BJO

HIP

Development of machine-learning algorithms for 90-day and one-year mortality prediction in the elderly with femoral neck fractures based on the HEALTH and FAITH trials

H. Dijkstra,
J. H. F. Oosterhoff,
A. van de Kuit,
F. F. A. IJpma,
J. H. Schwab,
R. W. Poolman,
S. Sprague,
S. Bzovsky,
M. Bhandari,
M. Swiontkowski,
E. H. Schemitsch,
J. N. Doornberg,
L. A. M. Hendrickx,
On behalf of the Machine Learning Consortium,
the HEALTH Investigators,
and the FAITH Investigators

From University Medical Centre Groningen, Groningen, the Netherlands

Aims

To develop prediction models using machine-learning (ML) algorithms for 90-day and one-year mortality prediction in femoral neck fracture (FNF) patients aged 50 years or older based on the Hip fracture Evaluation with Alternatives of Total Hip arthroplasty versus Hemi-arthroplasty (HEALTH) and Fixation using Alternative Implants for the Treatment of Hip fractures (FAITH) trials.

Methods

This study included 2,388 patients from the HEALTH and FAITH trials, with 90-day and one-year mortality proportions of 3.0% (71/2,388) and 6.4% (153/2,388), respectively. The mean age was 75.9 years (SD 10.8) and 65.9% of patients (1,574/2,388) were female. The algorithms included patient and injury characteristics. Six algorithms were developed, internally validated and evaluated across discrimination (c-statistic; discriminative ability between those with risk of mortality and those without), calibration (observed outcome compared to the predicted probability), and the Brier score (composite of discrimination and calibration).

Results

The developed algorithms distinguished between patients at high and low risk for 90-day and one-year mortality. The penalized logistic regression algorithm had the best performance metrics for both 90-day (c-statistic 0.80, calibration slope 0.95, calibration intercept -0.06, and Brier score 0.039) and one-year (c-statistic 0.76, calibration slope 0.86, calibration intercept -0.20, and Brier score 0.074) mortality prediction in the hold-out set.

Conclusion

Using high-quality data, the ML-based prediction models accurately predicted 90-day and one-year mortality in patients aged 50 years or older with a FNF. The final models must be externally validated to assess generalizability to other populations, and prospectively evaluated in the process of shared decision-making.

Cite this article: *Bone Jt Open* 2023;4-3:168–181.

Keywords: Machine learning, Hip fracture, Artificial intelligence, Prediction models, Shared decision-making

Correspondence should be sent to Hidde Dijkstra; email: h.b.dijkstra@umcg.nl

doi: 10.1302/2633-1462.43.BJO-2022-0162.R1

Bone Jt Open 2023;4-3:168–181.

Introduction

Each year more than 300,000 older adults are treated for hip fracture in the USA,[1] and over the next decade this number is expected to rise to approximately 400,000 cases.[2,3] Global incidence is expected to increase from 2.5 million in 2025 to more than 6 million cases occurring annually in 2050 due to an ageing population.[4–6] Most hip fracture patients are treated surgically, with mortality rates up to 35% in the first postoperative year.[7–10] In addition, hip fractures represent a significant financial burden to patients and the healthcare system, with $10.3 to

$15.2 billion spent annually in the USA.[11] Although mostly treated surgically, a recent study showed that a shared decision-making process including nonoperative management for a proximal femoral fracture might be a viable option for frail institutionalized patients with limited life expectancy,[12] which is mostly not the case for younger, vital, and mobile patients living at home.

Stratifying patients preoperatively to their mortality risk factors may aid treatment-related shared decision-making.[13,14] Therefore, numerous studies have identified risk factors for predicting mortality in this population using multivariable logistic regression.[15–19] Using Cox proportional hazard regression, Bzovsky et al[20] identified factors associated with 90-day and 24-month mortality based on patients enrolled in the Hip fracture Evaluation with Alternatives of Total hip arthroplasty versus Hemiarthroplasty (HEALTH) and the Fixation using Alternative Implants for the Treatment of Hip fractures (FAITH) trials (Supplementary Table i).[21,22] However, this study did not subsequently develop a clinical prediction model using the respective identified predictors. A clinical prediction model that includes such individualized prediction could empower patients and their families to make a personal decision for a surgical strategy that best fits their individual values and needs. This would facilitate data-driven individualized risk stratification for true shared decision-making between patient and care provider. Recently, in orthopaedic surgery, various studies have made use of machine-learning (ML) algorithms for developing clinical prediction models,[22] and these show good results in terms of accuracy.[23–37] The use of two high-quality large international databases (HEALTH and FAITH) with prospective follow-up data of patients aged 50 years or older with hip fractures creates opportunity to develop such ML prediction models to predict the risk of short- and longer-term mortality in patients with a femoral neck fracture (FNF).

We therefore asked: can we develop a prediction model using machine learning algorithms for 90-day and one-year mortality prediction in FNF patients who are 50 years or older based on the HEALTH and FAITH trials?

Methods

This study was performed according to the Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis Guideline (TRIPOD-Statement) (Supplementary Material 2)[38] and the JMIR Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research.[39]

Data safety. Our Machine Learning Consortium[33–37] adhered to the "Policy on Use and Sharing of Data Collected by World Health Organization (WHO) in Member States Outside the Context of Public Health Emergencies" of the WHO for safe multicentre data exchange and analysis.[40]

Ethical approval. The HEALTH trial (ClinicalTrials.gov NCT00556842) and the FAITH trial (ClinicalTrials.gov NCT00761813) was approved by the Hamilton integrated research ethics board (#06-151 and #06-402) and the participating clinical sites' research ethics boards.[21,22]

Study design and population. Data were derived from the HEALTH and FAITH trials.[21,22] The HEALTH trial was an international, expertise-based, randomized controlled trial (RCT) comparing total hip arthroplasty (THA) and hemiarthroplasty.[21] Between January 2009 and May 2017, HEALTH investigators included patients without a previous history with frank dementia, who were 50 years of age or older, who had low-energy (defined as fall from standing height) displaced fractures of the femoral neck which were planned to be treated surgically. This trial enrolled 1,441 patients with a mean age of 79 years (standard deviation (SD) 8) with 70% of patients (1,009/1,441) being female, and 49.8% of patients (718/1,411) treated with THA versus 50.2% (723/1,441) with hemiarthroplasty.

The FAITH trial was an international, multicentre, concealed RCT comparing fixation of FNF with cancellous screws versus a sliding hip screw (SHS).[22] Between March 2008 and March 2014, the FAITH investigators included mentally fit patients aged 50 years or older, with a low-energy fracture of the hip requiring fracture fixation. The FAITH trial enrolled 1,079 patients with a mean age of 72 years (SD 12) with 60% being female (648/1,079), and 50.2% (542/1,079) treated with a SHS versus 49.8% (537/1,079) with cancellous screws.

The HEALTH dataset and FAITH dataset combined resulted in a dataset of 2,520 patients. Patients with more than 5.0% missing data were excluded, leaving 2,388 patients for analysis (Table I). The 90-day mortality proportion was 3.0% (71/2,388) and the one-year mortality proportion was 6.4% (153/2,388). The mean age was 75.9 years (SD 10.8), 65.9% of patients (1,574/2,388) were female, and 89.3% (2,133/2,388) were white. The level of fracture line was subcapital in 62.9% of fractures (1,502/2,388), and 70.0% (1,671 of 2,388) were displaced. Other patient and fracture characteristics are displayed in Table I and Table II.

Primary outcomes. The primary outcome of this study was 90-day and one-year mortality (based on the date of surgery) following a FNF, as reported in the HEALTH and FAITH trials.

Candidate input variables. Variables that could be used for the development of the algorithms had to be available in both HEALTH and FAITH databases.[21,22] Based on previous work,[20,41–44] variables that were considered potentially important for predicting 90-day and one-year mortality are displayed in Table III. As a first step, variables potentially associated with 90-day and one-year mortality were identified with random forest algorithms.[45,46] Random forest is a commonly used technique that aims to work well for various classification and regression tasks. First, a model

Table I. Patient characteristics (n = 2,388).

Patient characteristics	Value
Mean age, yrs (SD)	75.9 (10.8)
Age, n (%)	
50 to 80	45 (39.6)
80+	1,443 (60.4)
Sex, n (%)	
Female	1,574 (65.9)
Male	814 (34.1)
Mean BMI, kg/m^2 (SD)	24.8 (4.7)
Ethnicity, n (%)	2,386
Indigenous	7 (0.3)
South Asian	138 (5.8)
East Asian	23 (1.0)
Hispanic/Latino	17 (0.7)
White	2,133 (89.4)
Black	65 (2.7)
Middle Eastern	3 (0.1)
ASA grade, n (%)	2,286
I	214 (9.4)
II	975 (42.7)
III	954 (41.7)
IV	143 (6.3)
Pre-fracture living status, n (%)	
nstitutionalized	113 (4.7)
Not institutionalized	2,275 (95.3)
Pre-fracture functional status, n (%)	
Able to walk without assistive device	1,807 (75.7)
Use of ambulatory assistive device	581 (24.3)
Current medications, n (%)	
None	510 (21.4)
NSAID	322 (13.5)
Opioid use	214 (9.0)
Osteoporosis medications (not bisphosphonates)	178 (7.5)
Anti-hypertension medications	1,301 (54.5)
Pulmonary (respiratory system) medications	292 (12.2)
General cardiac medications	899 (37.6)
Comorbidities, n (%)	
Osteoporosis	260 (10.9)
Heart disease	816 (34.2)
Lung disease	437 (18.3)
Diabetes	440 (18.4)
Kidney disease	228 (9.6)
Ulcer or stomach disease	264 (11.1)
Anaemia or other blood disease	197 (8.2)
Depression	320 (13.4)
Cancer	265 (11.1)
High blood pressure	1,430 (59.9)
Back pain	345 (14.5)
Osteoarthritis, degenerative arthritis	480 (20.1)
Rheumatoid arthritis	66 (2.8)
Mortality rates, n (%)	
90-day mortality	71 (3.0)
One-year mortality	153 (6.4)

ASA, American Society of Anesthesiologists; NSAID, nonsteroidal anti-inflammatory drugs; SD, standard deviation.

Table II. Fracture characteristics (n = 2,388).

Fracture characteristic	Value
Fracture side, n (%)	
Left	1,261 (52.8)
Right	1,127 (47.2)
Mechanism of injury, n (%)	2,387
Fall from standing	2,320 (97.2)
Spontaneous fracture	46 (1.9)
Fall from small height	9 (0.4)
Other low-energy trauma	12 (0.5)
Level of fracture line, n (%)	
Subcapital	1,502 (62.9)
Midcervical	748 (31.3)
Basal	138 (5.8)
Fracture displacement, n (%)	1,671 (70.0)
Garden classification, n (%)	
I	528 (22.1)
II	189 (7.9)
III	829 (34.7)
IV	842 (35.3)
Pauwels classification, n (%)	2,386
I	226 (9.5)
II	1,489 (62.4)
III	671 (28.1)
Any additional injuries (but not fractures), n (%)	195 (8.2)

is fitted with all variables, and then less relevant features are removed.[45,46] Nine variables were identified for developing algorithms for 90-day mortality (Figure 1) (in order of importance): prefracture functional status (the use of an assistive device for ambulation), American Society of Anesthesiologists (ASA)[47] grade, sex, osteoarthritis, age, anaemia or other blood disease, prefracture living status (institutionalized or not), ethnicity, and BMI.

For one-year mortality, 15 predictive variables were identified (Figure 2) (in order of importance): ASA grade, pre-fracture functional status, ethnicity, age, opioid use, antihypertension medication use, high blood pressure, heart disease, general cardiac medication use, sex, BMI, level of fracture line, back pain, pre-fracture living status, and cancer.

Missing data. Missing data were calculated in the pre-processing stage, and those variables with < 5% missing data were imputed using the missForest method.[48] This algorithm imputes missing values in continuous and categorical data, based on averaging regression trees. This algorithm has been chosen as it has been shown to outperform other methods of imputation, especially when complex interactions and non-linear relations are suspected.[48] We imputed missing values for BMI (2% (43/2,388)), ASA grade (4% (102/2,388)), and kidney disease (1% (15/2,388)).

Model development. First, the total dataset was divided into a training set (80%; n = 1,910) and a hold-out test set (20%; n = 478)) stratified on the outcomes (90-day and one-year mortality). Second, variables determined from

Table III. Candidate input variables.

Variable	Explanation
Age	Years
Sex	Male or female
BMI	kg/m^2
Ethnicity	Indigenous, South Asian, East Asian, Hispanic/Latino, White, Black or Middle Eastern
ASA grade	I to IV
Pre-fracture living status	Institutionalized or not institutionalized
Pre-fracture functional status	Independent ambulator or uses an ambulatory assistive device
Use of medication in general	Yes or no
NSAID	Yes or no
Opioid use	Yes or no
Other osteoporosis medications (not bisphosphonates)	Yes or no
Anti-hypertension medications	Yes or no
Pulmonary medications	Yes or no
General cardiac medications	Yes or no
Osteoporosis	Yes or no
Anaemia or other blood disease	Yes or no
Heart disease	Yes or no
Lung disease	Yes or no
Diabetes	Yes or no
Ulcer or stomach disease	Yes or no
Kidney disease	Yes or no
Depression	Yes or no
Cancer	Yes or no
High blood pressure	Yes or no
Back pain	Yes or no
Osteoarthritis, degenerative arthritis	Yes or no
Rheumatoid arthritis	Yes or no
Fractured hip	Left or right hip
Mechanism of injury	Fall from standing, spontaneous fracture, fall from small height, or other low-energy trauma
Level of fracture line	Subcapital, midcervical, or base of the femoral neck
Fracture displacement	Yes or no
Garden classification	I to IV
Pauwels classification	I to III
Any additional injuries (but not fractures)	Yes or no

ASA, American Society of Anesthesiologists; NSAID, nonsteroidal anti-inflammatory drugs.

the feature selection step were used to train pre-existing Microsoft Azure (USA) ML algorithms: 1) Bayes Point Machine, 2) Boosted Decision Tree, 3) Jungle Decision, 4) Penalized Logistic Regression (PLR), 5) Neural Network, and 6) Support Vector Machine (SVM). These algorithms were chosen based on prior ML studies and their binary classification capabilities.[31,49–51] The algorithms are supervised forms of ML, meaning that model development relies on the training of the algorithm with labelled data (the presence or absence of 90-day and one-year mortality). The reason why we train different algorithms is because you cannot predict beforehand which algorithm is performing best. Algorithm training and performance assessment were performed with ten-fold cross-validation repeated three times in the training set.[38,39] Cross-validation means dividing data into a selected number of groups, named folds. First, the data are divided into ten equally sized folds. Then, the algorithms were trained on nine of the ten folds (90% of the training data) and tested on the remaining fold (10% of the training data). Subsequently, algorithms with good performance in the training set were tested in the hold-out set whereafter performance metrics could be calculated, as described in the following section.

Performance measures. The following measures were used for the assessment of predictive performance of the algorithms: discrimination, calibration, and Brier-score (overall model performance).[52] Discrimination was assessed with the c-statistic (i.e. the area under the curve (AUC) of a receiver operating characteristic curve (ROC-curve)). The ROC-curve plots the sensitivity (true positive rate) against 1 - specificity (false positive rate). The c-statistic ranges from 0.50 to 1.0, with 1.0 indicating the highest discriminating score and 0.50 indicating the

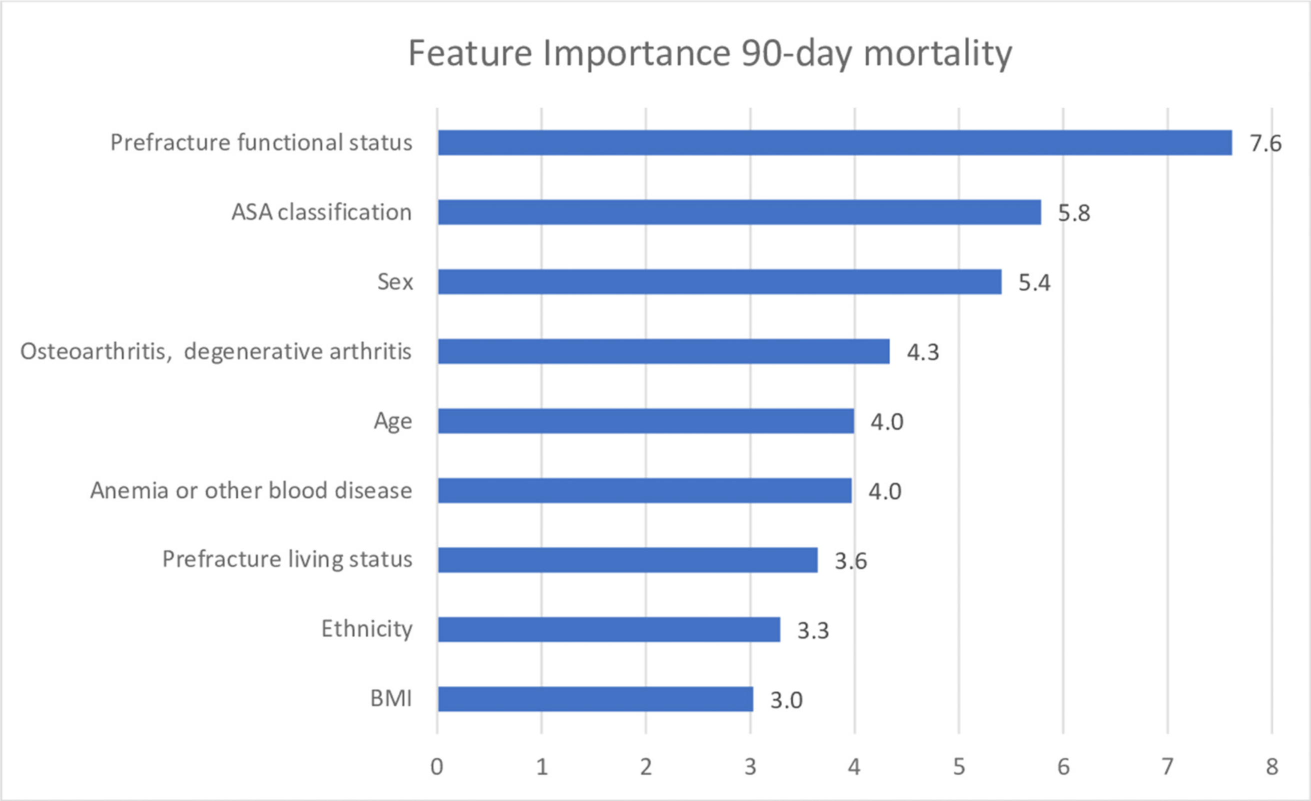


Fig. 1

Variable importance for 90-day mortality based on feature selection using random forest algorithms.

lowest discriminating score. This differentiates between patients who had the outcome of interest (i.e. 90-day and one-year mortality) from those who did not.[52]

Calibration was assessed with the calibration slope and calibration intercept of a calibration curve.[53] Calibration reflects the agreement between the observed outcome and the predicted probability. It can be assessed and visualized by plotting the predicted probability (x-axis) versus the actual probability (y-axis) creating a calibration curve. The intercept indicates that predictions are systematically too high (intercept < 0) or too low (intercept > 0). The slope of the calibration curve reflects whether predictions were too extreme (low predictions too low and high predictions too high; a slope smaller than 1) or not extreme enough (low predictions not low enough and high predictions not high enough; a slope larger than 1).[54] A perfect calibration curve has a slope of 1 and an intercept of 0.[54,55]

Finally, overall algorithm performance – a composite of discrimination and calibration – was assessed using the Brier score.[55,56] The Brier score is obtained by calculating the squared differences between the actual outcomes and the predictions. The score can range from 0, (indicating a perfect model) to 1 (indicating the worst possible). The upper limit of the Brier score is dependent on the incidence of outcomes in the dataset.[55,56] Therefore, the upper limit of the Brier score also was calculated and presented. First, the performance of the algorithms was evaluated in the training set. Thereafter, the performance was evaluated on the hold-out set based on the same performance measures as described above.

Sample size. Given the retrospective study design, post-hoc power analyses were conducted to evaluate the sample size of the study with an α value of 0.05. The post-hoc power analyses revealed 100% power in both evaluations (α = 0.05).

Web-application. The best performing algorithms across the performance metrics were incorporated into an online open-access prediction tool.

Statistical analysis. Baseline characteristics were calculated and presented as frequencies and percentages for categorical variables, whereas means and standard deviations (SDs) were used for continuous variables. Data pre-processing and analysis was performed using SPSS v. 25 (IBM, USA), and Azure (Microsoft, USA) was used for model training and development.

Results

Performance of ML algorithms predicting 90-day mortality and web application. In the training set (n = 1,910), the six algorithms showed c-statistics ranging from 0.73 to 0.78, calibration slopes ranging from 0.76 to 0.95,

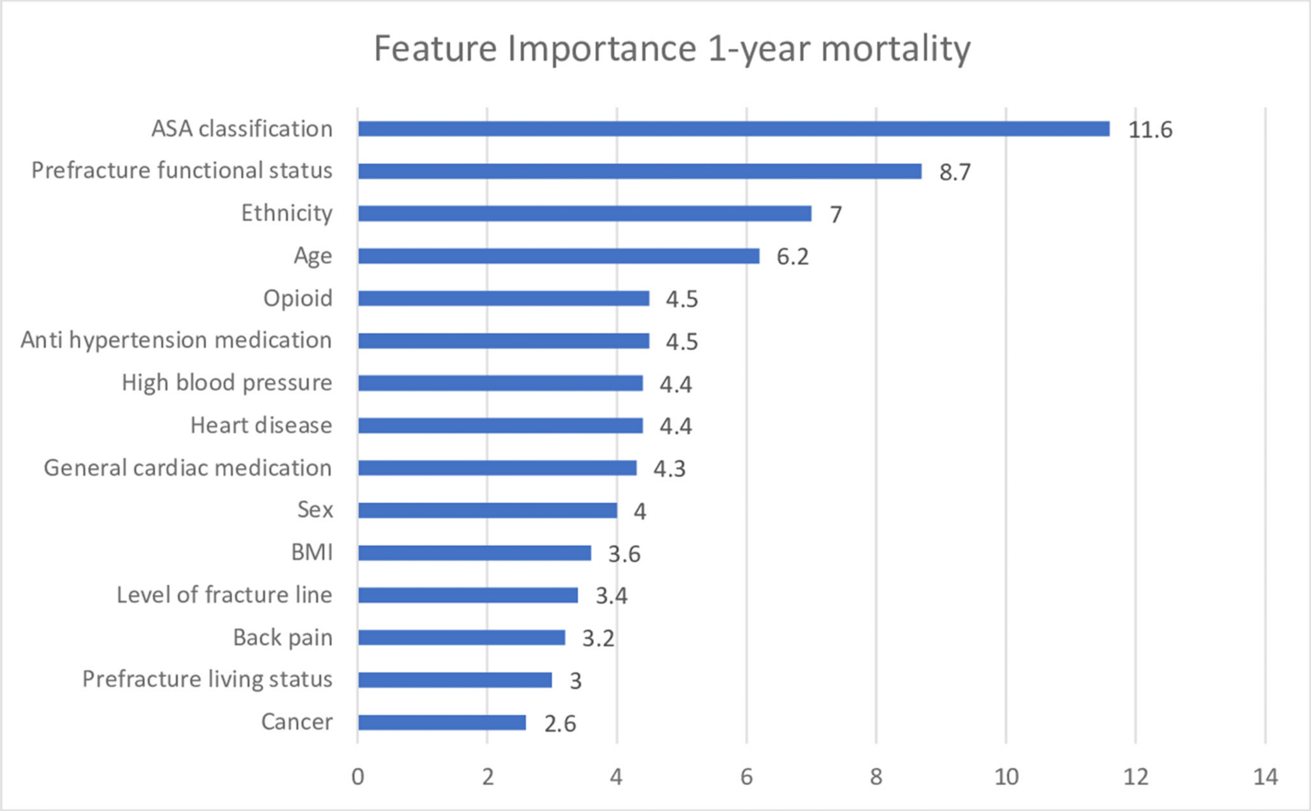


Fig. 2

Variable importance for one-year mortality based on feature selection using random forest algorithms.

Table IV. Performance of machine-learning algorithms in predicting 90-day and one-year mortality in the training set (n = 1,910) after ten-fold cross validation repeated three times.

Metric (95% CI)	Bayes Point Machine	Boosted Decision Tree	Jungle Decision Algorithm	Penalized Logistic Regression	Neural Network	Support Vector Machine
90-day mortality						
c-statistic	0.76 (0.73 to 0.79)	0.78 (0.75 to 0.80)	0.78 (0.75 to 0.80)	0.78 (0.75 to 0.81)	0.77 (0.74 to 0.80)	0.73 (0.70 to 0.76)
Calibration slope	0.90 (0.78 to 1.03)	0.76 (0.66 to 0.86)	0.95 (0.82 to 1.08)	0.93 (0.81 to 1.05)	0.86 (0.75 to 0.98)	0.79 (0.66 to 0.92)
Calibration intercept	-0.05 (-0.18 to 0.08)	0.00 (-0.14 to 0.13)	-0.02 (-0.15 to 0.12)	0.00 (-0.13 to 0.13)	-0.18 (-0.32 to -0.05)	0.00 (-0.13 to 0.13)
Brier score*	0.039	0.040	0.039	0.039	0.039	0.040
One-year mortality						
c-statistic	0.78 (0.76 to 0.80)	0.78 (0.76 to 0.80)	-	0.78 (0.76 to 0.80)	0.78 (0.76 to 0.80)	0.73 (0.70 to 0.75)
Calibration slope	0.90 (0.81 to 0.99)	0.78 (0.70 to 0.86)	-	0.95 (0.86 to 1.05)	0.78 (0.70 to 0.86)	0.82 (0.72 to 0.92)
Calibration intercept	-0.03 (-0.13 to 0.07)	0.00 (-0.10 to 0.10)	-	0.00 (-0.10 to 0.10)	0.16 (0.06 to 0.26)	0.00 (-0.10 – 0.10)
Brier score*	0.073	0.074	-	0.073	0.073	0.076

Dashes signify negative (below zero) values.
*Upper Limit Brier score for 90-day mortality was 0.041 and for one-year mortality 0.079.
CI, confidence interval.

and calibration intercepts ranging from -0.18 to 0.00 (Table IV). Moreover, the Brier score was either 0.039 or 0.040 relative to the upper limit of 0.041.

Five of the six algorithms were further tested in the hold-out set (n = 478) based on better performance in the training set (Table V). The five algorithms showed

Table V. Performance of machine-learning algorithms in predicting 90-day and one-year mortality in hold-out set (n = 478).

Metric	Bayes Point Machine	Boosted Decision Tree	Jungle Decision Algorithm	Penalized Logistic Regression*	Neural Network	Support Vector Machine
90-day mortality						
c-statistic	0.77	0.80	0.81	0.80	0.79	-
Calibration slope	0.94	0.77	1.25	0.95	0.87	-
Calibration intercept	-0.13	-0.58	-0.66	-0.06	-0.43	-
Brier score†	0.039	0.041	0.038	0.039	0.040	-
One-year mortality						
c-statistic	0.74	0.74	-	0.76	0.72	0.66
Calibration slope	0.74	0.75	-	0.86	0.52	0.60
Calibration intercept	-0.43	-0.43	-	-0.20	-0.81	-0.85
Brier score†	0.075	0.075	-	0.074	0.077	0.079

Dashes signify negative (below zero) values.
*This algorithm was chosen as the final prediction model for 90-day and one-year mortality prediction.
†Upper limit Brier score for 90-day mortality was 0.042 and for one-year mortality 0.079.

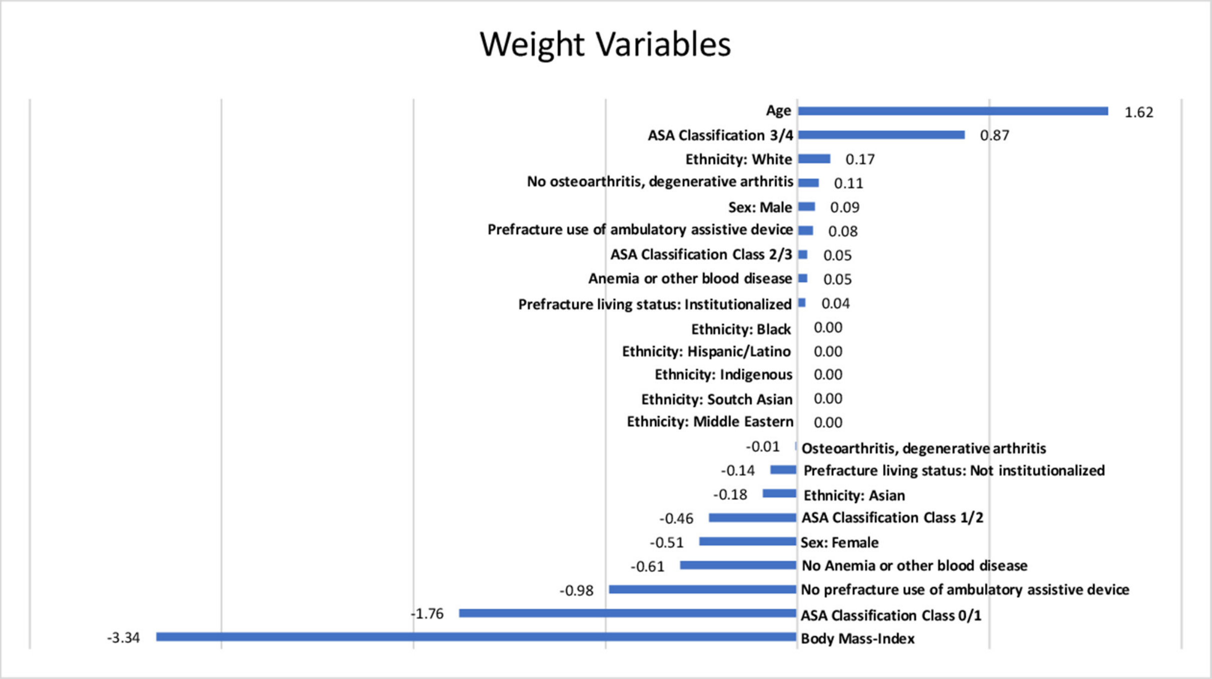


Fig. 3

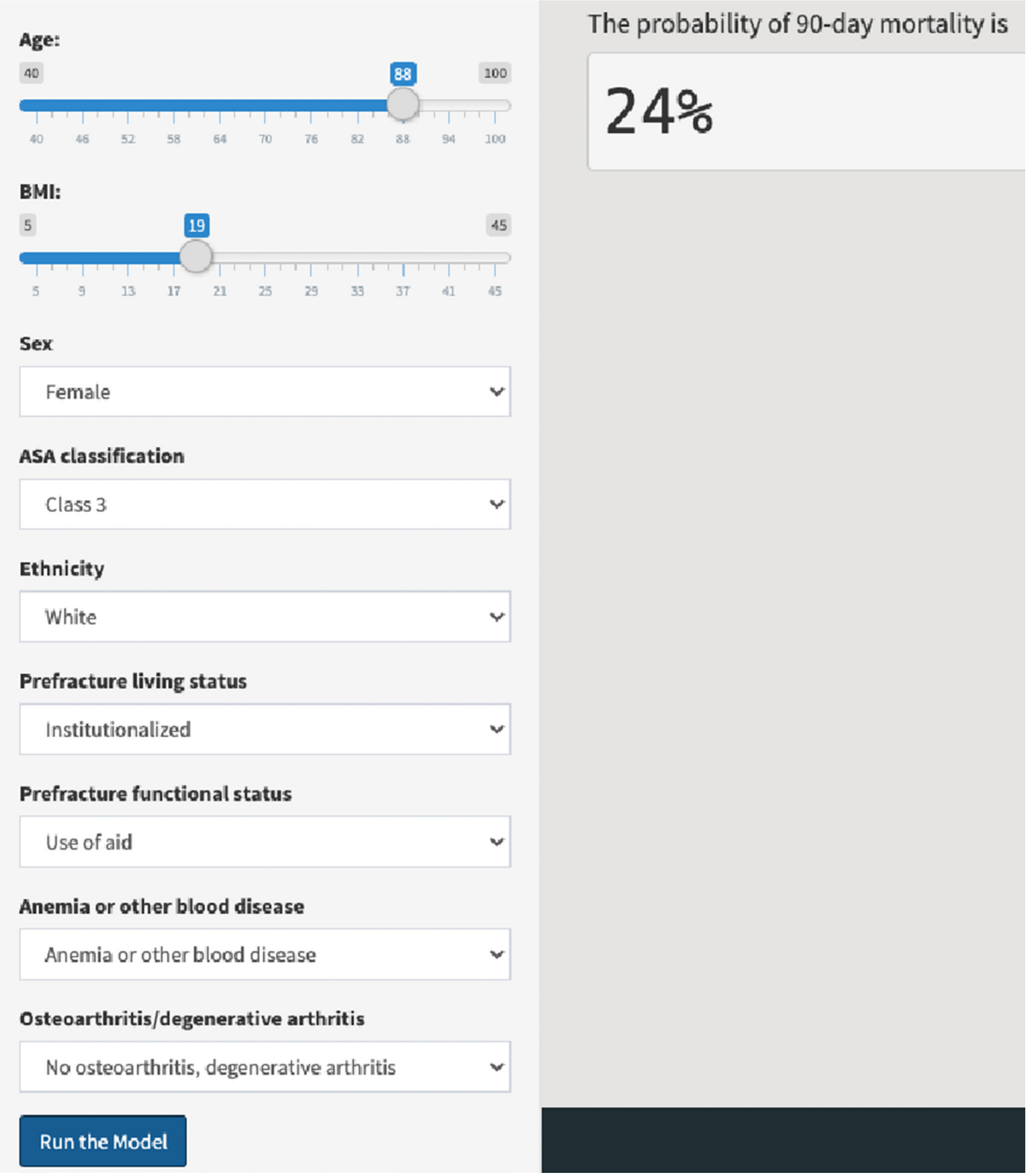
Importance of variables for 90-day mortality prediction based on the Penalized Logistic Regression algorithm. ASA, American Society of Anesthesiologists.

c-statistics ranging from 0.77 to 0.81, calibration slopes ranging from 0.87 to 1.25, calibration intercepts ranging from -0.66 to -0.06, and Brier scores ranging from 0.038 to 0.041, relative to the upper limit of 0.042.

For 90-day mortality prediction, based on better calibration in the hold-out set (calibration slope 0.95 vs 0.94 and intercept -0.06 to -0.13), we chose the PLR algorithm as the final prediction model. Age, ASA grade III/IV, white ethnicity, and osteoarthritis were the strongest predictors in this model (Figure 3). This algorithm was incorporated in an online open access prediction tool (see Figure 4 for a case example).[57]

Performance of ML algorithms predicting one-year mortality and web application. In the training set (n = 1,910), four of the five algorithms showed c-statistics of 0.78 and one showed a c-statistic of 0.73. The calibration slopes ranged from 0.78 to 0.95, and calibration intercepts ranged from -0.03 to 0.16. The Brier scores ranged from 0.073 to 0.076 relative to the upper limit of 0.079 (Table IV).

All five algorithms were further tested in the hold-out set (n = 478) and showed c-statistics ranging from 0.66 to 0.76, calibration slopes ranging from 0.52 to 0.86, and calibration intercepts ranging from -0.20 to -0.85. Brier



Age: 88

BMI: 19

Sex
Female

ASA classification
Class 3

Ethnicity
White

Prefracture living status
Institutionalized

Prefracture functional status
Use of aid

Anemia or other blood disease
Anemia or other blood disease

Osteoarthritis/degenerative arthritis
No osteoarthritis, degenerative arthritis

The probability of 90-day mortality is

24%

Fig. 4

Probability generated by the prediction tool for 90-day mortality for a fictious case. For an 88-year-old white female patient with a BMI of 19 kg/m^2, who lives in a nursing home, makes use of an ambulatory assistive device, has anaemia, and no osteoarthritis, the prediction tool generated a 24% probability of mortality in 90 days after the femoral neck fracture.

scores ranged from 0.074 to 0.079 relative to the upper limit of 0.079 (Table V).

For one-year mortality prediction, we found better calibration in the training set (calibration slope 0.95 vs

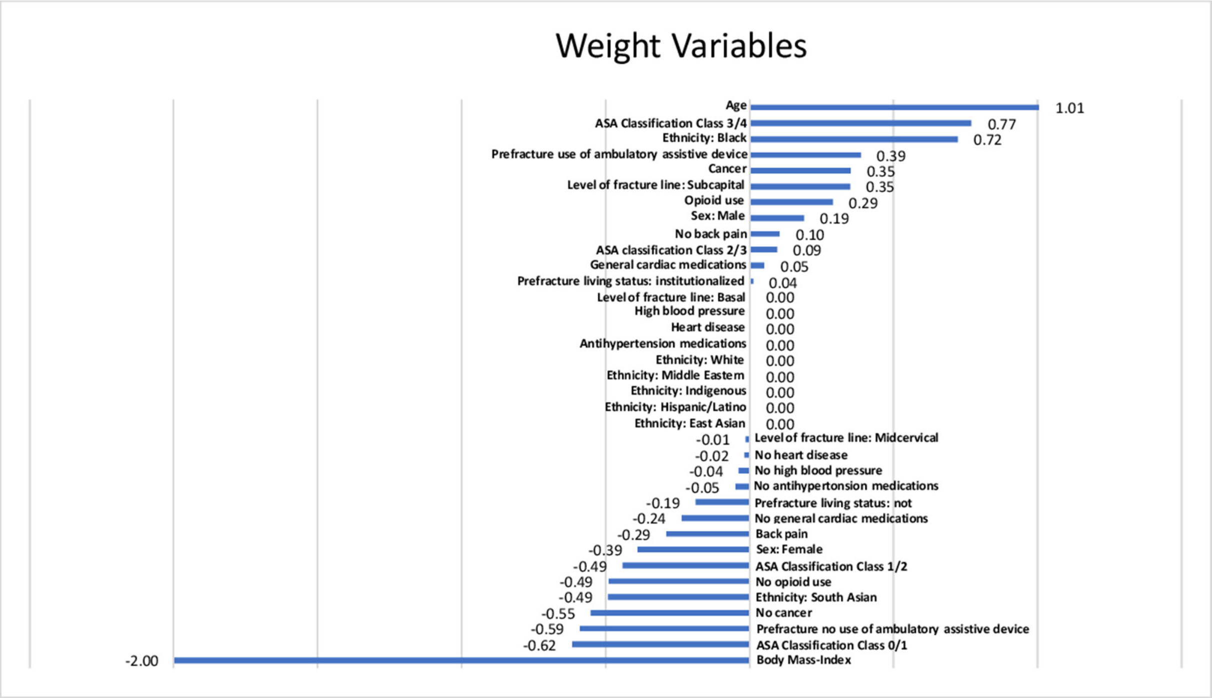


Fig. 5

Importance of variables for one-year mortality prediction based on the Penalized Logistic Regression algorithm. ASA, American Society of Anesthesiologists.

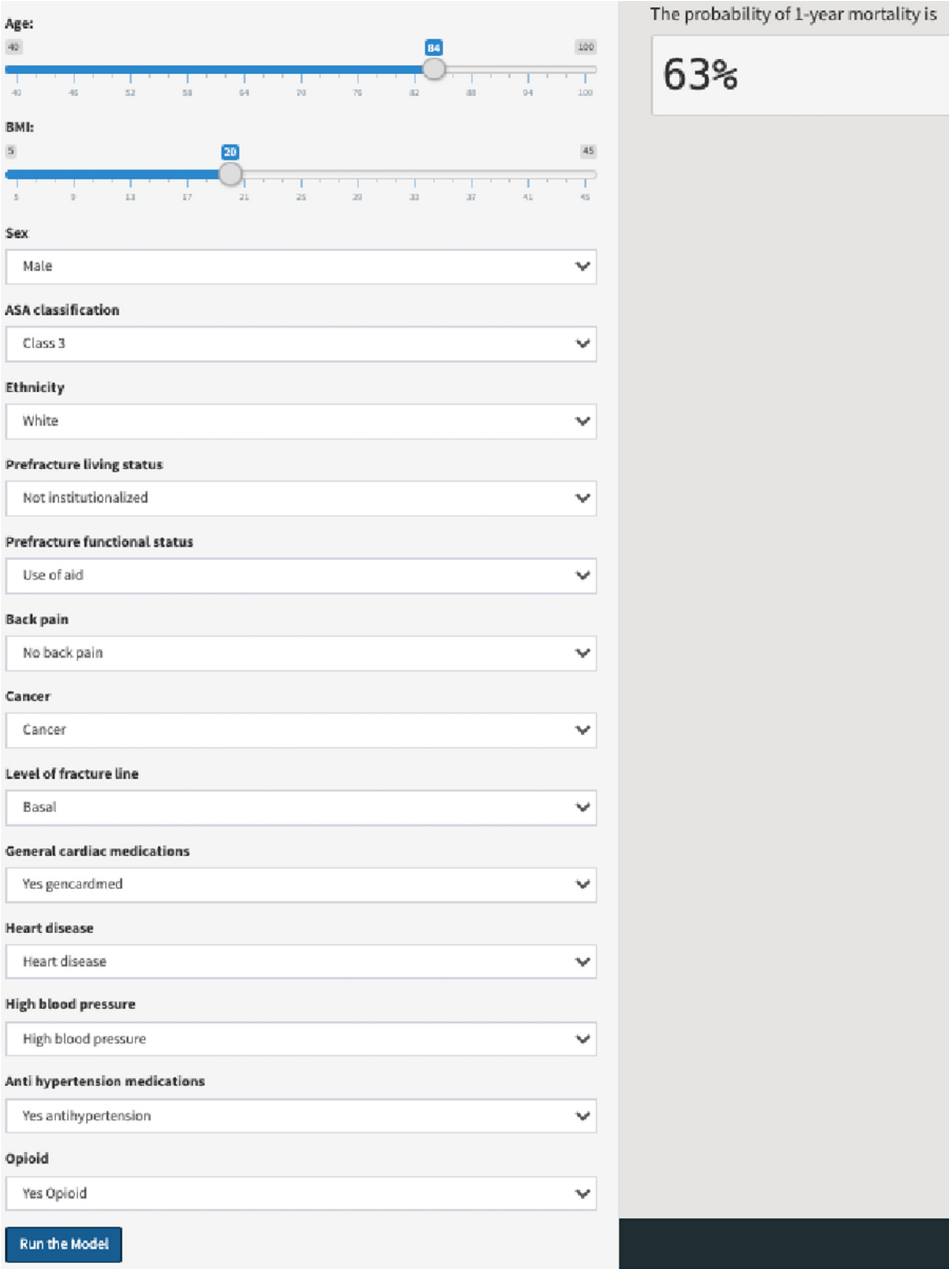
0.90) and hold-out set (calibration slope 0.86 vs 0.75 and intercept -0.20 vs -0.43) for the PLR algorithm, which was therefore chosen as the final prediction model. Age, ASA grade III/IV, black ethnicity, pre-fracture functional status, and cancer were the strongest predictors in this model (Figure 5). This algorithm was incorporated in an online open access prediction tool (see Figure 6 for a case example).[58] During the process of treatment decision-making, clinicians can generate personalized risk estimates with this web application, which might inform the clinician about the prognosis of the patient.

Discussion

Given the major public health concerns associated with FNFs in patients aged 50 years or older,[59] it is necessary to gain a further understanding of which patients are at risk for mortality at follow-up. This risk estimation might help surgeons to accurately inform patients about prognosis. Subsequently, a more accurate individualized risk stratification – devoid of surgeon bias – could empower patients and their families to make a personal decision for a surgical strategy that best fits their individual values and needs. This would facilitate true shared decision-making between patient and surgeon.

This study has several strengths. First, it used data from two high-quality RCTs of well-characterized patients with similar inclusion criteria and follow-up durations, which were designed by the same principal investigators.[21,22] Second, a large sample of patients from 12 countries and diverse clinical healthcare settings were included in both studies; this heterogeneity increases external validity of the algorithms. Third, the best-fit algorithms showed good performance, which is promising for external validation studies.

The results of this study should also, however, be viewed in light of several limitations. First, the 90-day mortality rate was 3.0% and the one-year mortality rate was 6.4%, which is relatively low compared to the literature.[9,10] This may be explained by the inclusion of relatively healthy patients in both trials.[21,22] External validation of the algorithms should be carried out in patient cohorts with representative and higher mortality rates, ensuring transportability of the algorithms. Second, there are likely other variables associated with mortality which have not been collected as part of the HEALTH and FAITH trials. For example, smoking status, a known prognostic factor for mortality after hip fracture,[60–62] has only been reported in the FAITH trial and could therefore not be included in the combined dataset. We combined both datasets to retrieve a sufficient sample size for model development, however the cost of this was that not all variables from the datasets were overlapping, and therefore we carried out variable selection with variables that were present in both datasets. Future research efforts can harmonize their data collection strategies leading to common data elements, which makes it easier to combine datasets. Third, both the HEALTH and FAITH trials were not designed for developing ML algorithms, leading to indication bias.



Age: 84

BMI: 20

Sex: Male

ASA classification: Class 3

Ethnicity: White

Prefracture living status: Not institutionalized

Prefracture functional status: Use of aid

Back pain: No back pain

Cancer: Cancer

Level of fracture line: Basal

General cardiac medications: Yes gencardmed

Heart disease: Heart disease

High blood pressure: High blood pressure

Anti hypertension medications: Yes antihypertension

Opioid: Yes Opioid

The probability of 1-year mortality is

63%

Fig. 6

Probability generated by the prediction tool for one-year mortality for a fictitious case. For an 84-year-old white male patient with a BMI of 20 kg/m^2 and American Society of Anesthesiologists grade III, who lives at home, makes use of an ambulatory assistive device, has cancer, heart disease, high blood pressure, and no back pain, uses opioids, general cardiac, and antihypertension medications, and sustained a femoral neck fracture at the base of the neck, the prediction tool generated a 63% probability of mortality one year after the femoral neck fracture.

Fourth, patients with more than 5.0% missing data were excluded, and therefore this study may have been subject to selection and indication bias. Fifth, the current study solely investigated the mortality risk estimation; future research can focus on investigating additional outcomes such as patient-reported outcome measures (e.g. quality of life, symptoms of pain). This will lead to more patient-centred care and evaluation of the individual patient's needs. Lastly, patients included in both trials may not reflect the demographic and clinical characteristics of patients for whom the prediction models are ultimately used. However, as this study used high-quality data from 12 different countries and various healthcare settings, we expect that after external validation studies in specific geographical regions (i.e. geographical validation) and subsequent model development (e.g. model retraining), the models can be used in different geographical regions. Therefore, future validation and performance assessment studies are needed.

For 90-day mortality prediction, the PLR algorithm was chosen as the final prediction model and showed good performance. We found that our c-statistics were comparable for models developed for 30-day and 90-day mortality prediction showing c-statistics ranging from 0.73 to 0.92.[24,26,27,29,63] However, most studies did not report calibration metrics or Brier scores.[24–29] This study identified the following predictive variables for 90-day mortality: pre-fracture functional status, ASA grade, sex, osteoarthritis, age, anaemia or other blood disease, pre-fracture living status, ethnicity, and BMI. Using regression analysis in the same cohorts, Bzovsky et al[20] identified corresponding variables associated with the outcome 90-day mortality: older age (ten-year increase), low BMI (five-point decrease), ASA grade (III to V), the use of an ambulatory assistive device pre-injury, and kidney disease. However, this study identified additional associated variables and we subsequently developed an individualized clinical ML-based prediction model using these. These findings correspond with previous studies which identified ASA grade,[41] age,[41,42] sex,[41,42] and pre-fracture functional status[42] as variables associated with the outcome 90-day mortality. The high predictive value of reduced pre-fracture functional status may be associated with lower mobility before the fracture, hence rehabilitation takes longer, with patients not being able to return to their pre-fracture level of activity, leading to a higher risk of mortality.[43]

For one-year mortality, again the PLR algorithm yielded the best performance in the training- and hold-out set with c-statistics of 0.78 and 0.76, respectively; this corresponds with previous findings.[26,27] We identified 15 predictive variables which can be categorized into patient characteristics (age, ethnicity, BMI, and pre-fracture functional status), comorbidities (high blood pressure, back pain, heart disease, and cancer), medication use (opioids, general cardiac medications), and health status with fracture characteristics (ASA grade, level of fracture line). Therefore, as shown in two major reviews,[43,44] predicting one-year mortality is dependent on multicategory variables. Moreover, for developing the prediction models we not only used known predictors for this outcome (age,[64–66] pre-fracture living status,[67–69] cancer, and heart disease)[64,65,68] but also new predictors: ethnicity, BMI, high blood pressure, back pain, use of opioids, use of general cardiac medications, and level of fracture line. ASA grade, the most predictive variable for one-year mortality in this study, has previously thoroughly been identified as a significant one-year mortality predictor.[65,69–71] Although this study had comparable outcome performance measures for 90-day and one-year mortality prediction, we believe that the models developed in this study are of added value since this study complied with the TRIPOD statement,[38] used high-quality data, and developed an open-access web application for real-time estimate calculation. For both 90-day and one-year mortality, we identified ethnicity as a predictive variable. Literature shows that ethnicity is an important variable to include when building prediction models, hereby preventing racial bias, and future studies should be aware of this.[72] In the last few years, much research has been performed predicting mortality in FNF patients. Most studies developed prediction tools based on age, sex, and the general presence of comorbidity.[73] In addition, some studies included postoperative variables, such as early postoperative mobilization and postoperative lab values (for example haemoglobin).[19,74] In contrast to the broader presence of comorbidity, this study used the ability of ML algorithms to differ between the effects of different types of comorbidities to estimate the individual prediction value of each variable.

Treatment decision-making is challenging for FNF patients aged 50 years or older and their clinicians. The ML-based prediction models developed in this study only provide personalized risk estimates, which may help the clinician to estimate the individual prognosis of the patient. It would be for the clinician to take this estimate into account during the process of treatment decision-making, preferably in the emergency department before the choice of treatment is made. High mortality estimates might be communicated as 'likely to have a poor prognosis' and low mortality estimates as 'likely to have a better prognosis' in the short and longer term. However, the impact of the algorithms needs to be evaluated (i.e. which estimates exceed certain decision thresholds). Estimating mortality probability might support the choice of treatment (i.e. no surgical intervention vs fracture fixation vs hemiarthroplasty vs THA), discharge destination, and preferred intensity of rehabilitation, and should therefore be interpreted carefully. However, it is important to note that the mortality estimates the models provide are only developed to inform the clinician and, consequently, the patient. The ultimate treatment decision should not solely be based on the outcome of prediction models but is ideally made after a shared decision-making process with the patient and family. More importantly, we believe that other personalized outcomes such as patient-reported outcome measures (e.g. functional mobility, quality of life) and short-term morbidity (i.e. the existence of co-injuries) should also be taken into account in the process of treatment decision-making. In the future, ML-based prediction models may aid this process, potentially leading to a data-driven best fit treatment, based on multiple outcome variables, for an individual patient, and may therefore improve (end-of-life) care.

In conclusion, using high-quality data from the HEALTH and FAITH trials,[21,22] we have developed accurate ML algorithms for 90-day and one-year mortality prediction for individual patients with FNF based on patient and fracture characteristics. The final models must be externally validated to assess the generalizability to other populations, and future studies should prospectively evaluate these ML-driven probability calculators in the process of shared decision-making.

Take home message

- Pending external validation, machine-learning algorithms might support treatment-related decision-making.
- The algorithms have been deployed as open access probability calculators giving personalized and data-driven estimates.

Supplementary material

Table presenting factors associated with 90-day and 24-month mortality in patients of the HEALTH and FAITH trials using Cox proportional hazard regression from Bzovsky et al (2020).

References

1. **Swenning T**, **Leighton J**, **Nentwig M**, **Dart B**. Hip fracture care and national systems: The United States and Canada. *OTA Int*. 2020;3(1):e073.
2. **Cummings SR**, **Rubin SM**, **Black D**. The future of hip fractures in the United States. Numbers, costs, and potential effects of post- menopausal estrogen. *Clin Orthop Relat Res*. 1990;252:163–166.
3. **Stevens JA**, **Rudd RA**. The impact of decreasing U.S. hip fracture rates on future hip fracture estimates. *Osteoporos Int*. 2013;24(10):2725–2728.
4. **Quah C**, **Boulton C**, **Moran C**. The influence of socioeconomic status on the incidence, outcome and mortality of fractures of the hip. *J Bone Joint Surg Br*. 2011;93-B(6):801–805.
5. **Gullberg B**, **Johnell O**, **Kanis JA**. World-wide projections for hip fracture. *Osteoporos Int*. 1997;44(1997):407–413.
6. **Johnell O**, **Kanis JA**. An estimate of the worldwide prevalence, mortality and disability associated with hip fracture. *Osteoporos Int*. 2004;15(11):897–902.
7. **Omsland TK**, **Emaus N**, **Tell GS**, **et al**. Mortality following the first hip fracture in Norwegian women and men (1999-2008). A NOREPOS study. *Bone*. 2014;63:81–86.
8. **Prieto-Alhambra D**, **Reyes C**, **Sainz MS**, **et al**. In-hospital care, complications, and 4-month mortality following a hip or proximal femur fracture: the Spanish registry of osteoporotic femur fractures prospective cohort study. *Arch Osteoporos*. 2018;13(1):96.
9. **Abrahamsen B**, **van Staa T**, **Ariely R**, **Olson M**, **Cooper C**. Excess mortality following hip fracture: A systematic epidemiological review. *Osteoporos Int*. 2009;20(10):1633–1650.
10. **Mundi S**, **Pindiprolu B**, **Simunovic N**, **Bhandari M**. Similar mortality rates in hip fracture patients over the past 31 years. *Acta Orthop*. 2014;85(1):54–59.
11. **Gillespie WJ**. Extracts from "clinical evidence": hip fracture. *BMJ*. 2001;322(7292):968–975.
12. **Loggers SAI**, **Willems HC**, **Van Balen R**, **et al**. Evaluation of quality of life after nonoperative or operative management of proximal femoral fractures in FRAIL institutionalized patients: The FRAIL-HIP Study. *JAMA Surg*. 2022;157(5):424–434.
13. **McLaughlin MA**, **Orosz GM**, **Magaziner J**, **et al**. Preoperative status and risk of complications in patients with hip fracture. *J Gen Intern Med*. 2006;21(3):219–225.
14. **Karnuta JM**, **Navarro SM**, **Haeberle HS**, **Billow DG**, **Krebs VE**, **Ramkumar PN**. Bundled care for hip fractures: A machine-learning approach to an untenable patient-specific payment model. *J Orthop Trauma*. 2019;33(7):324–330.
15. **Jiang HX**, **Majumdar SR**, **Dick DA**, **et al**. Development and initial validation of a risk score for predicting in-hospital and 1-year mortality in patients with hip fractures. *J Bone Miner Res*. 2005;20(3):494–500.
16. **Karres J**, **Kieviet N**, **Eerenberg JP**, **Vrouenraets BC**. Predicting early mortality after hip fracture surgery: The hip fracture estimator of mortality Amsterdam. *J Orthop Trauma*. 2018;32(1):27–33.
17. **Maxwell MJ**, **Moran CG**, **Moppett IK**. Development and validation of a preoperative scoring system to predict 30 day mortality in patients undergoing hip fracture surgery. *Br J Anaesth*. 2008;101(4):511–517.
18. **Pugely AJ**, **Martin CT**, **Gao Y**, **Klocke NF**, **Callaghan JJ**, **Marsh JL**. A risk calculator for short-term morbidity and mortality after hip fracture surgery. *J Orthop Trauma*. 2014;28(2):63–69.
19. **Nijmeijer WS**, **Folbert EC**, **Vermeer M**, **Slaets JP**, **Hegeman JH**. Prediction of early mortality following hip fracture surgery in frail elderly: The Almelo Hip Fracture Score (AHFS). *Injury*. 2016;47(10):2138–2143.
20. **Bzovsky S**, **Comeau-Gauthier M**, **Schemitsch EH**, **et al**. Factors associated with mortality after surgical management of femoral neck fractures. *J Orthop Trauma*. 2020;34 Suppl 3(11):S15–S21.
21. **Bhandari M**, **Einhorn TA**, **et al**, **HEALTH Investigators**. Total hip arthroplasty or hemiarthroplasty for hip fracture. *N Engl J Med*. 2019;381(23):2199–2208.
22. **Nauth A**, **Creek AT**, **Zellar A**. Fracture fixation in the operative management of hip fractures (FAITH): an international, multicentre, randomised controlled trial. *Lancet*. 2017;389(10078):1519–1527.
23. **Vickers AJ**, **Van Calster B**, **Steyerberg EW**. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6.
24. **DeBaun MR**, **Chavez G**, **Fithian A**, **et al**. Artificial neural networks predict 30-day mortality after hip fracture: Insights from machine learning. *J Am Acad Orthop Surg*. 2021;29(22):977–983.
25. **Chen CY**, **Chen YF**, **Chen HY**, **Hung CT**, **Shi HY**. Artificial neural network and Cox regression models for predicting mortality after hip fracture surgery: A population-based comparison. *Medicina (Kaunas)*. 2020;56(5):1–10.
26. **Li Y**, **Chen M**, **Lv H**, **Yin P**, **Zhang L**, **Tang P**. A novel machine-learning algorithm for predicting mortality risk after hip fracture surgery. *Injury*. 2021;52(6):1487–1493.
27. **Cary MP**, **Zhuang F**, **Draelos RL**, **et al**. Machine learning algorithms to predict mortality and allocate palliative care for older patients with hip fracture. *J Am Med Dir Assoc*. 2021;22(2):291–296.
28. **Zhang Y**, **Huang L**, **Liu Y**, **Chen Q**, **Li X**, **Hu J**. Prediction of mortality at one year after surgery for pertrochanteric fracture in the elderly via a Bayesian belief network. *Injury*. 2020;51(2):407–413.
29. **Harris AHS**, **Kuo AC**, **Weng Y**, **Trickey AW**, **Bowe T**, **Giori NJ**. Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clin Orthop Relat Res*. 2019;477(2):452–460.
30. **Karhade AV**, **Thio Q**, **Ogink PT**, **et al**. Predicting 90-day and 1-year mortality in spinal metastatic disease: Development and internal validation. *Neurosurgery*. 2019;85(4):E671–E681.
31. **Thio Q**, **Karhade AV**, **Ogink PT**, **et al**. Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma? *Clin Orthop Relat Res*. 2018;476(10):2040–2048.
32. **Shah AA**, **Karhade AV**, **Bono CM**, **Harris MB**, **Nelson SB**, **Schwab JH**. Development of a machine learning algorithm for prediction of failure of nonoperative management in spinal epidural abscess. *Spine J*. 2019;19(10):1657–1665.
33. **Hendrickx LAM**, **Sobol GL**, **Langerhuizen DWG**, **et al**. A machine learning algorithm to predict the probability of (occult) posterior malleolar fractures associated with tibial shaft fractures to guide "malleolus first" fixation. *J Orthop Trauma*. 2020;34(3):131–138.

34. **Oosterhoff JHF**, **Doornberg JN**, **Machine Learning Consortium**. Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner's hype cycle. *EFORT Open Rev*. 2020;5(10):593–603.
35. **Oliveira e Carmo L**, **van den Merkhof A**, **Olczak J**, **et al**. An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics. *Bone Jt Open*. 2021;2(10):879–885.
36. **Machine Learning Consortium on behalf of the SPRINT Investigators**. A machine learning algorithm to identify patients at risk of unplanned subsequent surgery after intramedullary nailing for tibial shaft fractures. *J Orthop Trauma*. 2021;35(10):e381–e388.
37. **Machine Learning Consortium, on behalf of the SPRINT and FLOW Investigators**. A machine learning algorithm to identify patients with tibial shaft fractures at risk for infection after operative treatment. *J Bone Joint Surg Am*. 2021;103-A(6):532–540.
38. **Collins GS**, **Reitsma JB**, **Altman DG**, **Moons KGM**. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med*. 2015;13(1):1.
39. **Luo W**, **Phung D**, **Tran T**, **et al**. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res*. 2016;18(12):e323.
40. **No authors listed**. WHO policy on the use and sharing of data collected by WHO in Member States outside the context of public health emergencies. July 24, 2021. www.who.int/about/policies/publishing/data-policy (date last accessed 14 February 2023).
41. **Rosencher N**, **Vielpeau C**, **Emmerich J**, **Fagnani F**, **Samama CM**, **ESCORTE group**. Venous thromboembolism and mortality after hip fracture surgery: the ESCORTE study. *J Thromb Haemost*. 2005;3(9):2006–2014.
42. **Dolk T**. Influence of treatment factors on the outcome after hip fractures. *Ups J Med Sci*. 1989;94(2):209–221.
43. **Smith T**, **Pelpola K**, **Ball M**, **Ong A**, **Myint PK**. Pre-operative indicators for mortality following hip fracture surgery: A systematic review and meta-analysis. *Age Ageing*. 2014;43(4):464–471.
44. **Xu BY**, **Yan S**, **Low LL**, **Vasanwala FF**, **Low SG**. Predictors of poor functional outcomes and mortality in patients with hip fracture: A systematic review. *BMC Musculoskelet Disord*. 2019;20(1):568.
45. **Pavlov YL**. *Random forests*. De Gruyter 2000, 2019. https://www.degruyter.com/document/doi/10.1515/9783110941975/html#Vancouver
46. **Breiman L**. Random forests. *Machine Learning*. 2001;45:5–32.
47. **Saklad M**. Grading of patients for surgical procedures. *Anesthesiol*. 1941;2(3):281–284.
48. **Stekhoven DJ**, **Bühlmann P**. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112–118.
49. **Wainer J**. Comparison of 14 different families of classification algorithms on 115 binary datasets. 2016. arxiv.org/abs/1606.00930 (date last accessed 14 February 2023).
50. **Maroco J**, **Silva D**, **Rodrigues A**, **Guerreiro M**, **Santana I**, **de Mendonça A**. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes*. 2011;4:299.
51. **Fernández-Delgado M**, **Cernadas E**, **Barro S**, **Amorim D**. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res*. 2014;15:3133–3181.
52. **Steyerberg EW**, **Vergouwe Y**. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925–1931.
53. **Cox DR**. Two further applications of a model for binary regression. *Biometrika*. 1958;45(3/4):562.
54. **Van Calster B**, **Vickers AJ**. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making*. 2015;35(2):162–169.
55. **Steyerberg EW**, **Vickers AJ**, **Cook NR**, **et al**. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.
56. **Neeman T**. Clinical prediction models: A practical approach to development, validation, and updating by Ewout W. Steyerberg. *Int Stat Rev*. 2009;77(2):320–321.
57. **No authors listed**. 90-day Mortality in Hip Fractures - Artifical Intelligence Prediction Tool. traumaplatform-ai-prediction-tools.shinyapps.io/90daymortalityhip/ (date last accessed 15 February 2023).
58. **No authors listed**. 1-year Mortality in Hip Fractures - Artificial Intelligence Prediction Tool. Trauma Platform. traumaplatform-ai-prediction-tools.shinyapps.io/1yearmortality/ (date last accessed 15 February 2023).
59. **Leal J**, **Gray AM**, **Prieto-Alhambra D**, **et al**. Impact of hip fracture on hospital care costs: a population-based study. *Osteoporos Int*. 2016;27(2):549–558.
60. **Vosoughi AR**, **Emami MJ**, **Pourabbas B**, **Mahdaviazad H**. Factors increasing mortality of the elderly following hip fracture surgery: role of body mass index, age, and smoking. *Musculoskelet Surg*. 2017;101(1):25–29.
61. **Heyes GJ**, **Tucker A**, **Marley D**, **Foster A**. Predictors for 1-year mortality following hip fracture: a retrospective review of 465 consecutive patients. *Eur J Trauma Emerg Surg*. 2017;43(1):113–119.
62. **Hung LW**, **Tseng WJ**, **Huang GS**, **Lin J**. High short-term and long-term excess mortality in geriatric patients after hip fracture: A prospective cohort study in Taiwan. *BMC Musculoskelet Disord*. 2014;15(1):1–8.
63. **Oosterhoff JHF**, **Savelberg ABMC**, **Karhade AV**, **et al**. Development and internal validation of a clinical prediction model using machine learning algorithms for 90 day and 2 year mortality in femoral neck fracture patients aged 65 years or above. *Eur J Trauma Emerg Surg*. 2022;48(6):4669–4682.
64. **Davis FM**, **Woolner DF**, **Frampton C**, **et al**. Prospective, multi-centre trial of mortality following general or spinal anaesthesia for hip fracture surgery in the elderly. *Br J Anaesth*. 1987;59(9):1080–1088.
65. **Elliott J**, **Beringer T**, **Kee F**, **Marsh D**, **Willis C**, **Stevenson M**. Predicting survival after treatment for fracture of the proximal femur and the effect of delays to surgery. *J Clin Epidemiol*. 2003;56(8):788–795.
66. **Jamal Sepah Y**, **Umer M**, **Khan A**, **Ullah Khan Niazi A**. Functional outcome, mortality and in-hospital complications of operative treatment in elderly patients with hip fractures in the developing world. *Int Orthop*. 2010;34(3):431–435.
67. **Hommel A**, **Ulander K**, **Bjorkelund KB**, **Norrman PO**, **Wingstrand H**, **Thorngren KG**. Influence of optimised treatment of people with hip fracture on time to operation, length of hospital stay, reoperations and mortality within 1 year. *Injury*. 2008;39(10):1164–1174.
68. **Muraki S**, **Yamamoto S**, **Ishibashi H**, **Nakamura K**. Factors associated with mortality following hip fracture in Japan. *J Bone Miner Metab*. 2006;24(2):100–104.
69. **Paksima N**, **Koval KJ**, **Aharanoff G**, **et al**. Predictors of mortality after hip fracture: A 10-year prospective study. *Bull NYU Hosp Jt Dis*. 2008;66(2):111–117.
70. **Holvik K**, **Ranhoff AH**, **Martinsen MI**, **Solheim LF**. Predictors of mortality in older hip fracture inpatients admitted to an orthogeriatric unit in oslo, norway. *J Aging Health*. 2010;22(8):1114–1131.
71. **Kalra S**, **Williams A**, **Whitaker R**, **et al**. Subclinical thyroid dysfunction does not affect one-year mortality in elderly patients after hip fracture: A prospective longitudinal study. *Injury*. 2010;41(4):385–387.
72. **Sterling RS**. Gender and race/ethnicity differences in hip fracture incidence, morbidity, mortality, and function. *Clin Orthop Relat Res*. 2011;469(7):1913–1918.
73. **Morri M**, **Ambrosi E**, **Chiari P**, **et al**. One-year mortality after hip fracture surgery and prognostic factors: a prospective cohort study. *Sci Rep*. 2019;9(1):18718.
74. **Heiden JJ**, **Goodin SR**, **Mormino MA**, **et al**. Early ambulation after hip fracture surgery is associated with decreased 30-day mortality. *J Am Acad Orthop Surg*. 2021;29(5):e238–e242.

Author information:

- H. Dijkstra, BSc, PhD Candidate, Department of Orthopaedic Surgery, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands; Department of Trauma Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; Department of Geriatric Medicine, University Medical Center of Groningen, University of Groningen, Groningen, The Netherlands.
- J. H. F. Oosterhoff, MD, Assistant Professor, Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA; Department of Orthopaedic Surgery, Amsterdam Movement Sciences, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands; Department of Engineering Systems and Services, Faculty Technology Policy Management, Delft University of Technology, Delt, Netherlands.
- A. van de Kuit, MSc, Resident Orthopaedic Surgery, Researcher, Department of Orthopaedic Surgery, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands; Department of Trauma Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
- F. F. A. IJpma, MD, PhD, Trauma Surgeon, Department of Trauma Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
- J. H. Schwab, MD, MS, Orthopaedic Surgeon, Department of Orthopaedic Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
- R. W. Poolman, MD, PhD, Professor, Orthopaedic Surgeon, Department of Orthopaedic Surgery, Leiden University Medical Center, Leiden, The Netherlands;

Department of Orthopaedic Surgery, Onze Lieve Vrouw Gasthuis, Amsterdam, The Netherlands.

- S. Sprague, PhD, Associate Professor
- M. Bhandari, MD, PhD, FRCSC, Professor, Orthopaedic Surgeon
Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, Canada; Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada.
- S. Bzovsky, MSc, Statistical Analyst Orthopaedics, Division of Orthopaedic Surgery, Department of Surgery, McMaster University, Hamilton, Canada.
- M. Swiontkowski, MD, Professor, Orthopaedic Surgeon, Department of Orthopaedic Surgery, University of Minnesota, Minneapolis, Minnesota, USA.
- E. H. Schemitsch, MD, FRCSC, Professor, Orthopaedic Surgeon, Department of Surgery, Western University, London, Canada.
- J. N. Doornberg, MD, PhD, Professor, Orthopaedic Trauma Surgeon, Department of Orthopaedic Surgery, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands.
- L. A. M. Hendrickx, MD, Resident Orthopaedic Surgery, Department of Orthopaedic Surgery, Amsterdam Movement Sciences, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands.

Author contributions:

- H. Dijkstra: Methodology, Formal analysis, Investigation, Writing – original draft.
- J. H. F. Oosterhoff: Methodology, Investigation, Writing – original draft, Writing – review & editing.
- A. van de Kuit: Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.
- F. F. IJpma: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- J. H. Schwab: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- R. W. Poolman: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- S. Sprague: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- S. Bzovsky: Design, data acquisition, data analysis, interpretation, writing and revising. Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- M. Bhandari: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- M. Swiontkowski: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- E. Schemitsch: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- J. N. Doornberg: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.
- L. A. M. Hendrickx: Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing.

- H. Dijkstra and J. H. F. Oosterhoff are joint first authors.

Funding statement:

- The authors disclose receipt of the following financial or material support for the research, authorship, and/or publication of this article: The HEALTH trial was supported by research grants from the Canadian Institutes of Health Research (CIHR) (MCT-90168), National Institutes of Health (NIH) (1UM1AR063386-01), ZorgOnderzoek Nederland-medische wetensehappen (ZonMw) (17088.2503), Sophies Minde Foundation for Orthopaedic Research, McMaster Surgical Associates, and Stryker Orthopaedics. The FAITH trial was supported by research grants from the Canadian Institutes of Health Research (MOP-106630 and MCT-87771), National Institutes of Health (1R01AR055267-01A1), Stichting NutsOhra (SNO-T-0602-43), the Netherlands Organisation for Health Research and Development (80-82310-97-11032), and Physicians' Services Incorporated. The FAITH trial was also supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health (award number R01AR055267-01A1). Research reported in this publication was also supported by the County Durham and Tees Valley Comprehensive Local Research Network, which operates as part of the National Institute for Health Research Comprehensive Clinical Research Network in England. Each author certifies that there are no funding or commercial associations (consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted article related to the author or any immediate family members. The FAITH trial was also supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number R01AR055267-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Research reported in this publication was also supported by The County Durham & Tees Valley Comprehensive Local Research Network which operates as part of the National Institute for Health Research Comprehensive Clinical Research Network in England. The funding sources had no role in design or conduct of the study; the collection, management, analysis, or interpretation of the data; or the preparation, review, or approval of the manuscript.

Acknowledgements:

- **Members of the Machine Learning Consortium include:** Britt Barvelink, Joost Colaris, Chris DiGiovanni, Andrew Duckworth, Hamid Ghaednia, Daniel Guss, Merilyn Heng, Sanne Hoeksema, Mike Hogervorst, Gino Kerkhoffs, Charlotte Laane, Koen Oude Nijhuis, Peter van Ooijen, Jacobien Oosterhoff, David Ring, Jospeph Schwab, Sheila Sprague, Vincent Stirler, Matthieu Wijffels.
Members of the HEALTH study include:
Steering Committee: Mohit Bhandari (Chair, McMaster University), P.J. Devereaux (McMaster University), Thomas A. Einhorn (New York University Langone Medical Center), Frede Frihagen (Oslo University Hospital), Ernesto Guerra-Farfán (*Hospital Vall d'Hebrón)*, Kenneth J. Koval (Orlando Regional Medical Centre), Rudolf W. Poolman (OLVG), Emil H. Schemitsch (University of Western Ontario), Kevin Tetsworth (University of Queensland), Lehana Thabane (McMaster University), Stephen D. Walter (McMaster University), Gordon H. Guyatt (McMaster University).
Global Methods Centre: Mohit Bhandari (Principal Investigator); Sheila Sprague (Research Director); Paula McKay (Manager); Kim Madden, Sofia Bzovsky, Kerry Tai, Taryn Scott, Marilyn Swinton (Research Coordination); Naveen Khan (Adjudication Coordination); Diane Heels-Ansdell (Statistical Analysis); Lisa Buckingham, Aravin Duraikannan (Data Management); (McMaster University).
United States Methods Centre: Thomas A. Einhorn (Principal Investigator); Heather Silva (Research Coordination) (Boston University Medical Center). Thomas A. Einhorn, David Novikov, Daniel P. Waren (New York University).
Netherlands Methods Centre: Ydo V. Kleinlugtenbelt, Ellie B.M. Landman (Deventer Ziekenhuis; Definitive). Martin J. Heetveld, Rudolf W. Poolman, Esther M.M. Van Lieshout, Paul T.P.W. Burgers (Erasmus MC, University Medical Center Rotterdam, Rotterdam; Pilot)
United Kingdom Methods Centre: Amar Rangan (The James Cook University Hospital, South Tees Hospitals NHS Foundation Trust, University of Oxford, University of York), Birgit Hanusch, Lucksy Kottam (The James Cook University Hospital, South Tees Hospitals NHS Foundation Trust).
Central Adjudication Committee: Robert D. Zura (Chair, Lousiana State University Health Sciences Center), Victoria Avram (McMaster University), Ajay Manjoo (McMaster University)
Data and Safety Monitoring Board (CIHR): John Antoniou (Chair), Tim Ramsay, Earl R. Bogoch, Andrew Trenholm.
Data and Safety Monitoring Board (NIH): Stephen Lyman (Chair), Madhu Mazumdar, Kevin J. Bozic, Mark Luborsky, Stuart Goodman, Susan Muray, Benjamin K. Potter
Statistical Support: Diane Heels-Ansdell, Lehana Thabane, Nathan N. O'Hara, Gerard P. Slobogean, Quazi Ibrahim.
Members of the FAITH study include:
Writing Committee: Mohit Bhandari, PJ Devereaux, Gordon Guyatt, Lehana Thabane, Stephen D. Walter, Martin J. Heetveld, Kyle J. Jeray, Susan Liew, Emil H. Schemitsch, Paul Tornetta III, Gregory J. Della Rocca, Richard E. Buckley, Robert McCormack, Todd M. Oliver, Michiel J.M. Segers, Amar Rangan, Martin Richardson, Sheila Sprague, Gerard P. Slobogean, Taryn Scott, Julie Agel, Alisha Garibaldi, Qi Zhou, Diane Heels-Ansdell, Helena Viveiros, Stephanie M. Zielinski, Esther M.M. Van Lieshout, Herman Johal, Birgit C. Hanusch, and Marc Swiontkowski.
Steering Committee: Mohit Bhandari (Chair), Marc Swiontkowski, PJ Devereaux, Gordon Guyatt, Martin J. Heetveld, Kyle Jeray, Susan Liew, Martin Richardson, Emil H. Schemitsch, Lehana Thabane, Paul Tornetta III, and Stephen D. Walter.
Global Methods Centre: Mohit Bhandari (Principal Investigator); Sheila Sprague (Research Methodologist); Paula McKay (Manager); Taryn Scott, Alisha Garibaldi, Helena Viveiros, Marilyn Swinton (Research Coordination); Mark Gichuru, Sofia Bzovsky (Adjudication Coordination); Diane Heels-Ansdell, Qi Zhou (Statistical Analysis); Lisa Buckingham, Aravin Duraikannan (Data Management); Deborah Maddock, Nicole Simunovic (Grants Management).
United States Methods Centre: Marc Swiontkowski (Principal Investigator); Julie Agel (Research Coordination).
Netherlands Methods Centre:Martin J. Heetveld (Principal Investigator); Esther M.M. Van Lieshout (Research Coordination); Stephanie M. Zielinski (Trial Coordination).
United Kingdom Methods Centre: Amar Rangan (Principal Investigator); Birgit C. Hanusch, Lucksy Kottam, Rachel Clarkson (Research Coordination).
Adjudication Committee: Gregory J. Della Rocca (Chair), Robert Haverlag, Susan Liew, Gerard P. Slobogean, Kyle Jeray.

Ethical review statement:

- This study is based on secondary analyses of existing datasets from the HEALTH and FAITH trials. Ethical approval was not required for the current study.

Open access funding

- The authors have no support or funding to report for the open access fee of the current study. Funding of the original data collections have been reported in the HEALTH and FAITH trials.