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

Christian Michelsen1, 3, Christoffer C Jørgensen2, 3, Mathias Heltberg1, Mogens H Jensen1, Alessandra Luchetti1, Pelle B Petersen2, Troels Petersen1, Henrik Kehlet2.

1. Niels Bohr Institute, University of Copenhagen, Denmark

2. Section of Surgical Pathophysiology 7621, Rigshospitalet, Copenhagen, Denmark

3. Joint first authors

**Corresponding author:**

NAME Christoffer C Jørgensen

Email address: christoffer.calov.joergensen@regionh.dk

**Text pages** (including tables):

Number of Words (excluding references):

Number of Tables: ; Number of Figures:

**Conflict of interest**: No conflict of interest.

**Funding**: None.

# Abstract

# INTRODUCTION

Prediction of postoperative morbidity and requirement for hospitalisation is important for planning of health care resources. With regard to the common surgical procedures of primary total hip and knee arthroplasty (THA and TKA), the introduction of enhanced recovery or fast-track programs have led to a significant reduction of postoperative length of stay (LOS) as well as morbidity and mortality.1-4 However, despite this progress, a fraction of patients still have postoperative complications leading to prolonged LOS or readmissions.1, 4, 5 Consequently, in order to prioritize perioperative care, many efforts have been published to preoperatively predict LOS and morbidity using traditional risk factors such as age, preoperative cardio-pulmonary disease, anaemia, diabetes, frailty, etc.5-10 These efforts have been based on traditional statistical methods including multiple regression analyses and essentially concluding that it is “better to be young and healthy than old and sick”. Also, despite being statistically significant, conventional risk-stratification based on such studies have had a relatively limited clinically relevant ability to predict potentially preventable morbidity and LOS.5-11

More recently, machine learning (ML) 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 anaesthetic and surgical conditions12, 13 This is also the case in in THA and unicompartmental knee replacement (UKA), where several publications on ML-algorithms for prediction of LOS14, 15, complications,16 disability,17 potential outpatient setup,18 readmissions19 or payment models20, 21, have shown promising predictive value compared to conventional statistical methods.22

However, few of such papers have considered the presence of enhanced recovery programs, as most are based on large database cohorts with the presence of risk factors and complications often being based on diagnostic codes and limited information on perioperative care, follow-up and discharge destination. Despite that such an approach may be suboptimal compared to patient specific comorbidity and outcome assessments independent from economic reimbursement systems, ML predictive models seem 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 from a well-established fast-track THA/TKA setup in a national public health-care system and with a median LOS of 1 day,1 to investigate whether a ML-algorithm was able to improve prediction of medical complications resulting in prolonged LOS and readmissions compared to a traditional logistic regression model.

# Method

Reporting of the study is done in accordance with the Transparent reporting of multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement23 and the Clinical AI Research (CAIR) checklist proposal.24  
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 between 2010 and 2017. The database is registered on ClinicalTrials.gov as a study registry (NCT01515670). 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). Patients completed the preoperative questionnaire with nurse assistance if needed and additional information on reimbursed prescriptions 6 months prior to surgery was acquired using the Danish National Registry of Reimbursed Prescriptions which records all dispensed prescriptions with reimbursement in Denmark.25 Finally, data was crossed with the Danish National Patient Registry (DNPR) for information on LOS (counted as nights spent in hospital), 90-days readmissions with overnight stay and mortality. In case of LOS >4 days or readmissions 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 always reviewed to identify cause of death. Evaluation of discharge and medical records was by CJ until 2015, and then by PP supervised by CJ. In case of disagreement, records were conferred with HK. Subsequently, causes of LOS >4, readmissions or mortality were classified as “medical” when related to perioperative care (renal failure, falls, pain, thrombosis, anaemia, disproved 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 LOS >4 days with a standard discharge summary describing a successful postoperative course, it was assumed that no clinically relevant postoperative complications had occurred.   
For the present study, only cases between 2014 and 2017 were used in order to provide the most up-to date data. All patients had elective unilateral THA or TKA in dedicated arthroplasty departments with similar fast-track protocols, including multimodal opioid sparing analgesia with high-dose (125mg) methylprednisolone, preference for spinal anaesthesia, only in-hospital thromboprophylaxis when LOS ≤5 days, early mobilization, functional discharge criteria and discharge to own home.1 There are 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 their THA or TKA, and patients with severe congenital joint disorder and cancer surgery.

## Outcomes

The primary outcome was to compare accuracy in prediction when using a machine learning algorithm to predict the occurrence of “medical” complications resulting in a LOS >4 days or readmission compared to a traditional logistic regression model (outcome A). Secondarily, we investigated whether a parsimonious model with a limited number of covariates would be equal to the full-scale model with all available variables and whether the effect of age per se would compare to the full-scale ML-model. Finally, we investigated how inclusion of cases with a LOS >4 days but no reported medical complication as positive outcome influenced the model (outcome B).

## Statistical Analysis

Data was initially trimmed by removing 156 patients (1.7%) who either had missing data or were outliers with regards to weight (<30 kg or >250 kg) and height <100 cm or >210 cm). To reduce the risk of overfitting, data was split into a training set consisting of 18.013 procedures with surgery between 2014-2016 and a test set of 3914 procedures with surgery in 2017, using time series analyses techniques.

We used a Boosted Decision Trees (LightGBM [Ref]) for the machine learning model (ML) as it is a tree-based method and, unlike neural networks, works well with categorical data and missing values. We tried using both normal cross entropy and FocalLoss [Ref] as the objective function for the ML model. The reason behind also testing FocalLoss was to allow the ML model to focus more on the (few) positives.

The ML model is trained and hyperparameter optimized using state of the art ML 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 are predicted by the model to experience complications, i.e. a positive predictive fraction (PPF) of 20%.

To investigate the importance of the included variables , we computed the normalized SHAP values which provide estimates on which variables contribute most to the risk score predictions.26, 27 Based on clinical experience we also investigated potential relation between for Prescrip\_Card vs. reported hypertension, Prescrip\_Anticoag vs. reported family VTE, Prescrip\_Psych vs. reported psychiatric disorder and Prescrip\_Resp vs. age.  
For comparing model performance we used a classical logistic regression (LR) model. Cases of missing values in the logistic regression were handled by multiple imputations using the median of present values and all variables were normalized.

We also trained two parsimonious models using ML and LR using only the 10 most important features (ML10, LR10) (Figure 3a) Finally, we specifically explored the influence of increasing age, by constructing a model based only on age (Age), and a ML model based on all variables but age (ML-NoAge). To compare the different classifiers, we applied a Bayesian metric comparison.28   
For evaluating model performance we evaluated area under the receiver operating characteristic curve (AUROC), area under the precision recall curve (AUPRC), and the probability that the model performed better than the ML model relative to the TPR metric ( P(TPR)). For two equal models, the P(TPR) value would be approximately 50%.28

Since the data is 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).  
We also calculated number of true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), true positive rate (TPR) or sensitivity, positive predictive value (PPV) or precision in order to evaluate clinical relevance.

# **Results**

Of the 3913 patients median age was 69 years (IQR 62-75), 59% were female and had 58% had THA (table 1)`. Median LOS was 1 (IQR: 1-2) day with 7.6% 90-days readmissions and outcome A ocurring in 182 (4.6%) patients.  
When applying a PPF of 20% 782 patients qualified as “risk-patients” with a ML\_all risk-score threshold of 0.403 (figure 1a). Of the 782 patients, 106 and 98 had outcome A resulting in a PPV of 13.6% and 12.5% for ML\_all and LR\_all, respectively. Correspondingly TPR was 58.2% for ML\_all and 53.8% for LR\_all. (table 2). ML\_all was superior on all parameters compared to any of the other model, although the differences in PPV compared to, LR\_all, ML10 and LR10 were ≤1.5% (table 2). Both the ML\_NoAge and Age models had fewer TP and lower PPV than the models including age and other variables (table 2, Figure 1b), however the TPR in ML\_NoAge and Age were similar, being 48.4% and 47.8%. respectively (table 2). The highest P(TPR) was in the ML\_10 (26.1%), indicating a low likelihood of any model having a better than the performance than ML\_All. The results were similar when using a PPF of 25% and 30%, however TPR improved to 64.4% and 69.2% while PPV was reduced to 12.0% and 10.7% for a PPF of 25% and 30%, respectively. (Table 2)

When evaluating feature importance we found a strong correlation between ML\_All and LR\_All.(Figure 2a), with age and use of walking aids being the most important variables in both models. However, the combined importance the variables outside the top ten was less in ML\_all (22.5%) than in LR\_all (31.6%). In ML\_all specifically, there was a clear signal that increasing age and number of reimbursed prescriptions, declining haemoglobin and presence of comorbidity all contributed to an increased risk-score. In contrast, recent date of surgery and high BMI seemed to reduce the calculated risk. (Figure 2b)

Individual analysis of the SHAP interaction values of the anticoagulant prescriptions vs. having a family member with VTE revealed that prescriptions on VKA or ADP-antagonists increased the ML\_all risk score, while ASA and DOAK prescriptions reduced the calculated risk independently of reported familiar VTE (Figure 4a). In contrast, SHAP analysis of prescribed cardiac medicine vs. reported hypertension demonstrated that patients with no reported hypertension generally had no prescriptions and reduced risk. However, having prescriptions on Calcium antagonists, or betablokkers with 1 or 2 (but not 3) other antihypertensives increased risk. Finally, prescriptions on nitrates, other antihypertensives and antiarrhythmics also increased risk score. (Figure 4b) Prescriptions on psychotropics vs. reported psychiatric disorder (PsD) revealed that most patients reporting PsD had reimbursed prescriptions, except for those using “other psychotropics” (group 8). Having psychotropic prescriptions increased the ML\_All risk score except for those using antipsychotics (0.6%) which had no influence on the risk score, and for SSRI users where the ML\_all risk score was increased in patients > xx years but decreased with age < xx. (Appendix figure A1) Finally, ML\_all risk score was increased in those with prescriptions on inhalation steroid and beta-agonists, however there was a clear association with increased age as all patients with prescriptions on such pulmonary medicine were >70 years (Appendix Figure A1). The results for including patients with a LOS >4 days, but no reported postoperative complications (Outcome B) were similar as for Outcome A. Thus, ML\_all was superior than LR\_all. However, most model performance indicators were slightly worse for in models, except for the PPV. Thus, although TPR declined to 52.8% with an AUROC of 75.3%, PPV increased to 15.5% for Outcome B when applying the ML\_all with a PPF of 20%. (Appendix table A1 and Figure A2a+b)

# Discussion

We found that using a ML algorithm including all available variables (ML\_all) and a parsimonious ML-algorithm (ML\_10) encompassing only the 10 most important predictors improved prediction of patients at increased risk of having a LOS >4 days or readmissions due to medical complications (outcome A) compared to traditional logistic regression models. In contrast, when also including patients having a LOS >4 days, but without mentioned complications as an outcome (Outcome B), the ML\_all remained superior, but the parsimonious ML-algorithm was worse than a traditional logistic regression model including all variables (LR\_all). We also found that although age was the single most an important predictor of both outcome A and B, it was not well suited for prediction of postoperative medical complications after fast-track THA and TKA on its own. Finally, we demonstrated how the chosen calibration point of the ML algorithm influenced model performance through an increase in TPR at the cost of decreased PPV.

A previous systematic review also found that ML algorithms can provide accurate prediction of postoperative outcomes in THA and TKA.31 However, the authors concluded that such models performed best at predicting postoperative complications, pain and patient reported outcomes but were less accurate at predicting readmissions and reoperations.31 That ML 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 ML framework to predict selected major complications after THA.16 However, it 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 approximately 20% of patients in whom about 60% of all medical complications would occur. This we believe is within the means of the Danish socialized healthcare system to allocate additional resources for intensified perioperative care.  
In contrast to many other ML studies,32 our dataset included only preoperative data and the only paraclinical data was preoperative haemoglobin. Although the inclusion of other laboratory tests such as preoperative albumin, sodium and alkaline phosphatase has been found to be of importance in ML algorithms for home discharge in UKA15 and spine surgery,12 they are not standard in our fast-track protocols. Furthermore, as there is a need to prioritize the limited health-care resources, most decisions on which patients may benefit from more extensive postoperative care will be conducted preoperatively. Thus, although postoperative information such as duration of surgery, perioperative blood loss or postoperative haemoglobin have been included in other studies32, we decided against the use of any peri- and postoperative data. The same approach has been used by Ramkumar and colleagues who used U.S. National Inpatient Sample data, including 15 preoperative variables in order to predict LOS, patient charges and disposition after both TKA33 and THA.21 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 reason for increased LOS or non-home discharge.33 Wei et al. used an artificial neural network model to predict same-day discharge after TKA, based on the NSQUIP database from 2018. They also compared with logistic regression and found that six of the ten most important variables were similar much like in our study. However, patients with one-day LOS were intentionally excluded due to variations in in-patient vs. outpatient registration.34   
Age has traditionally been a major consideration with regards to surgical procedures why we choose to specifically evaluate its effect on our risk-prediction. That age is important for risk-prediction was illustrated by ML\_noAge being comparable to Age\_only. However, although elderly patients had increased risk of postoperative complications, likely related to decline of physical reserves,35 using chronological age alone as a selection criteria for being a surgical “risk-patient” was much inferior compared to both ML and LR models incorporating comorbidity and functional status. We used the SHAP values for both local and global estimation of the feature importance, thus providing a better understanding of the otherwise “black-box” ML model. SHAP values show which variables, across all patients, contribute most to the risk score predictions. In this context our data on reimbursed prescriptions 6 months prior to surgery unsurprisingly found that risk increased with increased number of prescriptions. A Canadian study in elective non-cardiac surgery found decreased survival and increased LOS and readmissions and costs in patients with polypharmacy.36 However, this is a complex relationship where some patients benefit from their treatment while other may suffer from undesirable side-effects and the authors cautioned against altering perioperative practices based on current evidence.36 However, the information from SHAP values may provide inspiration for new hypothesis generating studies, i.e. investigating the potential differences in risk-profile between having preoperative prescribed VKA and DOAK as seen in our study.

Another requirement for ML-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 ML-algorithm with only the 10 most important variables was considerably worse at predicting outcome B than LR\_All. A reason for this could be that when including those patients who had a LOS >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 data as possible in order to full-fill the true potential of ML-algorithms.

Our study has some limitations. Thus, we included only a limited number of, often binary, preoperative variables. As analysis of multilevel continuous data is one of the strengths of ML compared to logistic regression, this may limit full realisation of our ML-model. As previously mentioned, we also excluded intraoperative information, including type of anaesthesia, prosthesis etc. all of which may influence postoperative outcomes. The observational design of study means that we cannot exclude unmeasured confounding or confounding by indication. Also. Despite that the DNPR has a near complete registration of dispensed medicine in Denmark some types or drugs, especially benzodiazepines, are exempt from general reimbursement and thus not captured.25 Finally, classification of a complication being “medical” depended on review of the discharge records which can also introduce bias. However, we believe this approach to be superior than depending on diagnostic codes which often are inaccurate and provide limited details on whether the complication may be attributed to a medical or surgical adverse event. Strengths of our study include use of national registries, with high degree of completion (>99% of all somatic admissions in case of the DNPR),37 prospective recording of comorbidity and similar perioperative care in all departments.

In summary, our results indicate that ML-algorithms may provide better and clinically relevant improvements for defining a “high-risk” population for medical complications in 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 may benefit the most from intensified perioperative care.

# Tables

Table 1. patient demographics

| **Characteristic** | **+outcome A** | **-outcome A** |
| --- | --- | --- |
| Characteristics based on preoperative questionaire |  |  |
| age |  |  |
| female gender |  |  |
| THA |  |  |
| weight |  |  |
| height |  |  |
| body mass index |  |  |
| need of walking aid  missing |  |  |
| living alone  with others  institution  missing |  |  |
| haemoglobin |  |  |
| >2 units of alcohol/day  missing |  |  |
| reported active smoker  missing |  |  |
| reported cardiac disease  missing |  |  |
| reported hypercholesterolaemia  missing |  |  |
| reported hypertension  missing |  |  |
| reported pulmonary disease  missing |  |  |
| reported previous cerebral attack |  |  |
| reported previous VTE  missing |  |  |
| reported active cancer treatment  reported previous cancer treatment  missing |  |  |
| reported chronic kidney disease  missing |  |  |
| reported family member with VTE |  |  |
| reported regular snoring  reported uncertain about snoring |  |  |
| reported not feeling rested morning  reported uncertain about being rested |  |  |
| reported psychiatric disorder |  |  |
| Characteristics based on combination of questionnaire and DNPR | | |
| diabetes  diet treated diabetes1  oral antidiabetics  insulin treated diabetes2 |  |  |
| Characteristics based on 6 months reimbursed prescriptions from the DNPR | | |
| total number of reimbursed prescriptions on drugs listed below |  |  |
| Anticoagulants  none  VKA  Heparin+ASA  DOAK  ASA  Dipyradimol  ADP  ASA+Dipyradimol  VKA+ASA  DOAK+ASA  VKA+ADP  DOAK+ADP  VKA+Heparin  DOAK+ASA+ADP  ASA+ADP  ASA+ADP+Heparin  ASA+ADP+Dipyradimol |  |  |
| Cardiac drugs  none  diuretics  angiotensin 2 or ACE-inhibitors  calcium antagonists  betablockers  nitrates  other antihypertensives  other drugs for IHD  other antiarrhythmics  2 antihypertensives  Betablocker +1 antihypertensive  3 antihypertensives  Betablocker +2 antihypertensives  Betablocker +3 antihypertensives  4 antihypertensives  betablocker+4 antihypertensives  other antihypertensive+1 or more regular antihypertensives  any hypertensive +nitrate  any hypertensive +other drugs for IHD and/or nitrate  other antiarrhythmics +antihypertensives |  |  |
| Anticholesterol drugs  None  Statins  other antilipids  Statins +other antilipids |  |  |
| Systemic steroids |  |  |
| Antirheumatics  none  disease modifying drugs  other antirheumatics |  |  |
| Pulmonary drugs  none  SABA only  LABA or LAMA  inhalation steroid only  SABA +Ipratopium (+/- others)  LABA +steroids  LABA +LAMA +steroids  LAMA +steroids  LABA+LAMA  other pulmonary drugs  other pulmonary drugs +steroids  SABA + LABA or LAMA -steroids  SABA +LABA or LAMA +steroids |  |  |
| Psychotropic drugs  none  SSRI/SNRI/NaRI  other antidepressants  antipsychotics  benzodiazepines3  anti-dementia drugs  anti-ADHD drugs  NaSSA  other psychotropics  SSRI +other antidepressants  SSRI +NaSSA  SRRI +antipsychotics  SRRI + other psychotropics  benzodiazepines +any other psychotropic  antipsychotics +any other psychotropic  anti-ADHD +any other psychotropic  NaSSA +any other psychotropic  other psychotropics +any listed psychotropic |  |  |
| 1Reported diabetes but no registered prescriptions 2 +/- oral antidiabetics 3 likely underreported due to limited reimbursement for benzodiazepines in Denmark | | |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Performance of the six different models with positive prediction fractions (PPF) of 20%, 25% and 30% | | | | | | | | | | |
| PPF 20 | TP | FP | FN | TN | TPR | PPV | MCC | AUROC | AUPRC | P (TPR) |
| ML\_all | 106 | 676 | 76 | 3055 | 58.2% | 13.6% | 21.1% | 76.3% | 15.5% | - |
| LR\_all | 98 | 684 | 84 | 3047 | 53.8% | 12.5% | 18.7% | 74.5% | 15.7% | 19.7% |
| ML\_10 | 100 | 682 | 82 | 3049 | 54.9% | 12.8% | 19.3% | 75.9% | 17.3% | 26.1% |
| LR\_10 | 95 | 687 | 87 | 3045 | 52.2% | 12.1% | 17.8% | 73.7% | 13.6% | 12.4% |
| ML\_NoAge | 88 | 694 | 94 | 3037 | 48.4% | 11.3% | 15.7% | 72.3% | 13.6% | 3.1% |
| Age\_only | 87 | 676 | 95 | 3055 | 47.8% | 11.4% | 15.8% | 69.7% | 12.1% | 2.3% |
| PPF 25 |  |  |  |  |  |  |  |  |  |  |
| ML\_all | 117 | 861 | 65 | 2870 | 64.3% | 12.0% | 20.0% | 76.3% | 15.5% | - |
| LR\_all | 110 | 868 | 72 | 2863 | 60.4% | 11.2% | 18.1% | 74.5% | 15.7% | 23.1% |
| ML\_10 | 115 | 863 | 67 | 2868 | 63.2% | 11.8% | 19.5% | 75.9% | 17.3% | 41.2% |
| LR\_10 | 106 | 872 | 76 | 2859 | 58.2% | 10.8% | 17.0% | 73.4% | 15.5% | 11.8% |
| ML\_NoAge | 106 | 872 | 76 | 2859 | 58.2% | 10.8% | 17.0% | 72.3% | 13.6% | 11.8% |
| Age\_only | 94 | 824 | 88 | 2907 | 51.6% | 10.2% | 14.7% | 69.7% | 12.2% | 0.7% |
| PPF 30 |  |  |  |  |  |  |  |  |  |  |
| ML\_all | 126 | 1047 | 56 | 2684 | 69.2% | 10.7% | 18.9% | 76.3% | 15.5% | - |
| LR\_all | 120 | 1053 | 62 | 2678 | 65.9% | 10.2% | 17.3% | 74.5% | 15.7% | 25.2% |
| ML\_10 | 124 | 1049 | 58 | 2682 | 68.1% | 10.6% | 18.4% | 75.9% | 17.3% | 40.8% |
| LR\_10 | 115 | 1058 | 67 | 2673 | 63.2% | 9.8% | 16.0% | 73.7% | 15.5% | 11.1% |
| ML\_NoAge | 116 | 1057 | 66 | 2674 | 63.7% | 9.9% | 16.3% | 72.3% | 13.6% | 13.8% |
| Age\_only | 100 | 955 | 82 | 2776 | 54.9% | 9.5% | 13.9% | 69.7% | 12.2% | 0.2% |
| PPF: positive prediction fraction TP: true positives FP: false positives FN: false negatives TN: true negatives TPR: true positive rate PPV: positive predictive value MCC: Matthews correlation coefficient AURC: area under the ROC curve AUPRC: area under the precision recall curve P(TPR): probability that the model performs better than the ML model relative to TPR. Green/red colors indicates the model with the best/worst performance given that specific metric | | | | | | | | | | |

# Figures

Figure 1a+b

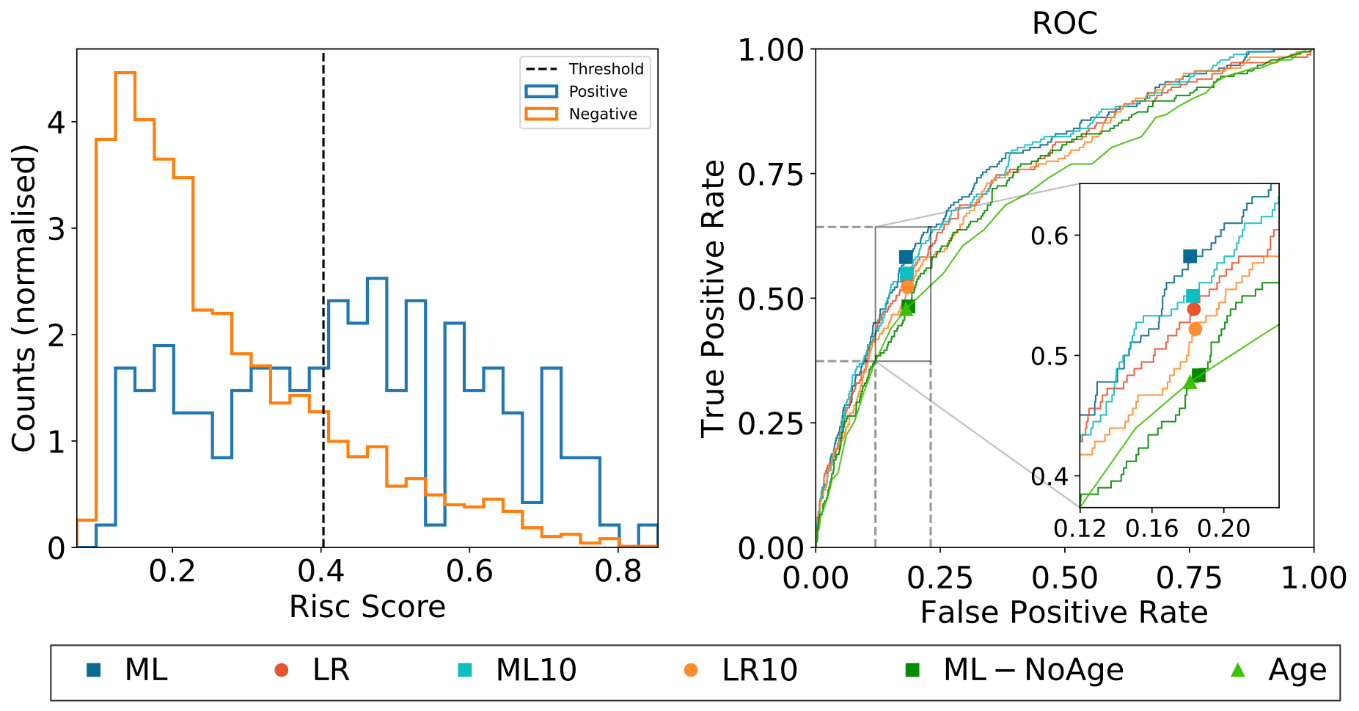


Figure 1a) Distribution of risk scores for patients +/- outcome A. The dashed lines shows the classification threshold of 20% PPF. Figure 1b) ROC curves for complete machine learning model (ML\_all), complete logistic regression model (LR\_all) , parsimonious machine learning model (ML\_10), parsimonious logistic regression model (LR\_10), machine learning using without age (ML\_NoAge) and the model age only model (Age).

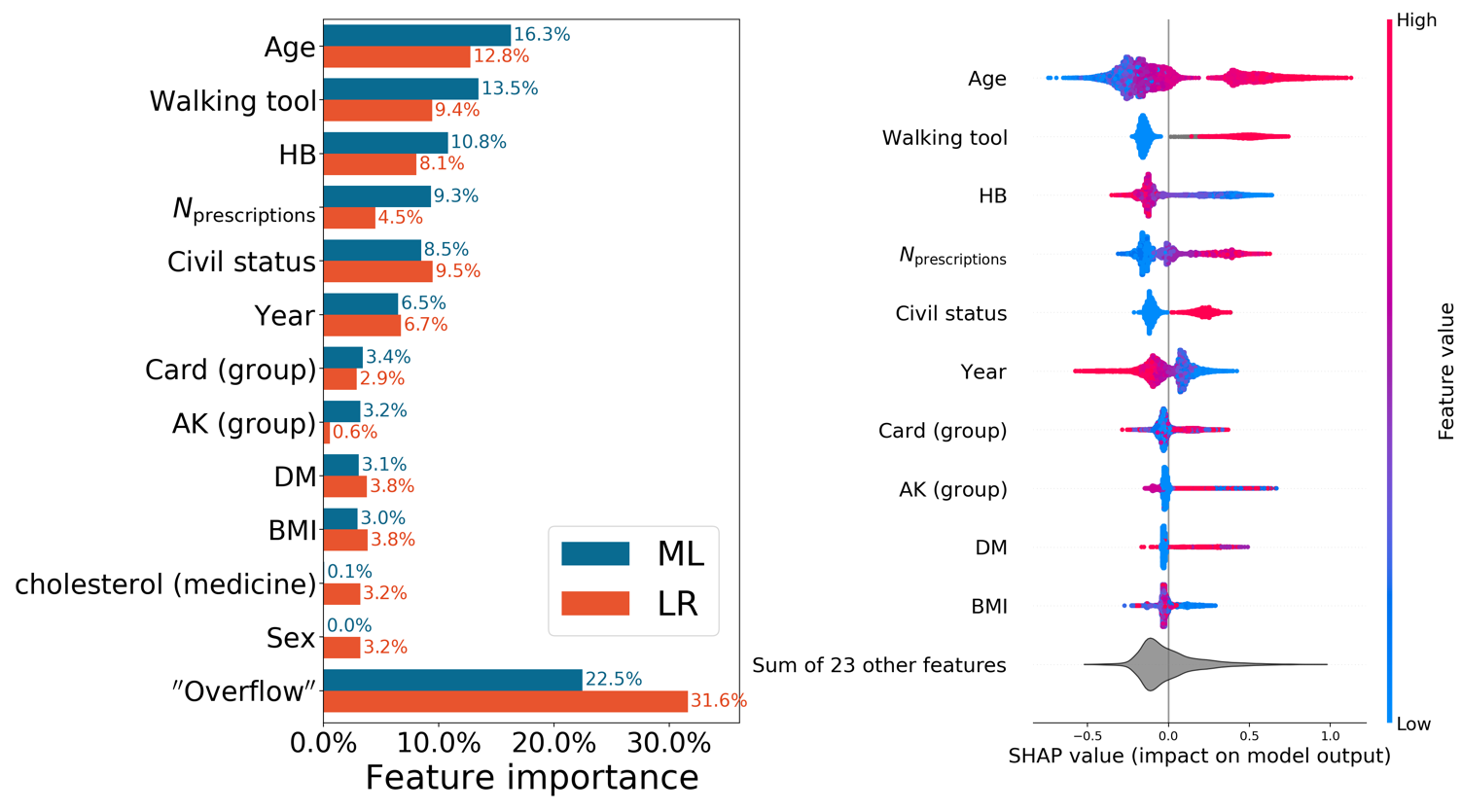
Figure 2a+b a) The overall importance of the 10 most important variables measured by the SHAP values for both the ML and LR models. The contributions of the remaining variables are summed in the “overflow” bar. b) The SHAP values for the ML model. Each individual point is a patient. The horizontal position is determined by the SHAP value and the color is related to the value of the variable. In addition to the overall importance of the different variables, this plot also shows the direction of the effect.

Figure 3a-b



SHAP beeswarm plot on the contributions individual values of a) Prescrip\_Anticoag vs. reported family disposition for venous thromboembolism (Fam.VTE) and b) Prescrip\_Cardiac vs. reported hypertension for the ML\_all model. Note Y-axis are not the same. See table 1. For explanation on group values.

# Appendix

Figure A1 SHAP beeswarm plots.



a) Prescrip\_Anticoag vs. reported family disposition for venous thromboembolism (Fam.VTE)  
b) Prescrip\_Cardiac vs. reported hypertension  
c) Prescrip\_Psych vs. reported psychiatric disorder (PsD)

d) Prescrip\_Resp vs. age.

See table 1. For explanation on group values.

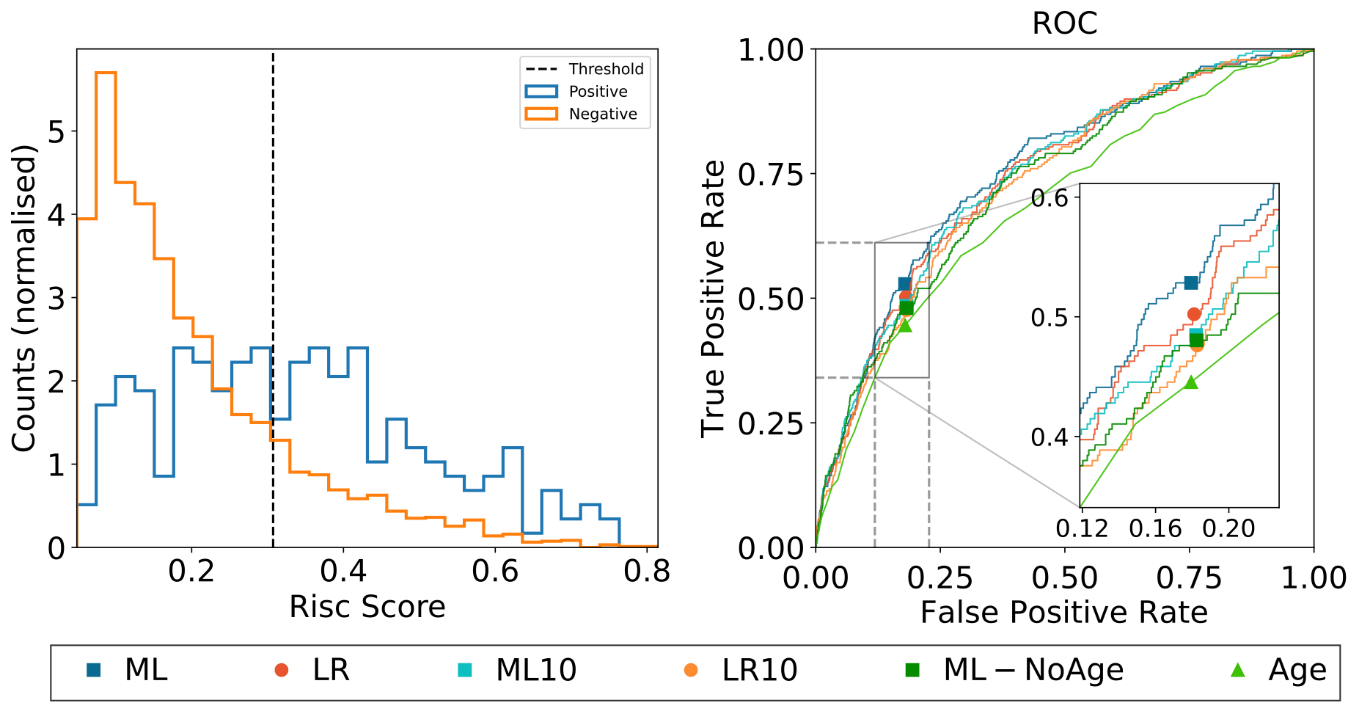
Figure A2a+b

Figure A2a) Distribution of risk scores for patients +/- outcome B. The dashed lines shows the classification threshold of 20% PPF. Figure 2Ab) ROC curves for complete machine learning model (ML\_all), complete logistic regression model (LR\_all) , parsimonious machine learning model (ML\_10), parsimonious logistic regression model (LR\_10), machine learning using without age (ML\_NoAge) and the model age only model (Age).

| Table A1 performance of different models for Outcome B | | | | | | | | | |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **TP** | **FP** | **FN** | **TN** | **TPR (%)** | **PPV (%)** | **MCC (%)** | **AUROC (%)** | **AUPRC (%)** | **P (TPR)** |
| PPF 20 | | | | | | | | | |  |
| ML\_all | 121 | 661 | 108 | 3023 | 52.8 | 15.5 | 20.5 | 75.3 | 17.1 | - |
| LR\_all | 115 | 667 | 114 | 3017 | 50.2 | 14.7 | 18.9 | 74.1 | 16.7 | 28.3% |
| ML\_10 | 111 | 671 | 118 | 3013 | 48.4 | 14.2 | 17.8 | 74.4 | 16.8 | 17.2% |
| LR\_10 | 109 | 673 | 120 | 3011 | 47.6 | 13.9 | 17.2 | 73.1 | 16.8 | 12.9% |
| ML\_NoAge | 110 | 672 | 119 | 3012 | 48.0 | 14.1 | 17.5 | 72.8 | 16.9 | 15.1% |
| Age\_only | 102 | 661 | 127 | 3023 | 44.5 | 13.4 | 15.8 | 68.7 | 13.4 | 3.8% |
| PPF 25 | | | | | | | | | |  |
| ML\_all | 140 | 838 | 89 | 2846 | 61.1 | 14.3 | 20.8 | 75.3 | 17.1 | - |
| LR\_all | 136 | 842 | 93 | 2842 | 59.4 | 13.9 | 19.8 | 74.1 | 16.7 | 35.3% |
| ML\_10 | 134 | 844 | 95 | 2840 | 58.5 | 13.7 | 19.3 | 74.4 | 16.8 | 28.3% |
| LR\_10 | 125 | 853 | 104 | 2831 | 54.6 | 12.8 | 17.0 | 73.1 | 16.8 | 7.8% |
| ML\_NoAge | 121 | 857 | 108 | 2827 | 52.8 | 12.4 | 16.0 | 72.8 | 16.9 | 3.6% |
| Age\_only | 113 | 805 | 116 | 2879 | 49.3 | 12.3 | 15.2 | 68.7 | 13.4 | 0.5% |
| PPF 30 | | | | | | | | | |  |
| ML\_all | 153 | 1020 | 76 | 2664 | 66.8 | 13.0 | 20.0 | 75.3 | 17.1 | - |
| LR\_all | 147 | 1026 | 82 | 2658 | 64.2 | 12.5 | 18.6 | 74.1 | 16.7 | 27.9% |
| ML\_10 | 147 | 1026 | 82 | 2658 | 64.2 | 12.5 | 18.6 | 74.4 | 16.8 | 27.7% |
| LR\_10 | 145 | 1028 | 84 | 2656 | 63.3 | 12.4 | 18.1 | 73.1 | 16.8 | 21.6% |
| ML\_NoAge | 140 | 1033 | 89 | 2651 | 61.1 | 11.9 | 17.0 | 72.8 | 16.9 | 10.2% |
| Age\_only | 122 | 933 | 107 | 2751 | 53.3 | 11.6 | 14.8 | 69.8 | 13.4 | 0.1% |
| PPF: positive prediction fraction TP: true positives FP: false positives FN: false negatives TN: true negatives TPR: true positive rate PPV: positive predictive value MCC: Matthews correlation coefficient AURC: area under the ROC curve AUPRC: area under the precision recall curve P(TPR): probability that the model performs better than the ML model relative to TPR. Green/red colors indicates the model with the best/worst performance given that specific metric | | | | | | | | | | |

**References**

1 Petersen PB, Kehlet H, Jørgensen CC. Improvement in fast-track hip and knee arthroplasty: a prospective multicentre study of 36,935 procedures from 2010 to 2017. *Scientific reports* 2020; **10**: 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**: 26-31

3 Savaridas T, Serrano-Pedraza I, Khan SK, Martin K, Malviya A, Reed MR. Reduced medium-term mortality following primary total hip and knee arthroplasty with an enhanced recovery program. *Acta Orthop* 2013; **84**: 40-3

4 Partridge T, Jameson S, Baker P, Deehan D, Mason J, Reed MR. 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**: 360-7

5 Jorgensen CC, Gromov K, Petersen PB, Kehlet H, Lundbeck Foundation Centre for Fast-track H, Knee Replacement Collaborative G. Influence of day of surgery and prediction of LOS > 2 days after fast-track hip and knee replacement. *Acta orthopaedica* 2021; **92**: 170-5

6 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**: e009813

7 Johns WL, Layon D, Golladay G, Kates S, Scott M, Patel NK. Preoperative Risk Factor Screening Protocols in Total Joint Arthroplasty: A Systematic Review. *J Arthroplasty* 2020

8 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**: 1466-73.e1

9 Johnson DJ, Castle JP, Hartwell MJ, D'Heurle AM, Manning DW. Risk Factors for Greater Than 24-Hour Length of Stay After Primary Total Knee Arthroplasty. *J Arthroplasty* 2020; **35**: 633-7

10 Shah A, Memon M, Kay J, Wood TJ, Tushinski DM, Khanna V. Preoperative Patient Factors Affecting Length of Stay following Total Knee Arthroplasty: A Systematic Review and Meta-Analysis. *J Arthroplasty* 2019; **34**: 2124-65.e1

11 Moonesinghe SR, Mythen MG, Das P, Rowan KM, Grocott MP. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. *Anesthesiology* 2013; **119**: 959-81

12 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

13 Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission. *Annals of surgery* 2020; **272**: 1133-9

14 Li H, Jiao J, Zhang S, Tang H, Qu X, Yue B. 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* 2020

15 Lu Y, Khazi ZM, Agarwalla A, Forsythe B, Taunton MJ. Development of a Machine Learning Algorithm to Predict Nonroutine Discharge Following Unicompartmental Knee Arthroplasty. *J Arthroplasty* 2021; **36**: 1568-76

16 Shah AA, Devana SK, Lee C, Kianian R, van der Schaar M, SooHoo NF. Development of a Novel, Potentially Universal Machine Learning Algorithm for Prediction of Complications After Total Hip Arthroplasty. *J Arthroplasty* 2021; **36**: 1655-62.e1

17 Sniderman J, Stark RB, Schwartz CE, Imam H, Finkelstein JA, Nousiainen MT. Patient Factors That Matter in Predicting Hip Arthroplasty Outcomes: A Machine-Learning Approach. *J Arthroplasty* 2021; **36**: 2024-32

18 Kugelman DN, Teo G, Huang S, Doran MG, Singh V, Long WJ. A Novel Machine Learning Predictive Tool Assessing Outpatient or Inpatient Designation for Medicare Patients Undergoing Total Hip Arthroplasty. *Arthroplasty today* 2021; **8**: 194-9

19 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**: e19761

20 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**: 632-7

21 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**: 2228-34.e1

22 Haeberle HS, Helm JM, Navarro SM, et al. Artificial Intelligence and Machine Learning in Lower Extremity Arthroplasty: A Review. *J Arthroplasty* 2019; **34**: 2201-3

23 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**: W1-73

24 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: 1-13

25 Johannesdottir SA, Horvath-Puho E, Ehrenstein V, Schmidt M, Pedersen L, Sorensen HT. Existing data sources for clinical epidemiology: The Danish National Database of Reimbursed Prescriptions. *ClinEpidemiol* 2012; **4**: 303-13

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

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

30 Chicco D, Tötsch 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**: 13

31 Lopez CD, Gazgalis A, Boddapati V, Shah RP, Cooper HJ, Geller JA. Artificial Learning and Machine Learning Decision Guidance Applications in Total Hip and Knee Arthroplasty: A Systematic Review. *Arthroplasty today* 2021; **11**: 103-12

32 Han C, Liu J, Wu Y, Chong Y, Chai X, Weng X. 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

33 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**: 2220-7.e1

34 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**: 1358-66

35 Griffiths R, Beech F, Brown A, et al. Peri-operative care of the elderly 2014: Association of Anaesthetists of Great Britain and Ireland. *Anaesthesia* 2014; **69 Suppl 1**: 81-98

36 McIsaac DI, Wong CA, Bryson GL, van Walraven C. Association of Polypharmacy with Survival, Complications, and Healthcare Resource Use after Elective Noncardiac Surgery: A Population-based Cohort Study. *Anesthesiology* 2018; **128**: 1140-50

37 Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clinical epidemiology* 2015; **7**: 449-90