PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Development of prediction models for complications after primary total hip and knee arthroplasty: a single-centre retrospective cohort study in the Netherlands.
AUTHORS	Sweerts, Lieke; Hoogeboom, Thomas J; van Wessel, Thierry; Van der Wees, Philip; van de Groes, Sebastiaan A.W.

VERSION 1 – REVIEW

REVIEWER	Mikko Venäläinen University of Turku
REVIEW RETURNED	10-Apr-2022

GENERAL COMMENTS	The manuscript by Sweerts et al. introduces risk prediction models for multiple different complications following total hip and knee arthroplasty using electronic health record data from a single medical center. Overall, the topic is of great relevance as the risk prediction applications in the field of orthopedics have been emerging with the premise of enabling and enhancing shared decision making. Although the models lack external validation, considerate effort with transparent reporting has been put into model development and demonstrating model performance thus giving confidence in the reported results. Not all developed models are successful, but the inclusion of these models is still interesting and provides valuable information for future development of risk prediction models for those outcomes. I have only few minor suggestions for this manuscript. Title – Consider adding the word primary as the revision surgeries were excluded from the study. Also, follow the TRIPOD guidelines regarding the information that should be provided there. Abstract (and anywhere where applicable) – "Area under the curve" is general terminology for calculating the area under any curve. Consider formatting it to "Area under the receiver operating characteristic curve". You may still use the same acronym though. Discussion – I would have expected more discussion on the poorly performing models, why weren't they as successful as the other models? Do you think there might still exist some variables that were not utilized here but could modify the risk of luxation, for example? I am specifically referring to the part where it is stated that "we expect that inclusion of more predictors would not lead to a considerably different model". Figure 1 – Please consider making the background of the figure transparent (or white) instead of black.

	Although the English language is good for the most part, please re-evaluate the text thoroughly for minor grammatical errors, especially for the use of hyphen.

REVIEWER	Alex Sox-Harris Stanford University
REVIEW RETURNED	26-Apr-2022

GENERAL COMMENTS	Infusing shared decision-making processes with patient-specific estimates of risks and benefits is a high priority. This is a strong rationale for the study. I have several concerns and comments for the authors to consider. 1. In the introduction, the authors say that “none of the currently available prediction models predict the risk for surgical complications, such as surgical site infections. This is remarkable, as discussing potential risks is an important aspect of SDM.” It would be remarkable if true. I encourage the authors to review these citations and web-accessible risk calculators that can be used for TKA and/or THA. Bozic KJ, Ong K, Lau E, et al. Estimating risk in Medicare patients with THA: an electronic risk calculator for periprosthetic joint infection and mortality. Clin Orthop Relat Res. 2013;471(2):574-583. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg. 2013;217(5):833-842 e831-833. https://riskcalculator.facs.org/RiskCalculator/ Harris AH, Kuo AC, Bowe T, Gupta S, Nordin D, Giori NJ. Prediction Models for 30-Day Mortality and Complications After Total Knee and Hip Arthroplasties for Veteran Health Administration Patients With Osteoarthritis. J Arthroplasty. 2018;33(5):1539-1545. Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori, NG (2019). Can Machine Learning Methods Produce Accurate and Easy to Use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty. Clinical Orthopaedics & Related Research, 477(2), 452-460. https://med.stanford.edu/s-spire/Resources/clinical-tools-.html These are not related to complications but also relevant Kuo A, Giori N, Bowe T, et al. Comparing Methods to Determine the Minimal Clinically Important Differences (MCIDs) for Elective Total Hip and Knee Arthroplasty in Veterans Health Administration Patients. JAMA Surgery. Fontana MA, Lyman S, Sarker GK, Padgett DE, MacLean CH. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? Clin Orthop Relat Res 2019;477:1267e79. 2. The internal validation method is unclear. I understand that they used bootstrapping to generate and shrink the coefficients, but I’m unsure if any k-fold cross-validation was used so the AUCs produced on test rather than training sets. 3. It looks like procedure type (TKA vs THA) wasn’t included as a potential predictor in the models. Is the risk of these outcomes very

	similar between procedures? If not then a main effect for procedure type might help with discrimination or calibration.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Mikko Venäläinen, University of Turku
Comments to the Author:
The manuscript by Sweerts et al. introduces risk prediction models for multiple different complications following total hip and knee arthroplasty using electronic health record data from a single medical center. Overall, the topic is of great relevance as the risk prediction applications in the field of orthopedics have been emerging with the premise of enabling and enhancing shared decision making. Although the models lack external validation, considerate effort with transparent reporting has been put into model development and demonstrating model performance thus giving confidence in the reported results. Not all developed models are successful, but the inclusion of these models is still interesting and provides valuable information for future development of risk prediction models for those outcomes. I have only few minor suggestions for this manuscript.

Comment 1: Title – Consider adding the word primary as the revision surgeries were excluded from the study. Also, follow the TRIPOD guidelines regarding the information that should be provided there.
Response: The reviewer raised an important point and we have incorporated this suggestion in our new Title. Together with the first suggestion of the editor, we revised the title as follows:

'Development of prediction models for complications after primary total hip and knee arthroplasty: a single-centre retrospective cohort study in the Netherlands.' Title page, page 1

Comment 2: Abstract (and anywhere where applicable) – "Area under the curve" is general terminology for calculating the area under any curve. Consider formatting it to "Area under the receiver operating characteristic curve". You may still use the same acronym though.
Response: We have incorporated this suggestion by formatting the first time we mention 'Area under the curve' in the abstract and the manuscript to 'Area under the receiver operating characteristic curve (AUC)'. Afterwards, we used the acronym AUC.

Comment 3: Discussion – I would have expected more discussion on the poorly performing models, why weren't they as successful as the other models? Do you think there might still exist some variables that were not utilized here but could modify the risk of luxation, for example? I am specifically referring to the part where it is stated that "we expect that inclusion of more predictors would not lead to a considerably different model".
Response: We think the reviewer raised an important point. We discussed all models and we stated that the model for luxation is of insufficient quality, but we should have discussed other factors of influence on this quality. In the discussion, we elaborated on possible other (non-modifiable patient) factors who could cause luxation. Furthermore, we slightly adjusted the discussion of the other models.

'We could not demonstrate diabetes to be a significant predictor for VTE.[3] For the risk of luxation, it is known that causes of dislocation are multifactorial and also caused by non-patient modifiable factors such as implant-related, surgery-related, and hospital-related factors. It is unclear to what extent these factors contribute to the occurrence of luxation, but we expect these factors to be of influence the model.[34 36] For these reasons, and the poor performance of the model for luxation, we consider this model of insufficient quality for use in patient information documents.' Discussion, page 19

'Furthermore, due to a low estimated event rate (1-3%) we needed a large population to have enough events to include predictors into our models. However, since not all

predictors were significant in our final models, we expect that inclusion of more predictors would not lead to a considerably different model, as also discussed above.' Discussion, page 20

Comment 4: Figure 1 – Please consider making the background of the figure transparent (or white) instead of black.

Response: We thank the reviewer for his attention. The background was transparent but became black while uploading the Figure in ScholarOne. We formatted the Figure to another file type and expect the layout will be as we proposed it to be.

Comment 5: Although the English language is good for the most part, please re-evaluate the text thoroughly for minor grammatical errors, especially for the use of hyphen.

Response: We double checked the text for grammatical errors and the use of hyphen. We hope this meets the reviewers point.

Reviewer: 2

Alex Sox-Harris
Comments to the Author:
Infusing shared decision-making processes with patient-specific estimates of risks and benefits is a high priority. This is a strong rationale for the study. I have several concerns and comments for the authors to consider.

Comment 1: In the introduction, the authors say that “none of the currently available prediction models predict the risk for surgical complications, such as surgical site infections. This is remarkable, as discussing potential risks is an important aspect of SDM.” It would be remarkable if true. I encourage the authors to review these citations and web-accessible risk calculators that can be used for TKA and/or THA.

Bozic KJ, Ong K, Lau E, et al. Estimating risk in Medicare patients with THA: an electronic risk calculator for periprosthetic joint infection and mortality. Clin Orthop Relat Res. 2013;471(2):574-583.
Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg. 2013;217(5):833-842 e831-833.
https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Friskcalculator.facs.org%2FRiskCalculator%2F&amp;data=05%7C01%7Clieke.sweerts%40radboudumc.nl%7C94e7bb81ca9a4ab32fca08da312e33c7%7Cb208fe69471e48c48d87025e9b9a157f%7C1%7C0%7C637876371748171892%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&amp;sdata=l1Qaf25HGt3TBC46lEUg7qfzHLW2G%2BonPA9V9H30Bbs%3D&amp;reserved=0
Harris AH, Kuo AC, Bowe T, Gupta S, Nordin D, Giori NJ. Prediction Models for 30-Day Mortality and Complications After Total Knee and Hip Arthroplasties for Veteran Health Administration Patients With Osteoarthritis. J Arthroplasty. 2018;33(5):1539-1545.
Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori, NG (2019). Can Machine Learning Methods Produce Accurate and Easy to Use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty. Clinical Orthopaedics & Related Research, 477(2), 452-460.
https://eur02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fmed.stanford.edu%2Fs-spire%2FResources%2Fclinical-tools-.html&amp;data=05%7C01%7Clieke.sweerts%40radboudumc.nl%7C94e7bb81ca9a4ab32fca08da312e33c7%7Cb208fe69471e48c48d87025e9b9a157f%7C1%7C0%7C637876371748171892%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&amp;sdata=HrllMwUOai%2FPN63QPbm7ny3BsrqiWNlfJv0qTgIsHYk%3D&amp;reserved=0
These are not related to complications but also relevant Kuo A, Giori N, Bowe T, et al. Comparing Methods to Determine the Minimal Clinically Important Differences (MCIDs) for Elective Total Hip and Knee Arthroplasty in Veterans Health Administration Patients. JAMA Surgery.
Fontana MA, Lyman S, Sarker GK, Padgett DE, MacLean CH. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? Clin Orthop Relat Res 2019;477:1267e79.

Response: We highly value the reviewer’s suggestions. We thoroughly evaluated the citations and we think the reviewer raises fair points regarding discussing potential risks for THA and TKA. The calculators and models are important to discuss in our manuscript and therefore we have incorporated these in the introduction and discussion. Furthermore we added the Minimal Clinically Important Difference to our suggestions for evaluation of the clinical impact because we agree that this is important point to consider.

‘Also useful electronic risk calculators predicting complications and mortality for patients and clinicians are available for specific populations.[13 14] *In one study, data of patients registered in the Medicare database, the federal health insurance program for*

individuals aged ≥65 years, are used for development of a risk calculator. However, the exact patient characteristics of the study population are not reported and the effect of the predictors remain unclear.[14] *In another study, regression models are based on the results of univariate analyses on a broad range of data as demographics, comorbidities, and laboratory, or test values of a mainly male Veteran population, and the authors reported suboptimal performance scores for prediction of most outomes.*[13] *Applicability of prediction models based on specific patient populations may be limited, and further evaluation of potential risk factors is needed to validate prediction models for complications after primary total hip- and knee replacement.*
As it is known from literature that personal factors including demographic characteristics and comorbidities have an impact on surgical complications,[3] *these assumed caused relationships might therefore serve as basis for a risk prediction model.' Introduction, page 5-6*

Additionally to our comparison to literature we added the reflection with the calculators:

'For the already available prediction model based on data of Veterans with osteoarthritis of Harris et al., independent variables of the model cannot be compared for SSI since these results have not been reported.[13] *We found a slightly better c-statistic (AUC) of 0.72 compared to 0.66 in their boosted model. Also comparison with Bozic et al., is difficult since applicability to non-Medicare population is questionable, as they also describe in their discussion.*[14] *Discussion, page 18*

We added the importance of determining the Minimal Clinically Important Difference to our Strengths and limitations section in the discussion. We added this to the part where we discuss the clinical impact of the model:

'For clinical impact it is also important to determine the Minimal Clinically Important Difference of the outcomes'. Discussion, page 20

Comment 2: The internal validation method is unclear. I understand that they used bootstrapping to generate and shrink the coefficients, but I'm unsure if any k-fold cross-validation was used so the AUCs produced on test rather than training sets.

Response: As the reviewer states, we indeed used bootstrapping to internally validate the model. This was performed to estimate the performance in future patients, and to adjust the model by shrinkage factors. This was performed to ensure future predictions to be less extreme. Bootstrap samples were drawn with replacement. Due to the drawing with replacement, a bootstrapped dataset may contain multiple instances of the same original cases, and may completely omit other original cases. K-fold cross-validation resamples without replacement and thus produces data sets that are smaller than the original. These data sets are produced in a systematic way so that after a pre-specified number k of surrogate data sets, each of the n original cases has been left out exactly once. Since we do not have a very large dataset, we wanted to prevent to create a relatively small validation data set. Therefore we did not use k-fold cross-validation. In addition, since we needed a quite large dataset for the development of the models, we were not able to split up the dataset for cross-validation. We addressed this point in the discussion of the manuscript because we think that this important to discuss. Furthermore, we clarified the methods section about internal validation to be more precise in our used methods.

'Due to the drawing with replacement, a bootstrapped dataset allows for containing the same original cases. Other validation methods resample without replacement and thereby such validation datasets are produced through a pre-specified number of surrogate datasets, and each of the original cases will be left out exactly once, which results in a smaller dataset. Since our dataset is not very large, we decided to use bootstrapping as internal validation method. Bootstrapping was performed to estimate the performance in future patients, and to adjust the model by the calculated shrinkage factor so that future predictions will be less extreme.[19]*' Methods, page 9*

'Another limitation is that we only performed internal validation by bootstrapping, and were not yet able to determine external validity and clinical impact of the models.' Discussion, page 20

We would like to mention that we are currently working on an external validation of the prediction models.

Comment 3: It looks like procedure type (TKA vs THA) wasn't included as a potential predictor in the models. Is the risk of these outcomes very similar between procedures? If not then a main effect for procedure type might help with discrimination or calibration.

Response: The risks for the outcomes are expected to be quite similar, except for the risk for luxation and nerve damage, which are uncommon in TKA. We decided to develop the models (except the model for luxation and nerve damage) over the pooled data because we assumed that the influence of patient characteristics, comorbidities and medication use is expected to be comparable for both surgical procedures, THA and TKA. It is known that for instance the risk on surgical site infection is somewhat higher for TKA compared to THA, but we consider this negligible for the influence of patient characteristics, comorbidities and medication use on the risk for complications. To be sure, we additionally conducted our analysis on separate THA and TKA data. The models with corresponding performance measures were still consistent with the main analysis.

We realized that we have not described our assumption that outcomes are expected to be similar between THA and TKA. We added this notification, and the notification about the additional analysis, to our discussion section.

'The models were developed based on pooled THA and TKA data. It is expected that the influence of patient characteristics, comorbidities and medication use is comparable for both THA and TKA.[41] The influence of comorbidities on outcomes is studied together quite often.[3] Furthermore, we tested this assumption by performing the analysis on THA and TKA data only. The models with corresponding performance measures were still consistent with the main analysis.' Discussion, page 20

VERSION 2 – REVIEW

REVIEWER	Mikko Venäläinen University of Turku
REVIEW RETURNED	14-Jun-2022

GENERAL COMMENTS	I would like to thank the authors for addressing all my comments. I have no further remarks to make.

REVIEWER	Alex Sox-Harris Stanford University
REVIEW RETURNED	31-May-2022

GENERAL COMMENTS	The authors have responded reasonably well to the previous suggestions and comments. I appreciate the updated literature review but the authors still missed the paper and calculator that is most similar to this work and population. Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori, NG (2019). Can Machine Learning Methods Produce Accurate and Easy to Use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty. Clinical Orthopaedics & Related Research, 477(2), 452-460. They might also briefly discuss the tension between risk calculators that are usable for a variety of procedures, like the ASC calculator, and more procedure-specific calculators like the one they developed.

	Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori, NG (2019). Can Machine Learning Methods Produce Accurate and Easy to Use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty. Clinical Orthopaedics & Related Research, 477(2), 452-460. The authors should be commended for contributing to this important area.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Mikko Venäläinen, University of Turku

Comments to the Author:
I would like to thank the authors for addressing all my comments. I have no further remarks to make.

Reviewer: 2
Alex Sox-Harris

Comments to the Author:
The authors have responded reasonably well to the previous suggestions and comments. I appreciate the updated literature review but the authors still missed the paper and calculator that is most similar to this work and population.
Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NG (2019). Can Machine Learning Methods Produce Accurate and Easy to Use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty. Clinical Orthopaedics & Related Research, 477(2), 452-460.

They might also briefly discuss the tension between risk calculators that are usable for a variety of procedures, like the ASC calculator, and more procedure-specific calculators like the one they developed.

Harris AHS, Kuo AC, Weng Y, Trickey AW, Bowe T, Giori NG (2019). Can Machine Learning Methods Produce Accurate and Easy to Use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty. Clinical Orthopaedics & Related Research, 477(2), 452-460.

The authors should be commended for contributing to this important area.

Response to comments from Reviewer 2:
We agree with the reviewer that this citation is valuable to discuss and refer to in our manuscript. We unintentionally missed your work before, we apologize as this work is very interesting. We have now added the citation to and some elaboration on your work to our introduction and discussion section:

> "*Harris et al. developed prediction models with machine learning techniques models to determine demographic and clinical predictors for prediction of postoperative complications and mortality. The authors were able to identify predictor variables for their three most accurate models predicting a postoperative renal complication, cardiac complication, and death. Further research is warranted to identify relevant predictors for different postoperative outcomes.[17]*" -- Introduction
>
> "*Similar variables as those used in our models, were used for the development of other models predicting postoperative complications as well, such as the models of Harris et al. Unfortunately, a direct comparison of the predictive capacity of these variables between the models of Harris et al. and our models is not possible, as the postoperative outcomes used in their prediction models were different to the postoperative outcomes used in our models.[17]*" – Discussion

We also agree with the reviewer that discussing the tension between universal surgical risk prediction models versus procedure specific prediction models is interesting. We assume that the reviewer refers to the study performed by Trickey et al., (full citation see reference 10 below) instead of the citation as presented above (which is the same as the first citation). To address this topic, we have added a brief differentiation between universal surgical prediction models and procedure specific prediction models to our introduction section. Furthermore, we added the citation of Trickey et al., and some elaboration on this work to our introduction section:

> "*To overcome this problem, models that can predict postoperative complications are frequently developed and applied. Several universal surgical prediction models have already been developed based on a big national database.[9] However, before applying these models to orthopedic surgical procedures, performance and accuracy on the specific surgical field needs to be determined. For total joint arthroplasty, this is performed by Trickey et al.[10] As shown by Trickey et al., and others, patients at risk of not benefitting from total hip- or total knee arthroplasty (THA or TKA) can be identified using prediction models based on preoperative data such as demographic factors, and pain scores, and physical functioning measured with Patient Reported Outcome Measures (PROMs).[10-13]*" – Introduction

References used in this point by point response

[9] Meguid RA, Bronsert MR, Juarez-Colunga E, et al. Surgical Risk Preoperative Assessment System (SURPAS): III. Accurate Preoperative Prediction of 8 Adverse Outcomes Using 8 Predictor Variables. Annals of Surgery 2016;264(1)

[10] Trickey AW, Ding Q, Harris AHS. How Accurate Are the Surgical Risk Preoperative Assessment System (SURPAS) Universal Calculators in Total Joint Arthroplasty? Clinical orthopaedics and related research 2020;478(2):241-51.

[11] Riddle DL, Golladay GJ, Jiranek WA, et al. External Validation of a Prognostic Model for Predicting Nonresponse Following Knee Arthroplasty. J Arthroplasty 2017;32(4):1153-58.e1.

[12] Barlow T, Dunbar M, Sprowson A, et al. Development of an outcome prediction tool for patients considering a total knee replacement--the Knee Outcome Prediction Study (KOPS). BMC musculoskeletal disorders 2014;15:451.

[13] Lungu E, Desmeules F, Dionne CE, et al. Prediction of poor outcomes six months following total knee arthroplasty in patients awaiting surgery. BMC musculoskeletal disorders 2014;15:299.

[17] Harris AHS, Kuo AC, Weng Y, et al. Can Machine Learning Methods Produce Accurate and Easy-to-use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty? Clinical orthopaedics and related research 2019;477(2):452-60.