# Does the SORG Orthopaedic Research Group Hip Fracture Delirium Algorithm Perform Well on an Independent Intercontinental Cohort of Patients With Hip Fractures Who Are 60 Years or Older?

Oosterhoff, Jacobien H. F.; Oberai, Tarandeep; Karhade, Aditya; Doornberg, Job N.; Kerkhoffs, Gino M. M. J.; Jaarsma, Ruurd L.; Schwab, Joseph H.; Heng, Marilyn

*Published in:*
Clinical Orthopaedics and Related Research

*DOI:*
10.1097/CORR.0000000000002246



*Publication date:*
2022

Clin Orthop Relat Res (2022) 480:2205-2213
DOI 10.1097/CORR.0000000000002246

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Clinical Research

# Does the SORG Orthopaedic Research Group Hip Fracture Delirium Algorithm Perform Well on an Independent Intercontinental Cohort of Patients With Hip Fractures Who Are 60 Years or Older?

**Jacobien H. F. Oosterhoff MD[1-3], Tarandeep Oberai PT[3], Aditya V. Karhade MD, MBA[1], Job N. Doornberg MD, PhD[4], Gino M.M.J. Kerkhoffs MD, PhD[2], Ruurd L. Jaarsma MD, PhD, FRACS, FAOrthA[3], Joseph H. Schwab MD, MS[1], Marilyn Heng MD, MPH, FRCSC[5]**

Received: 20 December 2021 / Accepted: 22 April 2022 / Published online: 10 May 2022


## Abstract

*Background* Postoperative delirium in patients aged 60 years or older with hip fractures adversely affects clinical and functional outcomes. The economic cost of delirium is estimated to be as high as USD 25,000 per patient, with a total budgetary impact between USD 6.6 to USD 82.4 billion annually in the United States alone. Forty percent of delirium episodes are preventable, and accurate risk stratification can decrease the incidence and improve clinical outcomes in patients. A previously developed clinical prediction model (the SORG Orthopaedic Research Group hip fracture delirium machine-learning algorithm) is highly accurate on internal validation (in 28,207 patients with hip fractures aged 60 years or older in a US cohort) in identifying at-risk patients, and it can

One of the authors (JHFO) certifies receipt of an amount less than USD 10,000 from ZonMW Translational Research (the Hague, the Netherlands).
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research®* editors and board members are on file with the publication and can be viewed on request.
Both databases used in this study are deidentified; the American College of Surgeons-National Surgical Quality Improvement Program database is exempt from institutional review board approval. Ethical approval was obtained from the Southern Adelaide Clinical Human Research Ethics Committee (OFR: 262.19).
This work was performed at Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

[1]Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[2]Amsterdam University Medical Centers, University of Amsterdam, Department of Orthopaedic Surgery, Amsterdam Movement Sciences, the Netherlands

[3]Department of Orthopaedic and Trauma Surgery, Flinders Medical Centre, Flinders University, Adelaide, SA, Australia

[4]Department of Orthopaedic Surgery, University Medical Centre Groningen, University of Groningen, the Netherlands

[5]Harvard Medical School Orthopedic Trauma Initiative, Massachusetts General Hospital, Boston, MA, USA

J. H. F. Oosterhoff ✉, Yawkey Building 3, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA, Email: joosterhoff@mgh.harvard.edu

Wolters Kluwer

facilitate the best use of preventive interventions; however, it has not been tested in an independent population. For an algorithm to be useful in real life, it must be valid externally, meaning that it must perform well in a patient cohort different from the cohort used to "train" it. With many promising machine-learning prediction models and many promising delirium models, only few have also been externally validated, and even fewer are international validation studies.

*Question/purpose* Does the SORG hip fracture delirium algorithm, initially trained on a database from the United States, perform well on external validation in patients aged 60 years or older in Australia and New Zealand?

*Methods* We previously developed a model in 2021 for assessing risk of delirium in hip fracture patients using records of 28,207 patients obtained from the American College of Surgeons National Surgical Quality Improvement Program. Variables included in the original model included age, American Society of Anesthesiologists (ASA) class, functional status (independent or partially or totally dependent for any activities of daily living), preoperative dementia, preoperative delirium, and preoperative need for a mobility aid. To assess whether this model could be applied elsewhere, we used records from an international hip fracture registry. Between June 2017 and December 2018, 6672 patients older than 60 years of age in Australia and New Zealand were treated surgically for a femoral neck, intertrochanteric hip, or subtrochanteric hip fracture and entered into the Australian & New Zealand Hip Fracture Registry. Patients were excluded if they had a pathological hip fracture or septic shock. Of all patients, 6% (402 of 6672) did not meet the inclusion criteria, leaving 94% (6270 of 6672) of patients available for inclusion in this retrospective analysis. Seventy-one percent (4249 of 5986) of patients were aged 80 years or older, after accounting for 5% (284 of 6270) of missing values; 68% (4292 of 6266) were female, after accounting for 0.06% (4 of 6270) of missing values, and 83% (4690 of 5661) of patients were classified as ASA III/IV, after accounting for 10% (609 of 6270) of missing values. Missing data were imputed using the missForest methodology. In total, 39% (2467 of 6270) of patients developed postoperative delirium. The performance of the SORG hip fracture delirium algorithm on the validation cohort was assessed by discrimination, calibration, Brier score, and a decision curve analysis. Discrimination, known as the area under the receiver operating characteristic curves (c-statistic), measures the model's ability to distinguish patients who achieved the outcomes from those who did not and ranges from 0.5 to 1.0, with 1.0 indicating the highest discrimination score and 0.50 the lowest. Calibration plots the predicted versus the observed probabilities, a perfect plot has an intercept of 0 and a slope of 1. The Brier score calculates a composite of discrimination and calibration, with 0 indicating perfect prediction and 1 the poorest.

*Results* The SORG hip fracture algorithm, when applied to an external patient cohort, distinguished between patients at low risk and patients at moderate to high risk of developing postoperative delirium. The SORG hip fracture algorithm performed with a c-statistic of 0.74 (95% confidence interval 0.73 to 0.76). The calibration plot showed high accuracy in the lower predicted probabilities (intercept -0.28, slope 0.52) and a Brier score of 0.22 (the null model Brier score was 0.24). The decision curve analysis showed that the model can be beneficial compared with no model or compared with characterizing all patients as at risk for developing delirium.

*Conclusion* Algorithms developed with machine learning are a potential tool for refining treatment of at-risk patients. If high-risk patients can be reliably identified, resources can be appropriately directed toward their care. Although the current iteration of SORG should not be relied on for patient care, it suggests potential utility in assessing risk. Further assessment in different populations, made easier by international collaborations and standardization of registries, would be useful in the development of universally valid prediction models. The model can be freely accessed at: https://sorg-apps.shinyapps.io/hipfxdelirium/.

*Level of Evidence* Level III, therapeutic study.

## Introduction

Hip fractures are one of the most serious and costly fall-related injuries experienced by people, most of whom are treated operatively [10, 14]. The number of hip fractures continues to rise worldwide and is predicted to rise to an incidence of 6.26 million fractures annually in 2050 [1]. Delirium is the most common complication in patients with hip fractures, occurring in 28% to 50% [5], and it is characterized by an acute and fluctuating course, inattention, altered level of consciousness, and evidence of disorganized thinking [28]. Although potentially reversible and by definition transient, delirium is one of the most frequent reasons for a patient referral to a geriatrician [26]. A patient with delirium may be disoriented to place and time, may not understand the severity of the injury, and may not adhere to therapy. This will lead to a longer in-hospital stay, higher risk of complications, and higher economic costs. Substantial additional costs occur after surgery because of the longer in-hospital stay, increased hospitalization, and rehabilitation after discharge [15]. According to estimates, the healthcare costs attributable to postoperative delirium can be as high as USD 25,000 per patient, with a total budgetary impact between USD 6.6 to USD 82.4 billion annually in the United States alone [12, 23]. Forty percent of delirium episodes are preventable, and accurate risk

stratification can decrease the incidence, improve clinical outcomes in patients, and reduce economic costs [15].

Many delirium prevention strategies have been described, with accurate (internally validated) tools in the intensive care unit population [6] and in the hip fracture population [22, 46]. However, only a few of the promising prediction models have been externally validated—a necessary step before clinical implementation [42]—with few external validation studies specific for the hip fracture population [11, 29]. External validation is required to assess the performance of the clinical prediction model and validate the promise in an independent population with similar injury and patient characteristics to confirm that the model is generalizable. Recently, a clinical prediction model using machine-learning algorithms was developed, showing promise in estimating the risk of postoperative delirium in 28,207 hip fracture patients aged 60 years or older in a North American cohort [31]. This clinical prediction model is available in a freely available internet application at https://sorg-apps.shinyapps.io/hipfxdelirium/. However, while many promising machine-learning prediction models have been developed in orthopaedic surgery, only few have also been externally validated, and even fewer are international validation studies [13]. International collaborations and standardization of international registries may allow for universally valid prediction models, which is the next step for moving prediction modeling from a single-country task to a coordinated global effort [17].

Therefore, we asked: Does the SORG hip fracture delirium algorithm, initially trained on a database from the United States, perform well on external validation in patients aged 60 years or older in Australia and New Zealand?

## Patients and Methods

This study followed the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis Guideline [7] and the Strengthening the Reporting of Observational Studies in Epidemiology [44] guidelines.

### *Study Design and Setting*

We developed a model in 2021 for assessing delirium risk in patients with hip fractures, using the records of 28,207 patients obtained from the American College of Surgeons National Surgical Quality Improvement Program. In this developmental cohort of 28,207 patients, 28% (8030) developed a postoperative delirium [31].

The clinical prediction model reached good discrimination (c-statistic = 0.79 [95% CI 0.77 to 0.80]), almost perfect calibration (intercept = -0.01, slope = 1.02), and excellent overall model performance (Brier score = 0.15). The following variables were included in the primary developed clinical prediction model: age, American Society of Anesthesiologists (ASA) classification, functional status (independent or partially or totally dependent for any activities of daily living), preoperative dementia, preoperative delirium, and preoperative need for a mobility aid. Further details of the original clinical prediction model can be found in the developmental study's report [31].

To assess whether this model could be applied elsewhere, we used records from an international hip fracture registry. The validation cohort originated from the Australian and New Zealand Hip Fracture Registry (ANZHFR), which was queried from June 2017 to December 2018. The ANZHFR is a prospective, multiinstitution database that collects preoperative, perioperative, and postoperative data on more than 50 independent variables from more than 67 participating hospitals in Australia and New Zealand. Data are acquired from medical records and operative notes. The data items collected by the ANZHFR are specified in the Australian and New Zealand Hip Fracture Registry Data Dictionary version 12.1_October 2019. A range of validated diagnostic tools for delirium have been deemed acceptable in the ANZHFR, including confusion assessment method (CAM), the CAM for the Intensive Care Unit (CAM-ICU), the 3-Minute CAM (3D-CAM), and the 4 As test (4AT) [3, 9, 16]. The selection criteria used in the developmental study [31] were applied; we included patients older than 60 years who underwent operative fixation of a femoral neck, intertrochanteric hip, or subtrochanteric hip fracture. Patients were excluded if they sustained a pathologic hip fracture or septic shock. The primary outcome of interest was postoperative delirium after surgical treatment of a hip fracture.

### *Participants' Baseline Characteristics*

Between June 2017 to December 2018, 6672 patients older than 60 years of age in Australia and New Zealand were treated surgically for a femoral neck, intertrochanteric hip, or subtrochanteric hip fracture and entered into the ANZHFR. Patients were excluded if they had sustained a pathological hip fracture or developed septic shock. Of all patients, 6% (402 of 6672) did not meet the inclusion criteria, leaving 94% (6270 of 6672) available for inclusion in this retrospective analysis, of whom 39% (2467 of 6270) had postoperative delirium. Seventy-one percent (4249 of 5986) of patients were aged 80 years or older, 68% were female (4292 of 6266), and 83% of patients were classified as ASA III/IV (4690 of 5661) (Table 1).

Baseline characteristics in the validation cohort differed from those in the original developmental cohort [31] in several regards (Table 1). The cohort from Australia and New Zealand were more likely to be older, men, and healthier (as evidenced by a lower ASA score). However, they were less

**Table 1.** Baseline characteristics of the developmental and validation cohorts

| Variable | Developmental cohort (n = 28,207) | Validation cohort (n = 6270)[a] | p value |
|---|---|---|---|
| Age in years | | | < 0.001 |
| 60 + | 11 (3151) | 6 (340) | |
| 70 + | 22 (6247) | 23 (1397) | |
| 80 + | 41 (11,691) | 47 (2838) | |
| 90 + | 25 (7118) | 24 (1411) | |
| Women | 70 (19,845) | 68 (4292) | < 0.01 |
| ASA class | | | < 0.001 |
| I | 0.4 (126) | 1 (60) | |
| II | 15 (4162) | 15 (848) | |
| III | 63 (17,631) | 60 (3373) | |
| IV | 22 (6288) | 23 (1317) | |
| Preoperative functional status | | | < 0.001 |
| Independent | 77 (21,672) | 71 (4389) | |
| Partially or totally dependent | 23 (6535) | 30 (1840) | |
| Preoperative need for mobility aid | 58 (16,239) | 57 (3527) | 0.55 |
| Preoperative delirium | 13 (3714) | 40 (1406) | < 0.001 |
| Preoperative dementia | 31 (8668) | 43 (2597) | < 0.001 |
| Preoperative bone protective medication | 32 (9047) | 37 (2317) | < 0.001 |
| Medical comanagement by geriatric medicine | 89 (25,136) | 97 (6086) | < 0.001 |
| Postoperative delirium | 28 (8030) | 39 (2467) | < 0.001 |

Data presented as % (n).

[a]Proportions were calculated accounting for the following missing values: age (5% [284 of 6270]), gender (0.06% [4 of 6270]), ASA class (10% [609 of 6270]), functional status (0.2% [13 of 6270]), preoperative need for a mobility aid (1% [86 of 6270]), preoperative delirium (44% [2767 of 6270]), and preoperative dementia (3% [187 of 6270]).

likely to live independently, and more likely to experience preoperative delirium and preoperative dementia. They were more likely to use bone protective medication (calcium and/or vitamin D only AND/OR bisphosphonates, denosumab, or teriparatide) prior to injury, and their hospital care was more likely to include medical co-management by a geriatrician or specialized nurse conducting preoperative medical assessment (in addition to an anesthetic review and orthopaedic assessment) (all $p < 0.05$). The proportion of postoperative delirium was higher in the validation cohort (39% [2467 of 6270]) than in the developmental cohort (29% [8030 of 28,207]; $p < 0.05$).

### *Missing Data*

Preprocessing of the validation cohort was performed by imputing missing values, using the missForest methodology [35] as previously applied by our group [4, 18–20, 39]. We imputed missing values for age (5% [284 of 6270]), gender (0.06% [4 of 6270]), American Society of Anesthesiologists (ASA) class (10% [609 of 6270]), functional status (0.2% [13 of 6270]), preoperative need for a mobility aid (1% [86 of 6270]), preoperative delirium (44% [2767 of 6270]), and preoperative dementia (3% [187 of 6270]). In addition, a complete case analysis was carried out to evaluate the effect when a variable has > 30% missing data [33].

### *Ethical Approval*

The study was approved by the Southern Adelaide Clinical Human Research Ethics Committee (OFR: 262.19).

### *Assessment of Model Performance and Statistical Analysis*

Model performance was evaluated according to a proposed framework for the evaluation of a clinical prediction model [38] that includes discrimination with the c-statistic,

calibration with a calibration slope and intercept, and the overall performance, assessed with the Brier score.

The c-statistic (area under the curve of a receiver operating characteristic curve) ranges from 0.50 to 1.0, with 1.0 indicating the highest discrimination score and 0.50 indicating the lowest. The receiver operating curve (ROC) plots the false positive rate (x-axis) and true positive rate (y-axis). In risk stratification, ideally there is a high true positive rate and a low false positive rate. The higher the discrimination score, the better the model's ability to distinguish between patients with the outcome and those who did not have the outcome [37]. In general, we used the following rule, depending on the context: a c-statistic of 0.5 suggests no discrimination (that is, the ability to predict patients with and without a postoperative delirium based on the model), 0.6 to 0.7 was considered poor, 0.7 to 0.8 was considered acceptable, 0.8 to 0.9 was considered excellent, and more than 0.9 was considered outstanding [27].

A calibration plot charts the predicted (x-axis) versus the true observed probabilities (y-axis, labeled outcomes) for the primary outcome. The concept is to evaluate the average predicted probability that corresponds with the true predicted probability for binned predictions (that is, a probability of 0.80 to 0.89 is one bin) and gives a certain confidence on the prediction (or the reliability of the algorithm) [32]. A perfect calibration plot has an intercept of 0 (< 0 reflects overestimation and > 0 reflects underestimating the probability of the outcome) and a slope of 1 (model is performing similarly in training and test sets) [38, 40]. In a small dataset, the slope is often < 1, reflecting model overfitting; probabilities are too extreme (low probability too low; high probability too high) [37].

The Brier score calculates a composite of discrimination and calibration, with 0 indicating perfect prediction and a Brier score of 1 representing the poorest prediction. The null-model Brier score (a score that equals the probability of delirium in the dataset) was used to benchmark the algorithm's Brier score. A Brier score lower than the null-model Brier score indicates superior performance of the model to this null benchmark. Perfect models would have a Brier score of 0 [38].

In addition, we undertook a decision curve analysis to investigate the net benefit (weighted average of true positives and false positives, formula = sensitivity x prevalence – (1 – specificity) x (1 – prevalence) x odds at the threshold probability) of the conducted algorithms over the range of risk thresholds for clinical decision-making [43]. With threshold probability we refer to the probability that an algorithm ranks a positive outcome over a negative outcome. If the threshold is set at 0.5, then patients with a probability > 0.5 are classified as positive and < 0.5 are classified as negative. If the threshold is set at 0.8, then patients with a probability > 0.8 are classified as positive and < 0.8 are classified as negative. The decision curve of the model is compared with decision curves of treating everyone as being at risk for postoperative delirium and treating no one as being at risk for postoperative delirium.

Baseline characteristics are presented as percentages and frequencies for dichotomous and categorical variables and median with interquartile range for continuous variables. Baseline characteristics in the developmental and validation cohort were compared using a bivariate analysis, where a p value of < 0.05 was considered significant. Data preprocessing and analysis were performed using R Version 4.0 ("R: A Language and Environment for Statistical Computing" The R Foundation) and R-studio Version 1.2.1335 (R-Studio).

### *Internet Application*

This clinical prediction model is available in a freely available internet application at https://sorg-apps.shinyapps.io/hipfxdelirium/.

## Results

### *External Validation of SORG Hip Fracture Delirium Algorithm in Australia and New Zealand*

The SORG hip fracture delirium algorithm achieved good discrimination in predicting postoperative delirium in hip fracture patients aged 60 years or older in the Australian and New Zealand cohorts. The c-statistic was 0.74 (95% confidence interval [CI] 0.73 to 0.76) (Table 2) and the ROC curve shows the graph of the model performance by plotting the false positive and true negative rates with an area under the curve (AUC) corresponding to the c-statistic with 0.74 (Fig. 1). The calibration plot of the algorithm in the validation cohort showed calibration metrics with an intercept of -0.28 (95% CI -0.35 to -0.21) and a calibration slope of 0.52 (95% CI 0.49 to 0.56) (Fig. 2). The calibration plot was highly accurate in the range of lower predicted probabilities. The Brier score was lower than the respective null-model Brier score (0.22 versus 0.24), indicating good overall performance of the SORG hip fracture delirium algorithm. According to the decision curve analysis, the SORG hip fracture delirium algorithm provided a positive net benefit compared with a strategy of treating all patients or no patients as being at risk of postoperative delirium (Fig. 3). The net benefit can be interpreted as reflecting the balance between a true positive prediction and the harm of a false positive prediction. Seeing no patients as being at risk is always 0 because the model will not predict anyone as being positive. Seeing all patients as being at risk of postoperative delirium will cross y = 0 at the prevalence of the validation cohort (39% in our study) [41]. A risk threshold can be interpreted as follows: with a risk threshold of 20% (1 to 5), each false positive should be weighed by the odds of 5 (the harm-to-benefit ratio). A

model is only clinically useful if the net benefit at a certain risk threshold T is higher than treat all or treat no patients. However, there is no single risk threshold that is universally acceptable, and the choice of a clinically appropriate threshold should not depend on the result of a decision curve analysis [21].

## Discussion

Patients aged 60 years or older undergoing hip fracture surgery have a high risk of developing postoperative delirium, leading to higher complications, longer in-hospital stays, and increased economic costs. Many delirium-preventive strategies exist, including prediction models that assess delirium risk. However, only a few delirium prediction models have been validated in an independent cohort, a necessary step before clinical implementation, and even fewer tools are externally validated specific for the hip fracture population. Previously, we developed a clinical prediction model (SORG hip fracture delirium algorithm) in a large North American cohort, and the purpose of this study was to externally validate the prediction model in an independent cohort. On external validation, the prediction model retained good discriminative ability and was shown to be accurate in distinguishing between low-risk patients (< 25%) and moderate to high-risk patients (> 25%) to make preventive interventions a priority. The internet-based tool suggests potential utility over treating everyone as being at risk.

**Table 2.** Model performance assessment on external validation in the Australian-New Zealand database (n = 6270)

| Metric | Elastic-net penalized logistic regression |
|---|---|
| c-statistic[a] | 0.74 (95% CI 0.73 to 0.76) |
| Intercept[b] | -0.28 (95% CI -0.35 to -0.21) |
| Slope[b] | 0.52 (95% CI 0.49 to 0.56) |
| Brier[c] | 0.22 (95% CI 0.21 to 0.23) |

Null-model Brier score = 0.24.
[a]A c-statistic of 0.5 indicates random guess and 1.0 indicates perfect discriminatory ability; a c-index of 0.7 to 0.8 is typically considered acceptable discriminatory ability.
[b]Calibration plots the predicted versus the observed probabilities; a perfect calibration plot has an intercept of 0 (< 0 reflects overestimation and > 0 reflects underestimation of the probability of the outcome) and a slope of 1 (model is performing similarly in training and test sets); if the slope is < 1 (often in small datasets), this reflects model overfitting; probabilities are too extreme (low probability too low; high probability too high).
[c]The Brier score of the prediction model should be compared with that of the null model; the null-model Brier score is a score calculated from the probability of delirium in the dataset and used to benchmark the algorithm's Brier score; a lower Brier score of the prediction model indicates good overall model performance.

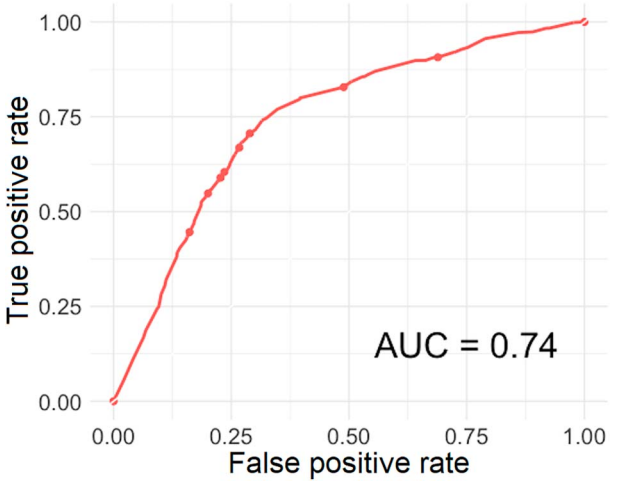


**Fig. 1** This graph shows discrimination of the elastic-net penalized logistic regression model on external validation in the ANZHFR (n = 6270); AUC = area under the receiver operating characteristic curve.

### *Limitations*

The results of this study should be viewed considering several limitations. First, although machine learning can work well at deriving associations and correlations, it cannot determine causation or assess whether those associations make physiologic sense. Second, as with any algorithm, the quality of machine learning is highly dependent on data quality; if

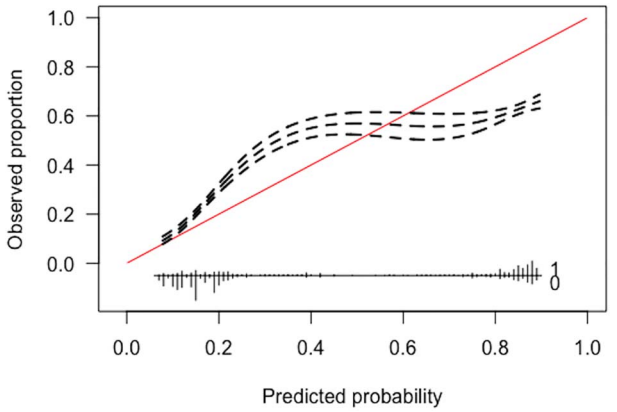


**Fig. 2** This graph shows the calibration of the elastic-net penalized logistic regression model on external validation in the ANZHFR (n = 6270).

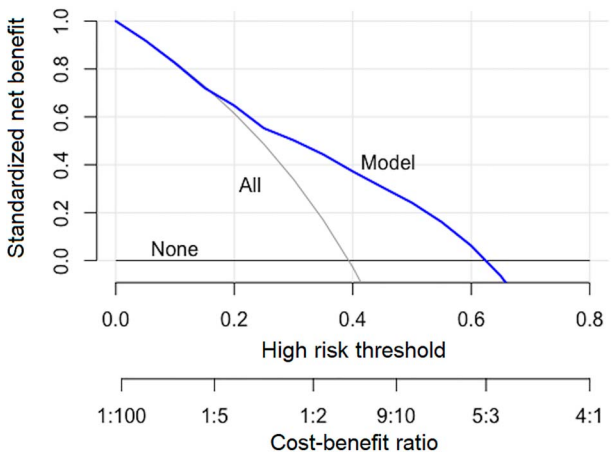


**Fig. 3** This graph shows a decision curve analysis for the penalized logistic regression for predicting postoperative delirium in the ANZHFR (n = 6270). The decision curve analysis represents the net benefit achieved by management changes based on the penalized logistic regression algorithm relative to default strategies. The net benefit is shown by the graph of the model following a higher and longer graph than the graphs of the default strategies that see all patients as being at risk for postoperative delirium or treating no patients as being at risk.

available data are poor, subjective, or incomplete, no algorithm can be expected to work well. Here, data on preoperative delirium were missing for almost half of the patient group, and we cannot be sure that our algorithm would not have performed differently if these data had been available. Further, generalizability of a prediction model cannot be assessed after a single external validation study, but it should be examined after thorough independent external validation for each population if the population differs considerably in setting, in patient demographics, or outcome incidence. This was a validation study in an Australian and New Zealand cohort. This might limit the reference value for other countries, for patients from other racially distinct regions, or patients with different background in terms of social determinants of health (such as, socioeconomic status, income level, or education). In addition, statistical models using machine learning are hypothesized to have the potential to provide more accurate estimates for the prediction of binary events compared with more traditional logistic regression algorithms. Our prediction model uses a penalized logistic regression algorithm, which is basically a logistic regression algorithm with more flexibility in the hyperparameters. This finding is in line with previous research, which has shown that the benefit of more complex machine-learning methods may be limited in this context for the prediction of binary outcome in orthopaedic trauma [30]. Moreover, the study designs of the development and validation cohort were country-wide registries, meaning the data were collected for quality outcome purposes rather than research. However, researchers can gain data-driven insights from these registry-based patient cohorts to better understand expected outcomes. Predictive analytics on registry-based data may play a significant role in the future with advances in computation to improve the prediction model's accuracy when, for example, combined with medical imaging or free-text notes leading to artificial intelligence–based registries [34]. In addition, the cohorts originated from different continents, which could lead to variation in treatment protocols and diversity in training programs for orthopaedic surgeons between countries. A previous study assessing a cross-cultural comparison of treatment outcomes in hip fracture patients found that although there were possible differences in clinical practices in two different countries, that did not influence the clinical outcomes [24], and we did not expect the differences from our cohort to influence treatment outcomes. Furthermore, the variable definition differed between both cohorts, including the assessment of postoperative delirium, which was defined as occurring within 7 days of surgery in the validation cohort compared with 30 days in the developmental cohort. Because the start of delirium is usually rapid (appearing within hours [2] and peaking between 1 and 3 days postoperatively [45]), we assumed all postoperative delirium events were captured within the 7-day period, and that we did not miss cases of postoperative delirium (Supplementary Table 1; http://links.lww.com/CORR/A803). Lastly, a high proportion of missing values was seen in our assessment of preoperative delirium. Therefore, we performed a complete case analysis, and the results were in line with model performance metrics for the total validation cohort with a c-statistic of 0.75 (95% CI 0.74 to 0.77) (Supplementary Table 2; http://links.lww.com/CORR/A804), comparable ROC curve (Supplementary Fig. 1; http://links.lww.com/CORR/A805), calibration plot (Supplementary Fig. 2; http://links.lww.com/CORR/A806), and decision curve analysis (Supplementary Fig. 3; http://links.lww.com/CORR/A807).

### *Discussion of Key Findings*

We found that the SORG hip fracture delirium algorithm, initially trained on a dataset from North America, performed equally well on a dataset from Australia and New Zealand.

However, in its current iteration, we did not find that the SORG hip fracture delirium algorithm performed better than other existing and validated instruments for assessing postoperative delirium risk. The current study is an external validation of a single prediction model, although many successful delirium prediction models have been described [6, 25]. Our study emphasizes the importance of externally validating a well-developed algorithm in an independent cohort, with similar patient and injury characteristics (patients with hip fractures who were 60 years or older). We believe international validation studies with transparent

reporting is an important step for moving prediction modeling from a single-country to a coordinated global effort [13]. More than 15 delirium prediction models are reported in the evidence [6], and only two studies externally validated a delirium prediction model specific to the hip fracture population [11, 29]. One of these two studies externally validated the Risk Model for Delirium score and reported a c-statistic of 0.73 (95% CI 0.68 to 77) but did not report calibration, Brier scores, or decision curve metrics, which is recommended in evaluating prediction models [29, 36]. Another study assessed the performance of the Delirium Elderly at Risk in hip fracture patients, reporting a positive predictive value ranging between 54% to 65% (that a positive prediction turns out to be a postoperative delirium) and a negative predictive value ranging between 76% to 90%. Discrimination, calibration, Brier scores, and decision curves were not reported [11].

The model in the current specific population has been shown to be highly accurate for distinguishing between low-risk patients (< 25%) and moderate to high-risk patients (> 25%). We recommend preventive measures be made a priority in patients who have a more than 25% probability of developing postoperative delirium after hip fracture surgery. Delirium is common, costly, and associated with complications; however, effective, multidisciplinary strategies can prevent it. Interventions in hospitalized older adults include regular orientation, therapeutic activities, frequent mobilization and exercise, and avoidance of psychoactive medications in favor of nonpharmacologic approaches for anxiety and sleep [8]. The prediction model should not be used as a standalone tool, and it does not replace clinical judgment nor screening measures. The prediction model may support assigning patients to a delirium prevention program when delirium prevention strategies are not standard practice, especially in smaller, nonacademic hospital and rural areas.

### *Conclusion*

Algorithms developed with machine learning are a potential tool for refining treatment of at-risk patients. If high-risk patients can be reliably identified, resources can be appropriately directed toward their care. Although the current iteration of SORG should not be relied on for patient care, it suggests potential utility in assessing risk. However, the current machine-learning algorithm did not perform any better than other existing and validated instruments for assessing postoperative delirium risk. Further assessment in different populations, made easier by international collaborations and standardization of registries, would be useful in the development of universally valid prediction models. The model can be freely accessed at: https://sorg-apps.shinyapps.io/hipfxdelirium/.

## References


1. American Academy Orthopaedic Surgeons. Management of hip fractures in the elderly: evidence-based clinical practice guideline. Available at: https://www.aaos.org/globalassets/quality-and-practice-resources/hip-fractures-in-the-elderly/management_of_hip_fractures_in_the_elderly-7-24-19.pdf. Accessed June 26, 2020.
2. Bellelli G, Mazzola P, Morandi A, et al. Duration of postoperative delirium is an independent predictor of 6-month mortality in older adults after hip fracture. *J Am Geriatr Soc.* 2014;62:1335-1340.
3. Bellelli G, Morandi A, Davis DHJ, et al. Validation of the 4AT, a new instrument for rapid delirium screening: a study in 234 hospitalised older people. *Age Ageing*. 2014;43:496-502.
4. Bongers MER, Thio QCBS, Karhade AV, et al. Does the SORG algorithm predict 5-year survival in patients with chondrosarcoma? An external validation. *Clin Orthop Relat Res.* 2019;477:2296-2303.
5. Brauer C, Morrison RS, Silberzweig SB, et al. The cause of delirium in patients with hip fracture. *Arch Intern Med.* 2000;160:1856-1860.
6. Chen X, Lao Y, Zhang Y, et al. Risk predictive models for delirium in the intensive care unit: a systematic review and meta-analysis. *Ann Palliat Med.* 2021;10:1467.
7. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.
8. Deeken F, Sánchez A, Rapp MA, et al. Outcomes of a delirium prevention program in older persons after elective surgery: a stepped-wedge cluster randomized clinical trial. *JAMA Surg.* 2022;157:e216370-e216370.
9. Ely EW, Margolin R, Francis J, et al. Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). *Crit Care Med.* 2001;29:1370-1379.
10. Fixation using Alternative Implants for the Treatment of Hip fractures (FAITH) Investigators. Fracture fixation in the operative management of hip fractures (FAITH): an international, multicentre, randomised controlled trial. *Lancet*. 2017;389:1519-1527.
11. Freter S, Dunbar M, Koller K, et al. Risk of pre- and postoperative delirium and the Delirium Elderly At Risk (DEAR) tool in hip fracture patients. *Can Geriatr J.* 2015;18:212-216.
12. Gou RY, Hshieh TT, Marcantonio ER, et al. One-year Medicare costs associated with delirium in older patients undergoing major elective surgery. *JAMA Surg.* 2021;156:430-442.
13. Groot OQ, BJJ Bindels, Ogink PT, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop.* 2021;92:385-393.
14. HEALTH Investigators, Bhandari M, Einhorn T, et al. Total hip arthroplasty or hemiarthroplasty for hip fracture. *N Engl J Med.* 2019;381:2199-2208.
15. Inouye SK, Bogardus ST, Charpentier PA, et al. A multicomponent intervention to prevent delirium in hospitalized older patients. *N Engl J Med.* 1999;340:669-676.
16. Inouye SK, Charpentier PA. Precipitating factors for delirium in hospitalized elderly persons: predictive model and interrelationship with baseline vulnerability. *JAMA*. 1996;275:852-857.
17. Johansen A, Golding D, Brent L, et al. Using national hip fracture registries and audit databases to develop an international perspective. *Injury*. 2017;48:2174-2179.

18. Karhade AV, Thio QCBS, Ogink PT, et al. Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation. *Clin Neurosurg.* 2019;85:E671-E681.
19. Karhade AV, Thio QCBS, Ogink PT, et al. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. *Clin Neurosurg.* 2019;85: E83-E91.
20. Karhade A V, Ogink PT, Thio QCBS, et al. Development of machine learning algorithms for prediction of prolonged opioid prescription after surgery for lumbar disc herniation. *Spine J.* 2019;19:1764-1771.
21. Kerr KF, Brown MD, Zhu K, et al. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *J Clin Oncol.* 2016;34:2534-2540.
22. Kim EM, Li G, Kim M. Development of a risk score to predict postoperative delirium in patients with hip fracture. *Anesth Analg.* 2020;130:79-86.
23. Kinchin I, Mitchell E, Agar M, et al. The economic cost of delirium: a systematic review and quality assessment. *Alzheimers Dement.* 2021;17:1026-1041.
24. Kusen JQ, van der Vet PCR, Wijdicks FJG, et al. Efficacy of two integrated geriatric care pathways for the treatment of hip fractures: a cross-cultural comparison. *Eur J Trauma Emerg Surg*. Published online March 10, 2021. DOI: 10.1007/s00068-021-01626-y.
25. Lindroth H, Bratzke L, Purvis S, et al. Systematic review of prediction models for delirium in the older adult inpatient. *BMJ Open*. 2018;8:e019223.
26. Lipowski ZJ. Transient cognitive disorders (delirium, acute confusional states) in the elderly. *Am J Psychiatry*. 1983;140: 1426-1436.
27. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol.* 2010;5:1315-1316.
28. Marcantonio ER. In the clinic. Delirium. *Ann Intern Med.* 2011; 154:6-16.
29. Moerman S, Tuinebreijer WE, de Boo M, et al. Validation of the risk model for delirium in hip fracture patients. *Gen Hosp Psychiatry*. 2012;34:153-159.
30. Oosterhoff JHF, Gravesteijn BY, Karhade AV, et al. Feasibility of machine learning and logistic regression algorithms to predict outcome in orthopaedic trauma surgery. *J Bone Joint Surg Am.* 2022;104:544-551.
31. Oosterhoff JHF, Karhade AV, Oberai T, et al. Prediction of postoperative delirium in geriatric hip fracture patients: a clinical prediction model using machine learning algorithms. *Geriatr Orthop Surg Rehabil.* Published online December 12, 2021. DOI: 10.1177/21514593211062277.
32. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12: 2825-2830.
33. Pigott TD. A review of methods for missing data. *Educ Res Eval.* 2001;7:353-383.
34. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med.* 2022;28:31-38.
35. Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28: 112-118.
36. Steyerberg E. Validation of prediction models. In: Gail M, Samet J, Singer B, eds. *Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating.* 2nd ed. Springer; 2019:309-323.
37. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35:1925-1931.
38. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128-138.
39. Thio QCBS, Karhade AV, Ogink PT, et al. Development and internal validation of machine learning algorithms for preoperative survival prediction of extremity metastatic disease. *Clin Orthop Relat Res.* 2020;478:322-333.
40. van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making*. 2015;35:162-169.
41. van Calster B, Wynants L, Verbeek JFM, et al. Reporting and interpreting decision curve analysis: a guide for investigators. *Eur Urol.* 2018;74:796-804.
42. van Meenen LCC, van Meenen DMP, de Rooij SE, et al. Risk prediction models for postoperative delirium: a systematic review and meta-analysis. *J Am Geriatr Soc.* 2014;62:2383-2390.
43. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26: 565-574.
44. von Elm E, Altman D, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370:1453-1457.
45. Whitlock EL, Vannucci A, Avidan MS. Postoperative delirium. *Minerva Anestesiol.* 2011;77:448-456.
46. Zhao H, You J, Peng Y, et al. Machine learning algorithm using electronic chart-derived data to predict delirium after elderly hip fracture surgeries: a retrospective case-control study. *Front Surg.* 2021;8:634629.