedics and Traumatology - Division of Hand Surgery and Upper Limb
REVIEW RETURNED	18-Jul-2023

GENERAL COMMENTS	1. In line 106, it is stated that the authors of primary studies will be requested to share their database. However, these participants did not consent to have their data used in a new study. This situation may not be appropriate, even with their data being anonymised. How will the authors manage this issue? 2. How will the authors provide feedback for patients who had their data included in the platform, and are identified with possible loosening of the implant?

REVIEWER	Bajpai, Ram Keele University, School of Primary, Community, and Social Care
REVIEW RETURNED	05-Aug-2023

GENERAL COMMENTS	I have reviewed this manuscript from statistical methodology point of view and identified several important points as listed below. All machine-learning (ML) methods are data hungry and should have been used when large sample size is available [https://doi.org/10.1186/1471-2288-14-137]. This simulation study suggests that ML models may need over 10 times as many events per variable to achieve a stable AUC and a small optimism than classical modelling techniques such as logistic regression (LR). Authors should make sure that they will have necessary sample size for developing a meaningful clinical prediction model. The 10 event per variable does not work with machine learning models. If authors have gone through the TRIPOD reporting guidelines for clinical prediction models, it suggests that do not randomly split data into training and test sets rather use full data for internal

	validation by using techniques such as bootstrapping or cross-validation (5-fold or 10-fold). Authors mentioned that it is not strictly necessary to externally validate the final algorithm. However, authors should know that internal validation does not provide guaranty for suitability in a different setting as it does depend on so many other factors. Therefore, it is also necessary to mention that this is a model development study as an independent data will be required for external validation. Nothing mentioned about the handling of model misclassification (over or under fitting) in the analysis plan. Most ML applications in medicine are still using classification methods rather than risk prediction. It is time for researchers to realise that classification provides limited information, is arbitrary, is more prone to patient sampling issues, and is at odds with medical decision making. I think authors should discuss as appropriate. Authors did not provide any rationale for selecting these machine learning prediction models. A proper reason should be provided for better clarity.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1		
1. In line 106, it is stated that the authors of primary studies will be requested to share their database. However, these participants did not consent to have their data used in a new study. This situation may not be appropriate, even with their data being anonymised. How will the authors manage this issue?	Thank you for this comment, it is important to carefully consider this point. The current study should not be viewed as an enrolment of patients in a new study (for which consent would be required). It is rather a meta-analysis of previously published results. No new prospective data is gathered, and researchers do not need to access any patient files. Data is also completely anonymous and not retraceable to individual patients. It can therefore be considered a compilation of previous works using a new analysis method, such as a meta-analysis or systematic review.	Line 107: anonymised Line 107: Only de-identified databases used for previous studies are included, authors are not required to gather additional data or access patient files. Line 209: , no additional prospective data is collected and contributing authors are not required to access any patient files,

	All included studies have been executed following local guidelines including IRB approval and a data-sharing agreement is signed between the parties providing their anonymised databases and our team.	
How will the authors provide feedback for patients who had their data included in the platform, and are identified with possible loosening of the implant?	It is impossible to retrace the data to an individual patient, as ethics required de-identification. Consequently, patients can not be notified of the predicted chance of implant loosening. However, once the online prediction tool is created, it will be made freely available online. We will share the tool with all authors that contributed to the study, and they will be able to use it for their patients if they wish so. In this way, patients of contributing authors will have the opportunity to make a personal prediction of implant loosening through their respective healthcare providers.	None
Reviewer 2		
All machine-learning (ML) methods are data hungry and should have been used when large sample size is available [https://doi.org/10.1186/1471-2288-14-137]. This simulation study suggests that ML models may need over 10 times as many events per variable to achieve a stable AUC and a small optimism than classical modelling techniques such as logistic regression (LR). Authors should make sure that they will have necessary sample size for	Thank you for reviewing our manuscript. Although we agree with your point that sample size is an important limitation and ML models should be used with caution in case of a limited sample size, there are two things to consider: First, the minimum events for a sample size remains a debated topic and differs per	Line 116: The minimum number of events per variable to achieve sufficient accuracy differs per model and is not clearly defined for each technique. We aim to include at least 30 events per variable, resulting in a sample size of 7500 patients for a model with up to 5 predictive variables. Line 230: However, most ML techniques require a larger

developing a meaningful clinical prediction model. The 10 event per variable does not work with machine learning models.	model and dataset. The study cited in this comment shows that ML models are more 'data hungry' but does not provide a clear minimum per model. In the book by the same author limits of 10 and 50 events per variable are recommended: "while most will agree that challenges arise especially in settings where we have less than 10 events (outcomes per patient) per predictor variable ("EPV < 10") [422]. On the other hand, having more than 50 events per variable (EPV > 50) allows for substantial freedom in modelling and will usually provide for limited model uncertainty" - *Steyerberg EW. Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer; 2019.* Second, the limitations should be considered in light of the clinical setting. Glenoid component loosening is a rare but very relevant outcome. With the methods we use in this study, we aim to collect the largest cohort published to date. Although the models may still be limited, this is currently the highest degree of accuracy we can achieve for this specific clinical outcome of interest. We also have to be realistic regarding the size of the cohort that can be collected.	sample size to achieve an accurate prediction compared to traditional regression models. The minimum events per variable is not clearly defined and differs per technique. Furthermore, (…)

	We have made a more conservative estimation of the cohort. We also acknowledge the limitations of the sample size in the methods and discussion, and added references for the sample size estimation.	
If authors have gone through the TRIPOD reporting guidelines for clinical prediction models, it suggests that do not randomly split data into training and test sets rather use full data for internal validation by using techniques such as bootstrapping or cross-validation (5-fold or 10-fold).	Thank you for this comment. It is not completely random; the division is stratified by data source (splitting each dataset separately) and outcome. We've added cross-validation to the methods. We've also copied the reference to the paper that we based our methods on.	Line 128: Each data set will be split into training (80%) and test (20%) subsets, stratified by outcome. Fivefold cross-validation of the training set will be used to develop the ML models.
Authors mentioned that it is not strictly necessary to externally validate the final algorithm. However, authors should know that internal validation does not provide guaranty for suitability in a different setting as it does depend on so many other factors. Therefore, it is also necessary to mention that this is a model development study as an independent data will be required for external validation.	The model is not trained on a single data source. The database will be compiled from multiple independent data sources from different settings and countries. This is what we considered when estimating the need for external validation. Adding one more data source for 'external' validation will not change this. Internal and external are relative terms in this case. However, we agree that external validation is important before applying the algorithm to a specific setting. Therefore, we've added this to the text.	Line 187: However, this study's primary aim is model development. External validation in a specific setting is advised before applying the algorithm to clinical practice.
Nothing mentioned about the handling of model	Thank you for this comment. Misclassification is included in the performance analysis (discrimination score,	Line 161: , positive predictive value (PPV), true positive rate (TPR), precision-recall curve,

misclassification (over or under fitting) in the analysis plan.	sensitivity, specificity), the misclassification rate could be calculated separately but we do not believe it is of added value to the other performance metrics. We believe it more relevant to add the PPV, TPR, and precision recall to the methods. If you would like to add the misclassification rate to the method section we would be happy to include it, do not hesitate to ask. It is not completely clear from your comment how this relates to under- or overfitting of the model. This is also included in the performance metrics (for example the Brier's score). For regression models the R^2 can be used to estimate over/underfitting. How would you like us to adjust the methods to make this point clear?	Line 167: The PPV is the proportion of true positive outcomes over the number of predicted positive outcomes. The TPR is the proportion of true positive outcomes over the number of observed positive outcomes. The precision recall curve is a plot of the PPV versus the TPR.
Most ML applications in medicine are still using classification methods rather than risk prediction. It is time for researchers to realise that classification provides limited information, is arbitrary, is more prone to patient sampling issues, and is at odds with medical decision making. I think authors should discuss as appropriate.	Thank you, we have highlighted this point in the discussion.	Line 241: Last, ML prediction models for a dichotomous outcome are limited to risk classification, the individual risk must be interpreted in the clinical context when used for medical decision making.
Authors did not provide any rationale for selecting these machine learning prediction models. A proper reason should be provided for better clarity.	Several different models were selected to account for the large variation in type of variables in the dataset. It is difficult to predict which type of model will perform best.	Line 144: Different ML models result in varying performance metrics based on the type of input data (continuous, categorical, dichotomous). Due to the variation in type of input variables in the dataset, several different ML techniques will be used and

	The techniques were based on previous studies with a similar methodology. A detailed argumentation for each technique is in our opinion beyond the scope of this protocol and would result in the manuscript exceeding the word limit. We hope this concern is sufficiently addressed, please let us know if you would like us to add more detail.	compared based on model performance.

VERSION 2 – REVIEW

REVIEWER	Matsunaga, Fabio Unifesp EPM, Orthopedics and Traumatology - Division of Hand Surgery and Upper Limb
REVIEW RETURNED	19-Sep-2023

GENERAL COMMENTS	Ok

REVIEWER	Bajpai, Ram Keele University, School of Primary, Community, and Social Care
REVIEW RETURNED	15-Sep-2023

GENERAL COMMENTS	Appreciate that authors have modified their manuscript. However, following minor points will help further to improve the plan for developing a prediction model. Under the variable selection it is mentioned that at least 10 events for each predictor variable will be included in the model, adhering to the rule of thumb in predictive models of binary variables. However, literature suggests that there is no rationale for 1 variable per 10 events criterion for binary logistic regression analysis (see https://doi.org/10.1186/s12874-016-0267-3). Rather, authors should use Riley et al (2019) approach (https://doi.org/10.1002/sim.7992) to identify minimum required events (of course then increase it to multiple folds for machine learning models). Authors mentioned that the best performing prediction algorithm that used to create a clinical prediction tool will be available in the form of a publicly available web application accessible on desktops, tablets, and smartphones. However, it does not suggest whether the analysis codes will also be available from reproducibility point of view if anyone want to externally validate the algorithm in a different setting. Providing a calculator is a different aspect from user point of view.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2 Dr. Ram Bajpai, Keele University Comments to the Author: Appreciate that authors have modified their manuscript. However, following minor points will help further to improve the plan for developing a prediction model.	Thank you for your review of our manuscript.	None.
Under the variable selection it is mentioned that at least 10 events for each predictor variable will be included in the model, adhering to the rule of thumb in predictive models of binary variables. However, literature suggests that there is no rationale for 1 variable per 10 events criterion for binary logistic regression analysis (see https://doi.org/10.1186/s12874-016-0267-3). Rather, authors should use Riley et al (2019) approach (https://doi.org/10.1002/sim.7992) to identify minimum required events (of course then increase it to multiple folds for machine learning models).	Thank you for this suggestion. We agree that using Riley's approach would be more thorough. The first step of this approach requires selection of potential predictors. This is something that we cannot reliably do since we are dependent on the databases that we receive and the completeness and accuracy of the variables in those databases. We do not know exactly which variables will be available to us. Furthermore, step 2 requires an R^2 estimation based on similar previous studies, which are not available in this case. Therefore, we believe that this approach would be based on too many uncertain variables to provide an accurate prediction. Please let us know if you think otherwise, we would be happy to revise again.	None.
Authors mentioned that the best performing prediction algorithm that used to create a clinical prediction tool will be available in the form of a publicly available web application accessible on desktops, tablets, and	This is a very good point, we have added this in the text.	Line 246: To facilitate reproduction of the results and external validation of the algorithm, the (anonymous) code of the developed predictive algorithms will be

smartphones. However, it does not suggest whether the analysis codes will also be available from reproducibility point of view if anyone want to externally validate the algorithm in a different setting. Providing a calculator is a different aspect from user point of view.		made available upon request with the authors.
Reviewer: 1 Dr. Fabio Matsunaga, Unifesp EPM Comments to the Author: Ok.	Thank you for your review.	