# Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine-learning based approach.

**Authors:**

Christian Michelsen, M.Sci.1, Christoffer C Jørgensen, M.D.2,3, Mathias Heltberg, M.Sci,1 Mogens H. Jensen, D.Sci., Prof.,1 Alessandra Lucchetti, M.Sci.,1 Pelle B Petersen, M.D., Ph.D,2,3 Troels Petersen, M.Sci, Ass.Prof.,1 Henrik Kehlet,M.D., Ph.D., Prof.,2,3 The Center for Fast-track Hip and Knee Replacement Collaborative group\*

1The Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17 2100 Copenhagen, Denmark

2Section of Surgical Pathophysiology 7621, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

3The Centre for Fast-track Hip and Knee Replacement, 7621, Rigshospitalet, Blegdamsvej 9, 2100 Copenhagen, Denmark

\*The members of the Center for Fast-track Hip and Knee Replacement Collaborative group are co-authors of the study and can be found under the subheading Collaborators.

This is a shared first authorship between MC and CJ.

**Corresponding author:**

Dr. Christoffer Calov Jørgensen

Section for Surgical Pathophysiology 7621

Rigshospitalet, Blegdamsvej 9,

DK-2100 Copenhagen, Denmark

Phone +45 3545 4616 Fax: +45 3545 6543

E-mail: christoffer.calov.joergensen@regionh.dk

**Word Count: 3762**

**ABSTRACT**

**Objectives**: Machine-learning models may improve prediction of length of stay (LOS) and morbidity after surgery. However, few studies include fast-track programs, and most rely on administrative coding with limited follow-up and information on perioperative care. This study investigates benefits of machine-learning models for prediction of postoperative morbidity in fast-track total hip (THA) and knee arthroplasty (TKA).

**Design:** Cohort study with prospective recording of comorbidity and prescribed medication. Information on length of stay and readmissions through the Danish National Patient Registry and medical records.

**Participants**: Consecutive unselected primary THA or TKAs between 2014-2017 from seven Danish centers with established fast-track protocols. Data was split into a training (n:18013) and a test set (n:3913).

**Outcomes:** Ability of a machine-learning model with 33 preoperative variables for predicting “medical” morbidity leading to LOS >4 days or 90-days readmissions vs. a logistic regression model. We also evaluated a machine-learning model excluding age, an age-only model and parsimonious machine-learning and logistic regression models using the ten most important variables. Model performances were analyzed using precision, area under receiver operating (AUROC) and precision recall curves (AUPRC). Variable importance was analyzed using Shapley Additive Explanations values.

**Results:** Using a threshold of 20% “risk-patients” (n:782), precision, AUROC and AUPRC were 13.6%, 76.3% and 15.5% vs. 12.5%, 74.5% and 15.7% for the machine-learning and logistic regression model, respectively. The parsimonious machine-learning model performed better than the full logistic regression model, while the remaining models performed worse. Of the top ten variables, eight were shared between the machine-learning and logistic regression models, but with a considerable age-related variation in importance of specific types of medication.

**Conclusion:** Machine-learning algorithms using preoperative characteristics and prescriptions slightly improved identification of patients in high-risk of “medical” complications after fast-track THA and TKA. Such algorithms could help identify patients who benefit from intensified perioperative care.

**STRENGTHS AND LIMITATIONS**

**Strengths**

* Fully implemented fast-track protocols with complete follow-up through nationwide registries and medical records.
* State of the machine-learning techniques
* Novel analysis on the importance of preoperative prescriptions in predicting postoperative morbidity.

**Limitations**

* Limited amount of multilevel continuous data, potentially limiting full realization of the machine-learning model.
* Registration of preoperative prescriptions dependent on reimbursement and lack of information on actual use on day of surgery

**INTRODUCTION**

Prediction of postoperative morbidity and requirement for hospitalization is important for planning of health care resources. With regard to the common surgical procedures of primary total hip (THA) and knee arthroplasty (TKA), the introduction of enhanced recovery or fast-track programs has led to a significant reduction of postoperative length of stay (length of stay) as well as morbidity and mortality.1-3 However, despite such progress, a fraction of patients still have postoperative complications leading to prolonged length of stay or readmissions.1 3 4 Consequently, in order to prioritize perioperative care, many efforts have been published to preoperatively predict length of stay and morbidity using traditional risk factors such as age, preoperative cardio-pulmonary disease, anemia, diabetes, frailty, etc.4-8 These efforts have been based on traditional statistical methods, most often multiple regression analyses, and essentially concluding that it is “better to be young and healthy than old and sick”. Consequently, despite being statistically significant, conventional risk-stratification based on such studies has had a relatively limited clinically relevant ability to predict and reduce potentially preventable morbidity and length of stay.4-8

More recently, machine-learning methods have been introduced with success in several areas of healthcare and where preliminary data suggest them to improve surgical risk prediction compared to traditional risk calculation in certain anesthetic and surgical conditions.9 10 This is also the case in THA, TKA and uni-compartmental knee replacement, where several publications on machine-learning algorithms for prediction of length of stay,11 12 complications,13 disability,14 potential outpatient setup,15 readmissions16 or payment models,17 18 have shown promising predictive value compared to conventional statistical methods.19

However, few papers have included fast-track programs, and most are based on large database cohorts with the presence of risk factors and complications often relying on administrative coding with limited information on perioperative care, follow-up and discharge destination. In our previous study of 9512 THA and TKAs within a fully implemented fast-track protocol and including the above information, we did not find advantages of machine-learning methods compared to logistic regression in predicting a length of stay > 2 days.20 However, this may have been due to data imbalance, lack of details on medication and the chosen outcome of length of stay of >2 days.20 Thus, machine-learning models remain promising and could provide an improved basis for identifying a potential “high-risk” surgical population who may benefit from more extensive preoperative evaluation and postoperative medical care.

Consequently, we used a large consecutive cohort of patients undergoing fast-track total hip and knee replacement within a national public health-care system1 to develop an improved machine-learning model for preoperative prediction of “medical” complications resulting in prolonged length of stay and readmissions. Model performances were subsequently compared to a traditional logistic regression model. In addition to well-defined patient-reported preoperative risk-factors, we also included information on dispensed reimbursed prescriptions 6 months prior to surgery using a nationwide registry.21

**METHODS**

Reporting of the study is done in accordance with the Transparent reporting of multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement22 and the Clinical AI Research (CAIR) checklist proposal.23  
The study is based on the Centre for Fast-track Hip and Knee Replacement database which is a prospective database on preoperative patient characteristics and enrolling consecutive patients from 7 departments between 2010 and 2017. The database is registered on ClinicalTrials.gov as a study registry (NCT01515670). Patients completed a preoperative questionnaire with nurse assistance if needed. Additional information on reimbursed prescriptions 6 months prior to surgery was acquired using the Danish National Database of Reimbursed Prescriptions (DNDRP) which records all dispensed prescriptions with reimbursement in Denmark.21 Finally, data were combined with the Danish National Patient Registry (DNPR) for information on length of stay (counted as postoperative nights spent in hospital), 90-days readmissions with overnight stay and mortality. In case of length of stay >4 days or readmission, patient discharge summaries were reviewed for information on postoperative morbidity and in case of insufficient information, the entire medical records were reviewed. Readmissions were only included if considered related to the surgical procedure, thus excluding planned procedures like cancer workouts, cataract surgery, etc. Readmissions due to urinary tract infection or dizziness after day 30 were also considered unrelated to the surgical procedure. In case of postoperative mortality the entire medical record, including potential readmissions, was reviewed to identify cause of death. Evaluation of discharge and medical records was performed by PP supervised by CJ. In case of disagreement, records were conferred with HK. Subsequently, causes of length of stay >4, readmissions or mortality were classified as “medical” when related to perioperative care (renal failure, falls, pain, thrombosis, anemia, venous thromboembolism or infection etc.) and “surgical” if related to surgical technique (prosthetic infection, revision surgery, periprosthetic fracture, hip dislocation, etc.).1 In case of a length of stay 4-6 days with a standard discharge summary describing a successful postoperative course, it was assumed that no clinically relevant postoperative complications had occurred. If length of stay was >6 days but with standard discharge summary, the entire medical record was evaluated to confirm that no relevant complications had occurred.  
For the present study, only cases between 2014 and 2017 were used to provide the most up-to date data. All patients had elective unilateral total hip and knee replacement in dedicated arthroplasty departments with similar fast-track protocols, including multimodal opioid sparing analgesia with high-dose (125mg) methylprednisolone, preference for spinal anesthesia, only in-hospital thromboprophylaxis when length of stay ≤5 days, early mobilization, functional discharge criteria and discharge to own home.1 There were no selection criteria for the fast-track protocol as it is considered standard of care, but we excluded patients with previous major hip or knee surgery within 90-days of THA or TKA and THA due to severe congenital joint disorder or cancer.

**Ethics and Permissions**

No Ethics Committee approval was necessary as the National Danish Ethics committee exempt non-interventional observational studies. Permission to review and store information from medical records without informed consent was acquired from Center for Regional Development (R-20073405) and the Danish Data Protection Agency (RH-2007-30-0623).

**Patient and Public Involvement**

There was no involvement of patients or the public in the planning or conduction of the study.

**Outcomes**

The primary outcome was to develop a machine-learning model to predict the occurrence of “medical” complications resulting in a length of stay >4 days or readmission and compare model performance with a traditional logistic regression model (primary outcome). Secondarily, we investigated how inclusion of cases with a length of stay >4 days but no reported “medical” complication as a positive outcome influenced the model (secondary outcome). For both outcomes, we also investigated whether a parsimonious model including only the top ten variables would perform equally well as the full model, and whether the effect of age per se would compare to the full machine-learning model. All figures and tables in the main text are based on the primary outcome; the corresponding figures for the secondary outcome are reported in the Supplemental Material.

**Statistical Analysis**

Data consisted of 33 input variables, of which 7 were continuous. All variables were collected prospectively, either through the patient completed questionnaire, through the DNDRP or a combination of both (table 1). Initially we trimmed the dataset by removing 156 patients (1.7%) who were outliers with regards to weight (<30 kg or >250 kg) and height (<100 cm or >210 cm) or where these data were missing. To reduce the risk of overfitting, data was subsequently split into a training set consisting of 18013 (82.2%) procedures from 2014-2016 and a test set of 3913 (17.8%) procedures from 2017 (Supplemental Material 1).

As reference model, we used classical logistic regression model with cases of missing values being handled by imputing missing values with the median of present values. All variables were then normalized.

In addition, we used Boosted Decision Trees (LightGBM)24 for the machine-learning models, as such methods work well with categorical data and missing values. We tried using both normal cross entropy and FocalLoss25 as the objective function for the machine-learning model. The reason for testing FocalLoss was to allow the machine-learning model to focus more on the (few) positives.

The full machine-learning model was trained and hyperparameter optimized using state of the art machine-learning methods. The models were trained on the training data and then used for making predictions on the unseen test data (see supplementary for details). The classification threshold was calibrated such that no more than 20% of the total number of patients were predicted as positive by the model (positive predictive fraction (PPF) of 20%). We also included results for PPF values of 25% and 30%. Furthermore, we trained two parsimonious models using machine-learning and logistic regression with only the 10 most important features. Finally, we specifically explored the influence of increasing age, by constructing a model based only on age (Age), and a machine-learning model based on all variables except for age.

To investigate the importance of the included variables, we computed the SHapley Additive exPlanations (SHAP) values, which provide estimates on which variables contribute most to the risk score predictions.26 27 Finally, we investigated a potential relation between reimbursed prescribed cardiac drugs, anticoagulants, psychotropics and pulmonary drugs and age, as the relation between polypharmacy and postoperative outcomes have mainly been found in older patients.28  
For evaluating model performance, we computed the number of true positives, false positives, false negatives, true negatives, sensitivity (true positive rate), precision (positive predictive value). Since the data was quite imbalanced (about a 1:20 positive:negative ratio) we also computed the Matthews Correlation Coefficient (MCC) which is independent of class imbalance.29 30 The MCC ranges between -1 (the 100% wrong classifier), 0 (the random classifier), and +1 (the perfect classifier). Finally, we computed the area under the receiver operating characteristic curve (AUROC) and the area under the precision recall curve (AUPRC). To evaluate the statistical difference between the classifiers, we applied a Bayesian metric comparison P(sensitivity),31 which is the probability that a model will perform better than the machine-learning model relative to the sensitivity. Thus, for two equally performing models P(sensitivity) is ≈ 50%.

**RESULTS**

Median age in the 3913 patients was 70 years (IQR 62-76), 59% were female and 58% had THA (table 1).

|  |  |  |
| --- | --- | --- |
| Table 1. patient demographics with and without theprimary outcome (length of stay >4 days or readmissions due to “medical” morbidity) in the combined test and training dataset. | | |
| Preoperative characteristics  n (%) unless otherwise specified | +primary outcome (n:1180) | -primary outcome (n:20837) |
| mean age (SD) | 75.0 (68.0-81.0) | 69.0 (62.0-75.0) |
| mean number of reimbursed prescriptions1 (SD) | 3.0 (1.0-4.0) | 2.0 (0.0-3.0) |
| female gender | 755 (64.0) | 12133 (58.2) |
| hip arthroplasty | 636 (53.9) | 11542 (55.4) |
| mean weight in kg (SD) | 78.0 (67.0-91.0) | 81 (70.0-93.0) |
| mean height in cm (SD) | 168 (162.0-175.0) | 170.0 (164.0-178.0) |
| mean body mass index (SD) | 27.3 (23.9-31.2) | 27.5 (24.6-31.1) |
| regular use of walking aid  missing | 552 (46.8)  29 (2.5) | 4398 (21.5)  359 (1.7) |
| living alone  with others  institution  missing | 578 (49.0)  571 (48.4)  24 (2.0)  7 (0.6) | 6717 (32.2)  13869 (66.6)  113 (0.5)  138 (0.7) |
| hemoglobin  missing | 8.2 (7.7-8.8)  11 (0.9) | 8.6 (8.1-9.2)  314 (1.5) |
| >2 units of alcohol/day  missing | 79 (6.7)  10 (0.8) | 1589 (7.6)  174 (0.8) |
| active smoker  missing | 130 (11.0)  11 (0.9) | 2751 (13.2)  141 (0.7) |
| cardiac disease  missing | 306 (25.9)  8 (0.8) | 2750 (13.2)  153 (0.7) |
| hypercholesterolemia  missing | 467 (39.6)  8 (0.7) | 6062 (29.1)  120 (0.6) |
| hypertension  missing | 738 (62.5)  64 (5.4) | 10141 (48.7)  663 (3.2) |
| pulmonary disease  missing | 182 (15.4)  5 (0.4) | 1841 (8.8)  96 (0.5) |
| previous cerebral attack  missing | 165 (14.0)  25 (2.1) | 1086 (5.2)  282 (1.4) |
| previous VTE  missing | 133 (11.3)  26 (2.2) | 1481 (7.1)  325 (1.6) |
| malignancy (undefined)  previous radically treated malignancy  missing | 557 (47.2)  127 (10.8)  14 (1.2) | 8843 (42.4)  2065 (9.9)  162 (0.8) |
| chronic kidney disease  missing | 50 (4.2)  35 (3.0) | 273 (1.3)  292 (1.4) |
| family member with VTE  missing | 155 (13.1)  1190 (16.1) | 2510 (12.0)  2569 (12.3) |
| regular snoring  uncertain about snoring  missing | 266 (22.5)  208 (17.6)  259 (21.9) | 5522 (26.5)  3781 (18.1)  3309 (15.9) |
| not feeling rested  uncertain about being rested  missing | 468 (39.7)  48 (4.1)  105 (8.9) | 9340 (44.8)  809 (3.9)  1230 (5.9) |
| psychiatric disorder  missing | 156 (13.2)  62 (5.3) | 1590 (7.6)  703 (3.4) |
| Characteristic based on combination of questionnaire and DNDRP | | |
| diabetes  diet treated diabetes2  oral antidiabetics  insulin treated diabetes3  missing | 29 (2.5)  137 (11.6)  60 (5.1)  7 (0.6) | 274 (1.3)  1448 (6.9)  413 (2.0)  98 (0.5) |
| SD: standard deviation VTE: venous thromboembolic event DNDRP: Danish National Database of Reimbursed Prescriptions.  1Antirheumatica, steroids, anticoagulants, cardiac, cholesterol lowering, respiratory and psychotropic drugs. 2Reported diabetes but no registered prescriptions 3 +/- oral antidiabetics | | |

Details on prescribed drug types are shown in Supplemental Material 2. Median length of stay was 2 (IQR: 1-2) days with 7.6% 90-days readmissions and theprimary outcome occurring in 182 (4.7%) patients. When applying any model with a positive prediction fraction of 20% to the 3913 patients, 782 qualified as “risk-patients”. The results are summarized in figure 1 and table 2.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Performance of the six different models with a predefined positive prediction fraction of 20% for primary outcome | | | | | | | | | | |
| Positive prediction fraction 20% | TP | FP | FN | TN | sensitivity | precision | MCC | AUROC | AUPRC | P (sensitivity) |
| Full machine-learning model | 106 | 676 | 76 | 3055 | 58.2% | 13.6% | 21.1% | 76.3% | 15.5% | - |
| Full logistic regression model | 98 | 684 | 84 | 3047 | 53.8% | 12.5% | 18.7% | 74.5% | 15.7% | 19.7% |
| Parsimonious machine-learning model | 100 | 682 | 82 | 3049 | 54.9% | 12.8% | 19.3% | 75.9% | 17.3% | 26.1% |
| Parsimonious logistic regression model | 95 | 687 | 87 | 3045 | 52.2% | 12.1% | 17.8% | 73.7% | 13.6% | 12.4% |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| TP: true positives FP: false positives FN: false negatives TN: true negatives MCC: Matthews correlation coefficient AUROC: area under the operating receiver curve AUPRC: area under the precision recall curve P(sensitivity): probability that a model performs better than the machine-learning model relative to sensitivity. | | | | | | | | | | |

When considering risk scores from the full machine-learning (figure 1a) and full logistic regression model leading to this risk-patient selection, 106 and 98 had the primary outcome, respectively. Correspondingly, the sensitivity and precision were 58.2% and 13.6% for the full machine-learning and 53.8% and 12.5% for the full logistic regression model, respectively. The full machine-learning model was superior (figure 1b) on essentially all parameters compared to any of the other models, although the differences were minor (table 2). Supplementary Material 3However, the machine-learning model excluding age performed as well as the age-only model d

The results were similar when using positive prediction fractions of 25% and 30%, but with the sensitivity for the full machine-learning model increasing to 64.4% and 69.2% and precision decreasing to 12.0% and 10.7%, respectively (Supplementary Material 3).

When evaluating feature importance, we found a strong correlation between the full machine-learning and full logistic regression model, with age and use of walking aids being the most important variables in both (figure 2a). From the combined importance of variables outside the top ten, the machine-learning approach extracted more information with fewer variables than logistic regression (figure 1b).

For the full machine-learning model, there was a clear signal that increasing age, number of reimbursed prescriptions, and presence of comorbidity, all contributed to an increased risk score. In contrast, a recent date of surgery and an increased hemoglobin level seemed to reduce the calculated risk (figure 2b). Individual analysis of the SHAP interaction values for types of anticoagulant prescriptions revealed that prescriptions on vitamin-K antagonists (VKA) or adenosine diphosphate (ADP) antagonists increased, while acetylic salicylic acid and direct oral anticoagulants (DOAC) reduced the risk score of the full machine-learning model, regardless of age (figure 3a). The SHAP analysis of prescribed cardiac drugs revealed that prescriptions on Ca2+-antagonists and betablockers in combination with one or two other antihypertensives increased the risk-score, as did prescriptions on nitrates, other antihypertensives and antiarrhythmics. For the remaining cardiac drugs, prescriptions either reduced or had minor influence, and with limited relation with age (figure 3b). Preoperative psychotropic prescriptions increased the risk-score except for antipsychotics (0.6%). For users of selective serotonin inhibitors there was a clear age-related distinction with the risk score being increased in elderly patients but decreased in those < 60 years (figure 3c). Finally, the risk score increased with prescriptions on inhalation steroid and β-blockers, and more accentuated in the younger patients (figure 3d).   
The results including patients with a length of stay >4 days, but no reported postoperative complications (secondary outcome) were similar as for the primary outcome. In general, we found that the full machine-learning model was superior to the others, although the differences were smaller than for the primary outcome. (Supplemental material 4 listing outcome parameters and Supplemental material 5 figure S1a-b showing distributions and ROC curves for the secondary outcome). While the ten most important variables for the full machine-learning model remained unchanged, familiar disposition for venous thromboembolism replaced gender as one of the top ten important variables in the full logistic regression model (Supplemental material 6 figure S2a-b showing SHAP values for the secondary outcome). Furthermore, the SHAP analysis on specific prescribed drugs demonstrated that the machine-learning model found no benefits from information on prescriptions on respiratory drugs, why all SHAP values were zero. In addition, the reduced risk with acetylsalicylic acid and DOAC prescriptions, as well as the influence of practically all cardiac drugs except for nitrates, other antihypertensives and antiarrhythmics, was attenuated (Supplemental material 7 figure S3a-d showing SHAP-values of prescriptions of specific drugs for the secondary outcome).

**DISCUSSION**

We found that using a machine-learning algorithm including all 33 available variables and a parsimonious machine-learning-algorithm encompassing only the 10 most important predictors improved prediction of patients at increased risk of having a length of stay >4 days or readmissions due to medical complications compared to traditional logistic regression models. In contrast, when also including patients having a length of stay >4 days but without a well-defined complication as an outcome, the parsimonious machine-learning model was slightly worse than a traditional logistic regression model including all variables. We also found that although age was the single most important predictor of both the primary and the secondary outcome, it was less suited for prediction of postoperative medical complications after fast-track THA and TKA on its own. Finally, we demonstrated how the chosen classification threshold of the machine-learning algorithm influenced model performance through an increase in sensitivity at the cost of decreased precision.

A previous systematic review also found that machine-learning algorithms may provide better prediction of postoperative outcomes in THA and TKA.32 However, the authors concluded that such models performed best at predicting postoperative complications, pain and patient reported outcomes and were less accurate at predicting readmissions and reoperations.32 That machine-learning algorithms may improve prediction of complications after THA and TKA compared to traditional logistic regression was also found by Shah *et al.* who used an automated machine-learning framework to predict selected major complications after THA.13 However, theirs was a retrospective study based on diagnostic and administrative coding and the selected complications occurred only in 0.61% of patients, potentially limiting clinical relevance. In contrast, we aimed at identifying a cohort which would comprise 20% of patients in which we found about 60% of all medical complications. This we believe, is within the means of the Danish socialized healthcare system to allocate additional resources for intensified perioperative care and with both patient-related and economic benefits due to potentially avoided complications and costs.  
In contrast to many other machine-learning studies,33 our dataset included not only preoperative data but also only one paraclinical variable, which was preoperative hemoglobin. Although the inclusion of other laboratory tests such as preoperative albumin, sodium and alkaline phosphatase has been found to be of importance in machine-learning algorithms for home discharge in uni-compartmental knee replacement12 and spine surgery,9 they are not standard in fast-track protocols and not easy to interpret from a pathophysiological point of view.   
Most decisions on which patients may benefit from more extensive postoperative care will likely need to be conducted preoperatively, as there is an increasing need to prioritize limited health-care resources. Thus, although postoperative information such as duration of surgery, perioperative blood length of stays or postoperative hemoglobin have been included in other studies33, we decided against the use of peri- and postoperative data. The same approach has been used by Ramkumar *et al.* who used U.S. National Inpatient Sample data including 15 preoperative variables, to predict length of stay, patient charges and disposition after both TKA34 and THA.18 However, these studies were not conducted in a socialized health care system, and the main focus was on the need for differentiated payment bundles and without specific information on the reason for increased length of stay or non-home discharge.34 Wei *et al.* used an artificial neural network model to predict same-day discharge after TKA, based on the NSQUIP database from 2018 and found that six of the ten most important variables were the same compared with logistic regression, similar to our findings.35 However, patients with one-day length of stay were intentionally excluded due to variations in in-patient vs. out-patient registration.35   
Age has traditionally been a major factor when predicting surgical outcomes which is why we choose to specifically evaluate its effect on our risk-prediction. That age is important for risk-prediction was further illustrated by the machine-learning model without age being comparable to the age-only model. Note that, although elderly patients had increased risk of postoperative complications, likely related to decline of physical reserves,36 the use of chronological age alone as a selection criteria for being a “risk-patient” was inferior compared to both machine-learning and logistic regression models incorporating comorbidity and functional status.

We used the SHAP values for estimation of feature importance, thus providing a better understanding of the otherwise “black-box” machine-learning model. The SHAP values showed which variables contribute most to the risk-score predictions. In this context, inclusion of specific data on reimbursed prescriptions 6 months prior to surgery based upon the unique Danish registries, unsurprisingly found increased risk-scores with increased number of prescriptions and with the majority being in elderly patients. Similarly, a Canadian study in elective non-cardiac surgery found decreased survival and increased length of stay and readmissions and costs in patients >65 years with polypharmacy.28 However, this is a complex relationship where some patients benefit from their treatments, while other may suffer from undesirable side-effects. Consequently, the authors cautioned against altering perioperative practices based on current evidence.28 However, the information from the included prescriptions with SHAP analysis may provide inspiration for new hypothesis-generating studies such as investigation of the potential differences in risk-profile between having preoperative prescribed VKA and DOAKs. Also, the age-related differences in risk from SSRI’s in our study could guide further studies on “deprescription”.

Another requirement for machine-learning-algorithms to be clinically useful is user friendliness and not depending on excessive additional data collection by the attending clinicians. In this context, it was a bit disappointing that the parsimonious machine-learning algorithm with only the ten most important variables was slightly worse at predicting the secondary outcome than the full logistic regression model. A reason for this could be that when including a length of stay >4 days but without described medical complications, the combination of all variables provides information not available by merely including the ten most important ones. This highlights the need for as much detailed, and preferably non-binary, data as possible to fulfill the true potential of machine-learning algorithms.

Our study has some limitations. First, one of the strengths of machine learning compared to logistic regression is the analysis of multilevel continuous data, whereas we included only a limited number of, often binary, preoperative variables. This could have limited the full realization of our machine-learning model. As previously mentioned, we excluded intraoperative information, including type of anesthesia, surgical approach etc. all of which may influence postoperative outcomes. The observational design of this study means that we cannot exclude unmeasured confounding or confounding by indication. Also, despite that the DNDRP has a near complete registration of dispensed medicine in Denmark, some types or drugs, especially benzodiazepines, are exempt from general reimbursement and thus not sufficiently captured.21 Furthermore, it is doubtful whether the patients used all types of drugs at the time of surgery (e.g. heparin which is rarely for long-term use). Finally, classification of a complication being “medical” depended on review of the discharge records which can also introduce bias. However, we believe our approach to be superior to depending only on diagnostic codes which often are inaccurate37 and provide limited details on whether the complication may be attributed to a medical or surgical adverse event. The strengths of our study include the use of national registries with high degree of completion (>99% of all somatic admissions in case of the DNDRP),38 prospective recording of comorbidity, extensive information on prescription patterns 6 months prior to surgery and similar established enhanced recovery protocols in all departments.

In summary, our results suggest that machine-learning-algorithms likely provide clinically relevant improved predictions for defining patients in high-risk of medical complications after fast-track THA and TKA compared to a logistic regression model. Future studies could benefit from using such algorithms to find a manageable population of patients who benefit from intensified perioperative care.

**Competing interests**: Prof. Kehlet is a board member of “Rapid Recovery”, by Zimmer Biomet. Mr. Heltberg is sponsored by a grant from the Lundbeck Foundation, independently of the present study. Dr. Petersen is an advisory member of Sanofi outside of the present study. The remaining authors declare no conflicts of interest.

**Funding**: The study received funding from the Lundbeck Foundation, Copenhagen, Denmark grant number R25-A2702.

**Author Contributions:** CJ and HK contributed to the original idea of the study. CJ, PB and HK contributed to data collection and review of medical records. CM, MH, MJ, AL and TP contributed to the statistical methods, designed the prediction models and conducted the statistical analysis. CM, CJ, HK and TP wrote the original draft. All authors contributed to revision of the initial draft and agreed on the final version of the manuscript. The members of the Centre for Fast-track Hip and Knee Replacement Database collaborative group all contributed by implementing the fast-track protocol at their respective departments and reviewing the final manuscript.

**Collaborators**:

Frank Madsen M.D. Consultant, Department of Orthopedics, Aarhus University Hospital, Aarhus, Denmark

Torben B. Hansen M.D., Ph.D., Prof. Department of Orthopedics, Holstebro Hospital, Holstebro, Denmark

Kirill Gromov, M.D., Ph.D., Ass.Prof. Department of Orthopedics Hvidovre Hospital, Hvidovre Denmark

Thomas Jakobsen, M.D., Ph.D., DM.Sci., Ass. Prof. Department of Orthopedics, Aalborg University Hospital, Farsø, Denmark

Claus Varnum, M.D., Ph.D., Ass. Prof. Department of Orthopedic Surgery, Lillebaelt Hospital - Vejle, University Hospital of Southern Denmark, Denmark

Soren Overgaard, M.D., DM.Sci., Prof, Department of Orthopedics, Bispebjerg Hospital, Copenhagen, Denmark

Mikkel Rathsach, M.D., Ph.D., Ass. Prof. Department of Orthopedics, Gentofte Hospital, Gentofte, Denmark

Lars Hansen, M.D., Consultant, Department of Orthopedics, Sydvestjysk Hospital, Grindsted, Denmark

**Data sharing:** Statistical code and dataset is not publicly available due to Danish data-protection law but can be acquired from the corresponding author by request.**References**

1. Petersen PB, Kehlet H, Jorgensen CC, et al. Improvement in fast-track hip and knee arthroplasty: a prospective multicentre study of 36,935 procedures from 2010 to 2017. *Scientific reports* 2020;**10**(1):21233.

2. Khan SK, Malviya A, Muller SD, et al. Reduced short-term complications and mortality following Enhanced Recovery primary hip and knee arthroplasty: results from 6,000 consecutive procedures. *Acta Orthop* 2014;**85**(1):26-31.

3. Partridge T, Jameson S, Baker P, et al. Ten-Year Trends in Medical Complications Following 540,623 Primary Total Hip Replacements from a National Database. *The Journal of bone and joint surgery American volume* 2018;**100**(5):360-67.

4. Jorgensen CC, Gromov K, Petersen PB, et al. Influence of day of surgery and prediction of LOS > 2 days after fast-track hip and knee replacement. *Acta orthopaedica* 2021;**92**(2):170-75.

5. Jorgensen CC, Petersen MA, Kehlet H. Preoperative prediction of potentially preventable morbidity after fast-track hip and knee arthroplasty: a detailed descriptive cohort study. *BMJ Open* 2016;**6**(1):e009813.

6. Johns WL, Layon D, Golladay GJ, et al. Preoperative Risk Factor Screening Protocols in Total Joint Arthroplasty: A Systematic Review. *J Arthroplasty* 2020;**35**(11):3353-63.

7. Adhia AH, Feinglass JM, Suleiman LI. What Are the Risk Factors for 48 or More-Hour Stay and Nonhome Discharge After Total Knee Arthroplasty? Results From 151 Illinois Hospitals, 2016-2018. *J Arthroplasty* 2020;**35**(6):1466-73 e1.

8. Shah A, Memon M, Kay J, et al. Preoperative Patient Factors Affecting Length of Stay following Total Knee Arthroplasty: A Systematic Review and Meta-Analysis. *J Arthroplasty* 2019;**34**(9):2124-65 e1.

9. Li Q, Zhong H, Girardi FP, et al. Machine Learning Approaches to Define Candidates for Ambulatory Single Level Laminectomy Surgery. *Global spine journal* 2021:2192568220979835.

10. Chiew CJ, Liu N, Wong TH, et al. Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission. *Annals of surgery* 2020;**272**(6):1133-39.

11. Li H, Jiao J, Zhang S, et al. Construction and Comparison of Predictive Models for Length of Stay after Total Knee Arthroplasty: Regression Model and Machine Learning Analysis Based on 1,826 Cases in a Single Singapore Center. *The journal of knee surgery* 2022;**35**(1):7-14.

12. Lu Y, Khazi ZM, Agarwalla A, et al. Development of a Machine Learning Algorithm to Predict Nonroutine Discharge Following Unicompartmental Knee Arthroplasty. *J Arthroplasty* 2021;**36**(5):1568-76.

13. Shah AA, Devana SK, Lee C, et al. Development of a Novel, Potentially Universal Machine Learning Algorithm for Prediction of Complications After Total Hip Arthroplasty. *J Arthroplasty* 2021;**36**(5):1655-62 e1.

14. Sniderman J, Stark RB, Schwartz CE, et al. Patient Factors That Matter in Predicting Hip Arthroplasty Outcomes: A Machine-Learning Approach. *J Arthroplasty* 2021;**36**(6):2024-32.

15. Kugelman DN, Teo G, Huang S, et al. A Novel Machine Learning Predictive Tool Assessing Outpatient or Inpatient Designation for Medicare Patients Undergoing Total Hip Arthroplasty. *Arthroplasty today* 2021;**8**:194-99.

16. Mohammadi R, Jain S, Namin AT, et al. Predicting Unplanned Readmissions Following a Hip or Knee Arthroplasty: Retrospective Observational Study. *JMIR medical informatics* 2020;**8**(11):e19761.

17. Ramkumar PN, Navarro SM, Haeberle HS, et al. Development and Validation of a Machine Learning Algorithm After Primary Total Hip Arthroplasty: Applications to Length of Stay and Payment Models. *J Arthroplasty* 2019;**34**(4):632-37.

18. Ramkumar PN, Karnuta JM, Navarro SM, et al. Preoperative Prediction of Value Metrics and a Patient-Specific Payment Model for Primary Total Hip Arthroplasty: Development and Validation of a Deep Learning Model. *J Arthroplasty* 2019;**34**(10):2228-34 e1.

19. Haeberle HS, Helm JM, Navarro SM, et al. Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty: A Review. *J Arthroplasty* 2019;**34**(10):2201-03.

20. Johannesdottir KB, Kehlet H, Petersen PB, et al. Machine learning classifiers do not improve prediction of hospitalization > 2 days after fast-track hip and knee arthroplasty compared with a classical statistical risk model. *Acta orthopaedica* 2022;**93**:117-23.

21. Johannesdottir SA, Horvath-Puho E, Ehrenstein V, et al. Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. *ClinEpidemiol* 2012;**4**:303-13.

22. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine* 2015;**162**(1):W1-73.

23. Olczak J, Pavlopoulos J, Prijs J, et al. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal. *Acta orthopaedica* 2021;**92**(5):513-25.

24. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T. LightGBM: a highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc, 2017:3149-57.

25. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. Secondary Focal Loss for Dense Object Detection 2018. <http://arxiv.org/abs/1708.02002>.

26. Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature machine intelligence* 2020;**2**(1):56-67.

27. Lundberg SMLSI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, ed. Adv Neural Inf Process Syst [Internet]: Curran Associates, Inc., 2017.

28. McIsaac DI, Wong CA, Bryson GL, et al. Association of Polypharmacy with Survival, Complications, and Healthcare Resource Use after Elective Noncardiac Surgery: A Population-based Cohort Study. *Anesthesiology* 2018;**128**(6):1140-50.

29. Chicco D. Ten quick tips for machine learning in computational biology. *BioData mining* 2017;**10**(1):35 (2017).

30. Chicco D, Totsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining* 2021;**14**(1):13 (2021).

31. Totsch N, Hoffmann D. Classifier uncertainty: evidence, potential impact, and probabilistic treatment. *PeerJ Computer science* 2021;**7**:e398.

32. Lopez CD, Gazgalis A, Boddapati V, et al. Artificial Learning and Machine Learning Decision Guidance Applications in Total Hip and Knee Arthroplasty: A Systematic Review. *Arthroplasty today* 2021;**11**:103-12.

33. Han C, Liu J, Wu Y, et al. To Predict the Length of Hospital Stay After Total Knee Arthroplasty in an Orthopedic Center in China: The Use of Machine Learning Algorithms. *Frontiers in surgery* 2021;**8**:606038.

34. Ramkumar PN, Karnuta JM, Navarro SM, et al. Deep Learning Preoperatively Predicts Value Metrics for Primary Total Knee Arthroplasty: Development and Validation of an Artificial Neural Network Model. *J Arthroplasty* 2019;**34**(10):2220-27 e1.

35. Wei C, Quan T, Wang KY, et al. Artificial neural network prediction of same-day discharge following primary total knee arthroplasty based on preoperative and intraoperative variables. *Bone Joint J* 2021;**103-B**(8):1358-66.

36. Griffiths R, Beech F, Brown A, et al. Peri-operative care of the elderly. *Anaesthesia* 2014;**69 Suppl 1**:81-98.

37. Bedard NA, Pugely AJ, McHugh MA, et al. Big Data and Total Hip Arthroplasty: How Do Large Databases Compare? *J Arthroplasty* 2018;**33**(1):41-45.e3.

38. Schmidt M, Schmidt SA, Sandegaard JL, et al. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clinical epidemiology* 2015;**7**:449-90.

**FIGURE LEGENDS**

**Figure 1a-b**

1a) Distribution of full machine learning model risk scores for patients +/- the primary outcome. The dashed line marks the classification threshold of 20% positive prediction fraction.   
1b) Receiver operating curves (ROC) for the full machine learning model (F-MLM), full logistic regression model (F-LRM), parsimonious machine learning model (P-MLM), parsimonious logistic regression model (P-LRM), machine learning excluding age (MLM -age) and the age-only model (AM).

**Figure 2a-b**

2a) The overall importance of the 10 most important variables measured by the SHAP-values for the full machine-learning and full logistic regression models on the primary outcome (LOS >4 days or readmission due to “medical” morbidity). Only the importance of prescribed anticholesterols and gender differ between the models. The contributions of the remaining variables are summed in the bottom bar.   
2b) The SHAP-values for the full machine-learning model on the primary outcome, where positive increase and negative values decrease the risk score. The color is related to the value of the variable with blue being lowest and red highest and each dot represents a patient.

**Figure 3a-d**

SHAP scatter-plot on the contributions to the full machine-learning model on the primary outcome (LOS >4 days or readmission due to “medical” morbidity), for individual types of prescribed anticoagulants, cardiac drugs, psychotropics and respiratory drugs stratified by age.

3a) Prescribed anticoagulants

VKA: vitamin K antagonists ASA: acetylsalicylic acid DOAC: direct oral anticoagulant ADP: Adenosine diphosphate ACE: angiotensin converting enzyme

3b) Prescribed cardiac drugs

ACE: angiotensin converting enzyme AHT: antihypertensive. Other AHT were defined as AHT different from diuretics ANG-II/ACE inhibitors or Ca2+antagonists. IHD: Ischemic heart disease

3c) Prescribed psychotropics

SSRI: Selective serotonin inhibitor SNRI: Serotonin and norepinephrine reuptake inhibitor NaRI: Norepinephrine reuptake inhibitor NaSSA: Norepinephrine and specific serotonergic antidepressants. AD: antidepressants BZ: Benzodiazepines (likely underreported due to limited general reimbursement in Denmark). ADHD: Attention-deficit/hyperactivity disorder

3d) Prescribed respiratory drugs

SABA: Short-acting beta agonist LABA: long-acting beta agonist LAMA: Long-acting muscarinic antagonist.