scientific reports

Check for updates

OPEN

Early prediction of functional impairment at hospital discharge in patients with osteoporotic vertebral fracture: a machine learning approach

Soichiro Masuda[1,2], Toshiki Fukasawa[2,3], Shoichiro Inokuchi[4], Bungo Otsuki[1], Koichi Murata[1], Takayoshi Shimizu[1], Takashi Sono[1], Shintaro Honda[1], Koichiro Shima[1], Masaki Sakamoto[1], Shuichi Matsuda[1] & Koji Kawakami[2]✉

Although conservative treatment is commonly used for osteoporotic vertebral fracture (OVF), some patients experience functional disability following OVF. This study aimed to develop prediction models for new-onset functional impairment following admission for OVF using machine learning approaches and compare their performance. Our study consisted of patients aged 65 years or older admitted for OVF using a large hospital-based database between April 2014 and December 2021. As the primary outcome, we defined new-onset functional impairment as a Barthel Index ≤ 60 at discharge. In the training dataset, we developed three machine learning models (random forest [RF], gradient-boosting decision tree [GBDT], and deep neural network [DNN]) and one conventional model (logistic regression [LR]). In the test dataset, we compared the predictive performance of these models. A total of 31,306 patients were identified as the study cohort. In the test dataset, all models showed good discriminatory ability, with an area under the curve (AUC) greater than 0.7. GBDT (AUC = 0.761) outperformed LR (0.756), followed by DNN (0.755), and RF (0.753). We successfully developed prediction models for new-onset functional impairment following admission for OVF. Our findings will contribute to effective treatment planning in this era of increasing prevalence of OVF.

Keywords Osteoporotic vertebral fracture, OVF, Prediction model, Functional impairment, Activities of daily living, Machine learning

The number of patients with osteoporotic vertebral fracture (OVF) has increased with the aging of society, and OVF is now considered a serious public health problem in developed countries[1]. While conservative treatment, such as pain medications and orthosis, is generally effective for OVF, approximately 30% of patients experience a decrease in activities of daily living (ADL)[2–4]. Some of these OVF patients with functional impairment require long-term medical and nursing care, experience reduced quality of life, and have increased mortality rates[5–7]. Recently, surgical treatment aimed at preventing this decline in ADL has become more common[8,9]. While early surgical intervention for OVF can improve clinical outcomes, prediction immediately after OVF injury whether conservative treatment will be effective is difficult[10,11]. Identification of those most likely to develop functional disability following conservative treatment of OVF is therefore particularly important to facilitating clinical decision making and reducing the number of patients with disability following OVF.

Although previous studies have reported several prognostic factors associated with a decrease in ADL[5,12,13], these studies has predominantly focused on identifying radiographic factors and exploring prognostic factors potentially linked to functional outcomes. Furthermore, we are unaware of any attempt to build a prediction

[1]Department of Orthopaedic Surgery, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [2]Department of Pharmacoepidemiology, Graduate School of Medicine and PublicHealth, Kyoto University, Kyoto, Japan. [3]Department of Digital Health and Epidemiology, Graduate School of Medicine andPublic Health, Kyoto University, Kyoto, Japan. [4]Research and Analytics Department, Real World Data Co Ltd., Kyoto, Japan. ✉email: kawakami.koji.4e@kyoto-u.ac.jp

model of functional impairment after OVF possibly due to the small number of patients, the limited study population, and a high risk of bias.

Machine learning models have recently shown promise in improving predictive ability in various conditions[14,15]. These approaches have the advantages of being able to handle more variables and build complex models that take account of interactions between variables and non-linear relationships between variables and outcomes. However, to our knowledge, there is no predictive models using machine learning related to functional outcomes after OVF.

Accordingly, we aimed to develop prediction models for new-onset functional impairment after OVF admission and to validate these models using a temporal test dataset. We used conventional and machine learning approaches and compared the predictive performance of these models.

Results

Study cohort

In the database, we identified 72,691 patients hospitalized for OVF between April 2014 and December 2021. After applying the exclusion criteria, a total of 31,306 patients admitted for OVF were included in the study (Fig. 1). The mean (SD) age was 81.6 (7.7) years and 28.6% were men. Of these, 24,603 (78.6%) were allocated to the training dataset (those admitted between April 2014 and December 2020), while 6703 (21.4%) were allocated to the test dataset (those admitted between January and December 2021).

The primary outcome of a Barthel Index ≤ 60 at discharge was observed in 33.2% (10,385 of 31,306) of patients. Table 1 shows the baseline characteristics and exploratory measures of the patients overall and by outcome status. Patients with a Barthel Index ≤ 60 at discharge tended to be older men, had a lower Barthel Index at admission and were more likely to have mild or moderate dementia than those with a Barthel Index > 60 at discharge. The length of hospital stay was 27.3 (14.9) days in the training dataset and 26.8 (14.9) days in the test dataset. The in-hospital mortality rate was 0.1% (23 of 24,603 in the training dataset and 9 of 6703 in the test dataset) in the two datasets, and discharge to home was achieved in 71.2% (17,522 of 24,603) in the training dataset and 69.4% (4654 of 6703) in the test dataset. Patients with a Barthel Index > 60 at discharge were more likely to be discharged home than those with a Barthel Index ≤ 60 at discharge.

Performance of models in the test dataset

The performance of the different models in the test dataset is shown in Table 2; Fig. 2. All models showed good discriminative ability, with an AUC above 0.7. Compared to the LR (AUC = 0.757 [95% CI 0.746–0.770]), only the GBDT (AUC = 0.760 [95% CI 0.748–0.771]) outperformed it, albeit without statistical significance (Delong's test P-value = 0.22), followed by RF (AUC = 0.756 [95% CI 0.744–0.767]), DNN (AUC = 0.702 [95% CI 0.689–0.715]) (Table 2). While the sensitivities of GDBT and RF were low (0.44), their specificities were high (0.86). As shown in the graph, the calibration of each model was similarly good for low-risk patients with a discharge Barthel Index ≤ 60, but the number of high-risk patients tended to be overestimated (Fig. 3).

Variable importance

The variable importance of the 10 most important predictors for the best-performing model (GBDT) is shown in Fig. 4, suggesting that Barthel Index at admission, dementia level, and age were important predictors.

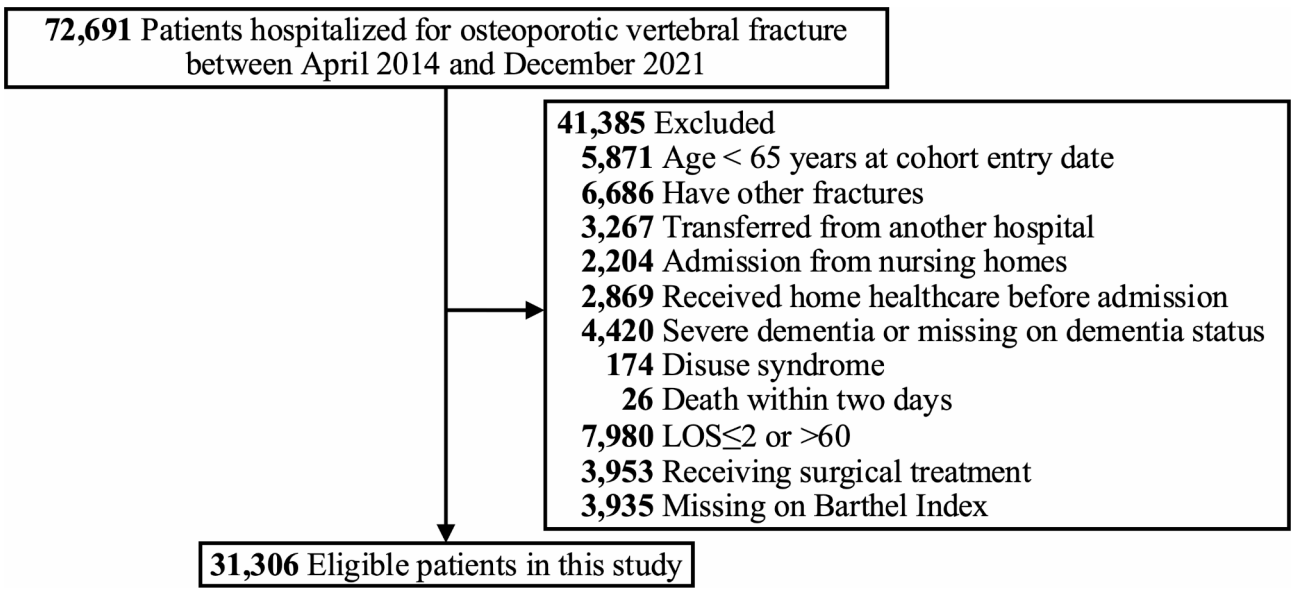

Fig. 1. Flow chart for cohort selection. *LOS* length of hospital stay.

	Training dataset ($n = 24,603$)			Test dataset ($n = 6703$)		
	Overall ($n = 24,603$)	Barthel Index > 60 at discharge ($n = 16,545$)	Barthel Index ≤ 60 at discharge ($n = 8058$)	Overall ($n = 6703$)	Barthel Index > 60 at discharge ($n = 4376$)	Barthel Index ≤ 60 at discharge ($n = 2327$)
Men	6,959 (28.3%)	4,546 (27.5%)	2,413 (29.9%)	1,939 (28.9%)	1,227 (28.0%)	712 (30.6%)
Age, mean (SD), years	81.5 (7.7)	80.2 (7.4)	84.2 (7.5)	82.1 (7.8)	80.7 (7.3)	84.6 (7.9)
Body Mass Index, mean (SD), kg/m^2	21.9 (3.3)	22.1 (3.2)	21.6 (3.3)	21.9 (3.3)	22.0 (3.2)	21.5 (3.3)
Unknown	4739	2837	1902	1329	762	567
Smoking index, mean (SD)	1004.8 (2,850.3)	982.7 (2,815.6)	1050.1 (2,919.9)	1111.2 (3,002.1)	1056.6 (2,922.3)	1213.7 (3,144.6)
Unknown	7	4	3	6	5	1
Barthel Index at admission	38.2 (34.4)	46.8 (35.6)	20.9 (24.0)	35.8 (33.6)	44.0 (35.3)	20.9 (23.9)
Unknown	2,774	1,971	803	645	458	187
Dementia level						
No dementia	17,644 (71.7%)	13,139 (79.4%)	4505 (55.9%)	4738 (70.7%)	3392 (77.5%)	1346 (57.8%)
Level I	4402 (17.9%)	2433 (14.7%)	1969 (24.4%)	1105 (16.5%)	617 (14.1%)	488 (21.0%)
Level II	2557 (10.4%)	973 (5.9%)	1,584 (19.7%)	860 (12.8%)	367 (8.4%)	493 (21.2%)
Ambulance use	11,227 (45.6%)	6627 (40.1%)	4600 (57.1%)	3381 (50.4%)	1935 (44.2%)	1446 (62.1%)
Charlson comorbidity index, mean (SD)	0.5 (1.1)	0.5 (1.1)	0.6 (1.2)	0.6 (1.2)	0.5 (1.2)	0.7 (1.3)
Diabetes mellitus	4191 (17.0%)	2747 (16.6%)	1444 (17.9%)	1198 (17.9%)	751 (17.2%)	447 (19.2%)
Rheumatoid arthritis	707 (2.9%)	472 (2.9%)	235 (2.9%)	162 (2.4%)	118 (2.7%)	44 (1.9%)
Hypertension	9806 (39.9%)	6390 (38.6%)	3416 (42.4%)	2597 (38.7%)	1655 (37.8%)	942 (40.5%)
Hyperlipidemia	4334 (17.6%)	3052 (18.4%)	1282 (15.9%)	1244 (18.6%)	878 (20.1%)	366 (15.7%)
Malignancy	1802 (7.3%)	1140 (6.9%)	662 (8.2%)	554 (8.3%)	345 (7.9%)	209 (9.0%)
Hemodialysis	357 (1.5%)	207 (1.3%)	150 (1.9%)	98 (1.5%)	51 (1.2%)	47 (2.0%)
Acute myocardial infarction	325 (1.3%)	194 (1.2%)	131 (1.6%)	91 (1.4%)	57 (1.3%)	34 (1.5%)
Congestive heart failure	1903 (7.7%)	1080 (6.5%)	823 (10.2%)	553 (8.3%)	309 (7.1%)	244 (10.5%)
Chronic pulmonary disease	1266 (5.1%)	808 (4.9%)	458 (5.7%)	372 (5.5%)	237 (5.4%)	135 (5.8%)
Cerebrovascular disease	2248 (9.1%)	1258 (7.6%)	990 (12.3%)	576 (8.6%)	298 (6.8%)	278 (11.9%)
Non-steroidal anti-inflammatory drugs	11,007 (44.7%)	7775 (47.0%)	3232 (40.1%)	2593 (38.7%)	1801 (41.2%)	792 (34.0%)
Acetaminophen	7344 (29.9%)	4584 (27.7%)	2760 (34.3%)	2707 (40.4%)	1676 (38.3%)	1031 (44.3%)
Tramadol	746 (3.0%)	501 (3.0%)	245 (3.0%)	192 (2.9%)	124 (2.8%)	68 (2.9%)
Steroid	579 (2.4%)	388 (2.3%)	191 (2.4%)	189 (2.8%)	115 (2.6%)	74 (3.2%)
Bisphosphonates	1620 (6.6%)	1132 (6.8%)	488 (6.1%)	440 (6.6%)	288 (6.6%)	152 (6.5%)
Denosumab	16 (0.1%)	9 (0.1%)	7 (0.1%)	2 (0.0%)	1 (0.0%)	1 (0.0%)
Romosozumab	12 (0.0%)	10 (0.1%)	2 (0.0%)	9 (0.1%)	6 (0.1%)	3 (0.1%)
Teriparatide	675 (2.7%)	498 (3.0%)	177 (2.2%)	143 (2.1%)	106 (2.4%)	37 (1.6%)
Vitamin D	3614 (14.7%)	2541 (15.4%)	1073 (13.3%)	1093 (16.3%)	743 (17.0%)	350 (15.0%)
Selective estrogen receptor modulation	449 (1.8%)	342 (2.1%)	107 (1.3%)	167 (2.5%)	116 (2.7%)	51 (2.2%)
Anti-diabetic drugs	2777 (11.3%)	1838 (11.1%)	939 (11.7%)	940 (14.0%)	587 (13.4%)	353 (15.2%)
Antithrombotic drugs	6673 (27.1%)	4213 (25.5%)	2460 (30.5%)	2115 (31.6%)	1305 (29.8%)	810 (34.8%)
Proton pump inhibitors	7165 (29.1%)	4793 (29.0%)	2372 (29.4%)	2299 (34.3%)	1471 (33.6%)	828 (35.6%)
Rehabilitation within 2 days after admission	13,367 (54.3%)	9060 (54.8%)	4307 (53.4%)	4273 (63.7%)	2790 (63.8%)	1483 (63.7%)
Thoracolumbar orthosis	2863 (11.6%)	1994 (12.1%)	869 (10.8%)	1228 (18.3%)	851 (19.4%)	377 (16.2%)
Bed size category						
20–99	599 (2.4%)	399 (2.4%)	200 (2.5%)	323 (4.8%)	190 (4.3%)	133 (5.7%)
100–199	6191 (25.2%)	4476 (27.1%)	1715 (21.3%)	1824 (27.2%)	1281 (29.3%)	543 (23.3%)
200–299	6127 (24.9%)	4469 (27.0%)	1658 (20.6%)	1704 (25.4%)	1203 (27.5%)	501 (21.5%)
300–499	8458 (34.4%)	5380 (32.5%)	3078 (38.2%)	2137 (31.9%)	1270 (29.0%)	867 (37.3%)
≥ 500	3228 (13.1%)	1821 (11.0%)	1407 (17.5%)	715 (10.7%)	432 (9.9%)	283 (12.2%)
Number of hospitalizations, mean (SD)	0.2 (0.7)	0.2 (0.7)	0.2 (0.7)	0.2 (0.8)	0.2 (0.8)	0.2 (0.8)
Number of hospitalizations for osteoporotic vertebral fracture, mean (SD)	0.0 (0.2)	0.0 (0.2)	0.0 (0.2)	0.1 (0.3)	0.1 (0.3)	0.1 (0.3)
Secondary outcomes						
Length of hospital stay, mean (SD), days	27.3 (14.9)	28.1 (14.6)	25.6 (15.2)	26.8 (14.9)	27.8 (14.9)	24.8 (14.7)
In-hospital death	23 (0.1%)	2 (0.0%)	21 (0.3%)	9 (0.1%)	1 (0.0%)	8 (0.3%)
Discharge location						
Home	17,522 (71.2%)	13,954 (84.3%)	3568 (44.3%)	4654 (69.4%)	3718 (85.0%)	936 (40.2%)
Continued						

	Training dataset ($n=24,603$)			Test dataset ($n=6703$)		
	Overall ($n=24,603$)	Barthel Index > 60 at discharge ($n=16,545$)	Barthel Index ≤ 60 at discharge ($n=8058$)	Overall ($n=6703$)	Barthel Index > 60 at discharge ($n=4376$)	Barthel Index ≤ 60 at discharge ($n=2327$)
Another hospital	5880 (23.9%)	2173 (13.1%)	3707 (46.0%)	1761 (26.3%)	570 (13.0%)	1191 (51.2%)
Nursing home	1112 (4.5%)	386 (2.3%)	726 (9.0%)	279 (4.2%)	87 (2.0%)	192 (8.3%)
Others	89 (0.4%)	32 (0.2%)	57 (0.7%)	9 (0.1%)	1 (0.0%)	8 (0.3%)

Table 1. Baseline characteristics of patients.

Model	AUC (95% CIs)	*P* value[a]	Sensitivity	Specificity	PPV	NPV
Gradient-boosting decision tree	0.761 (0.750–0.772)	0.57	0.46	0.85	0.62	0.75
Random forest	0.753 (0.741–0.767)	0.21	0.41	0.87	0.62	0.74
Deep neural network	0.755 (0.744–0.766)	0.45	0.46	0.85	0.62	0.75
Logistic regression	0.756 (0.744–0.768)	Reference	0.46	0.85	0.63	0.75

Table 2. Performance of the five models in the test dataset. [a]P values were calculated with Delong's test using logistic regression as a reference. *AUC* area under curve, *CI* confidence interval, *PPV* positive predictive value, *NPV* negative predictive value.

Discussion

In our study of 31,306 patients hospitalized for OVF who had no previous functional impairment, approximately one-third experienced new-onset functional impairment by the time of discharge. We successfully developed prediction models that achieved high predictive performance using data routinely collected within two days of admission for OVF only. To our knowledge, this is the first study to use conventional and machine learning approaches with a variety of predictors to target new-onset functional impairment after admission for OVF. Key potential risk factors for functional impairment were identified, with Barthel Index at admission, dementia level, ambulance use, and age at OVF highlighted as significant indicators.

Several previous studies have investigated prognostic factors for functional impairment after OVF, including the presence of middle column injury, cognitive decline, and nonunion[12,13,16,17]. Although most of these studies focused on the association of ADL decline with radiographic assessment by computed tomography or magnetic resonance imaging, the universal use of these diagnostic tools for every OVF patient is not feasible, considering cost and the typically favorable prognosis of OVF[12,13,16,17]. In stark contrast, the predictive model developed in our study uses routinely collected information only and does not require any additional testing. Thus, the present study builds on these previous reports and extends them by demonstrating the superior ability of modern machine learning approaches to the prediction of functional outcomes after OVF.

Although GBDT outperformed LR in our results, the difference was negligible. Other machine learning models did not outperform LR. This is consistent with a recent systematic review that reported that there was no statistically significant difference in predictive performance between traditional regression models and machine learning counterparts[18]. Our use of only 37 predictor variables, which is a limited number for machine learning techniques, may have influenced these results. In addition, most of our predictor variables were not continuous. The inclusion of continuous variables that may have non-linear relationships with functional regression, such as blood test results, might improve the performance of machine learning models. This suggests that future research with expanded predictor variables, including continuous variables and clinical information extracted using natural language processing, may improve the predictive capabilities of machine learning approaches[18].

Some of the important predictors selected in this study have provided us with meaningful insights. First, Barthel Index at admission and ambulance use were associated with increased functional impairment. We speculate that patients who experienced a decrease in Barthel Index at admission and used an ambulance experienced a greater severity of pain than those who did not, which is consistent with previous studies[12,19]. These previous studies reported that pain severity was associated with poor ADL prognosis after OVF[12,19]. Next, although we excluded patients with severe dementia, dementia level was nevertheless selected as an important predictor of new-onset functional impairment. Patients with dementia are generally less compliant with treatment, such as bed rest, orthosis, and rehabilitation; poor adherence might have contributed to functional impairment after OVF hospitalization[16].

The implications of our findings are profound, particularly for clinicians, patients, and their families. Most older adults with OVF expect to return to their pre-fracture functional level with bed rest, thoracolumbar orthosis, rehabilitation and, in some cases, surgery. Recently, a variety of surgical treatments have been used to treat OVF when conservative treatment is not effective in preventing a decline in ADL, ranging from vertebroplasty to corrective spinal fusion surgery[20]. However, these surgical treatments are associated with increased risk of complications, especially in the elderly[20,21]. Our results may assist clinicians in risk stratification and decision-making regarding treatment strategies for patients with OVF, and are therefore of critical importance to both clinicians and patients.

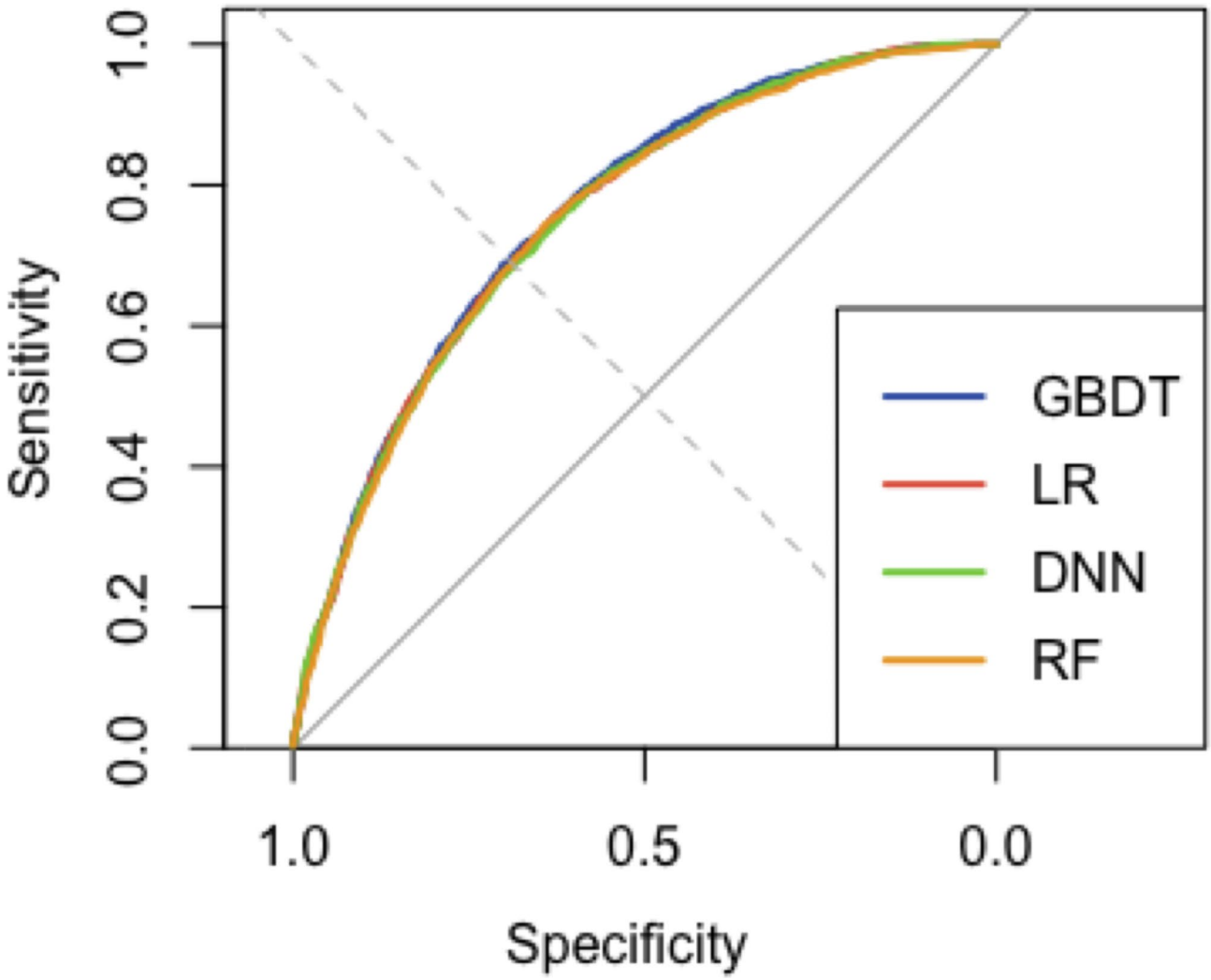

Fig. 2. Receiver operating characteristic curve of each model for the test dataset. *GBDT* gradient-boosting decision tree, *LR* logistic regression, *RF* random forest, *DNN* deep neural network, *AUC* area under curve.

This study has several limitations. First, information on radiographic findings was not available. As radiographic findings are associated with the prognosis of OVF[12,13,16,17], adding these variables to our prediction models may improve their performance. Future prospective study with radiographic data will be needed to improve our prediction models. Second, we focused on ADL at discharge, which has been associated with poor future ADL, higher rates of adverse events, and increased mortality[22,23]. Nevertheless, the true long-term outcomes remain unknown. Furthermore, different discharge policies between participating hospitals would likely have led to misclassification of outcomes. Third, although the methods of conservative treatment for OVF may have varied between hospitals[3], we could not account for clustering of patients within hospitals due to the lack of this type of information in our database. Fourth, the comorbidities we assessed were underestimated because only a maximum of four comorbidities can be recorded in the DPC database. In addition, the number of hospitalizations was also underestimated because our database did not include information on treatment history at other institutions. Fifth, our findings may not be generalizable to other settings. Although we assessed the performance of our prediction models using the test dataset, we were not able to validate the models using external datasets. Cultural or health system differences, such as long hospital stays, may hinder the application of our predictive models to other patients with OVF. Nevertheless, the variables used in our prediction models are common in other countries. Data from Japan, which is aging faster than the rest of the world, should be useful for other countries which will shortly experience similar levels of aging. Sixth, although we attempted to exclude patients with dependent ADL prior to admission for OVF, some of these patients may have been included as eligible due to misclassification.

In conclusion, using a large hospital-based database, we developed and validated machine learning models to predict the risk of new-onset functional impairment after OVF admission using predictor variables that are commonly available within two days of OVF admission. Our model may be useful for clinicians in identifying OVF patients at high risk of functional impairment, and in considering alternative treatment plans for these patients in place of conservative treatment.

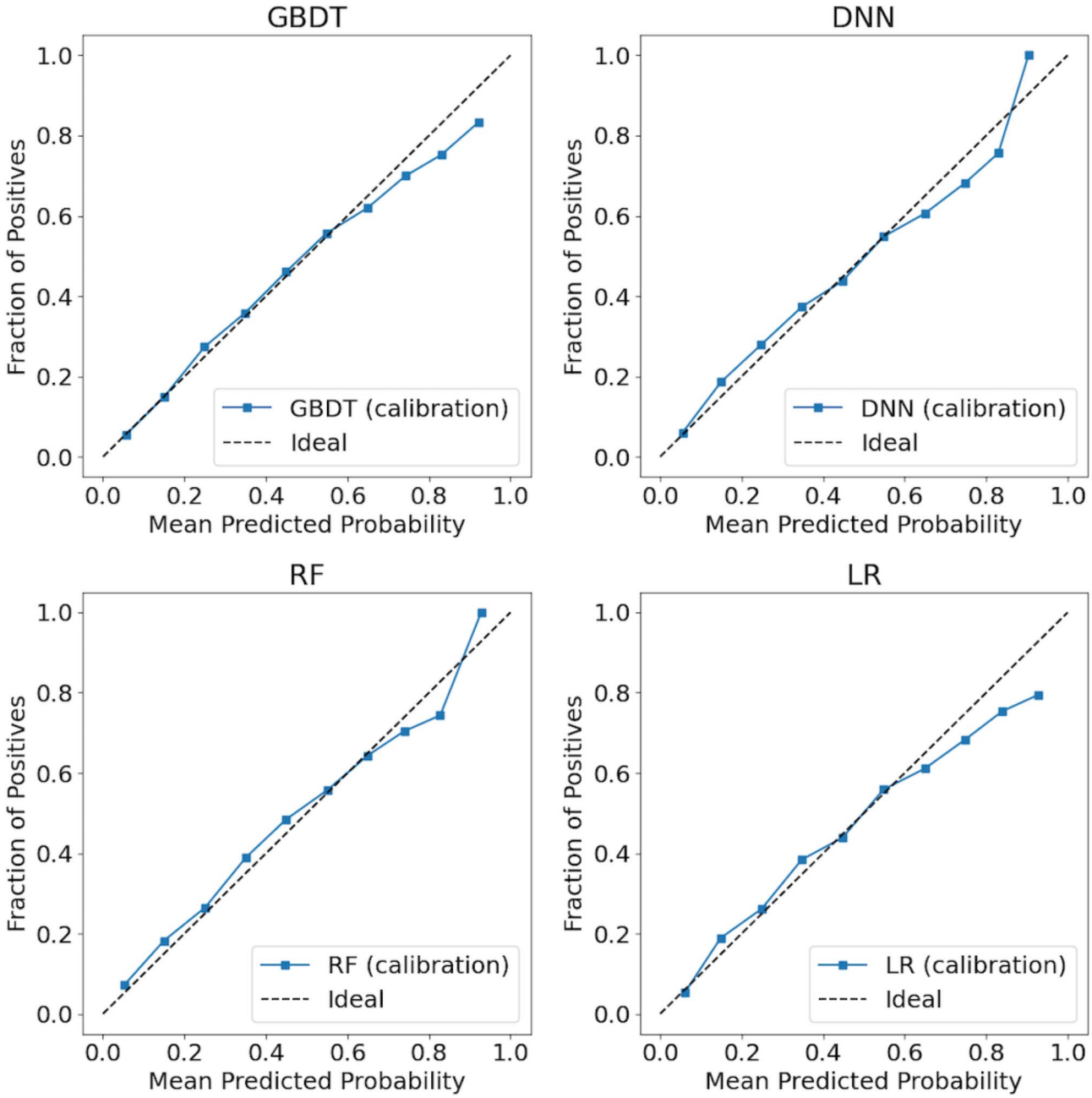

Fig. 3. Calibration plot of each model for the test dataset. *GBDT* gradient-boosting decision tree, *LR* logistic regression, *RF* random forest, *DNN* deep neural network.

Methods

The study was approved by the ethical committee of Kyoto University Graduate School and Faculty of Medicine (No. R4191) and performed in accordance with the Declaration of Helsinki. The requirement for individual informed consent was waived because the database is provided with all patient information anonymized by the ethical committee of Kyoto University Graduate School and Faculty of Medicine. We followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) as a reporting guideline[24].

Study design and setting

We conducted a prognostic study to develop a prediction model using a large hospital administrative database provided by JMDC Inc. (Tokyo, Japan)[25]. As of 2022, the database contained Diagnosis Procedure Combination (DPC) discharge summaries and administrative claims data on more than 18 million patients treated in 573 hospitals. The DPC-based Per Diem Payment system is the main medical system for acute inpatient care reimbursement in Japan[26]. Details of this database and the DPC system have been described elsewhere[25,27].

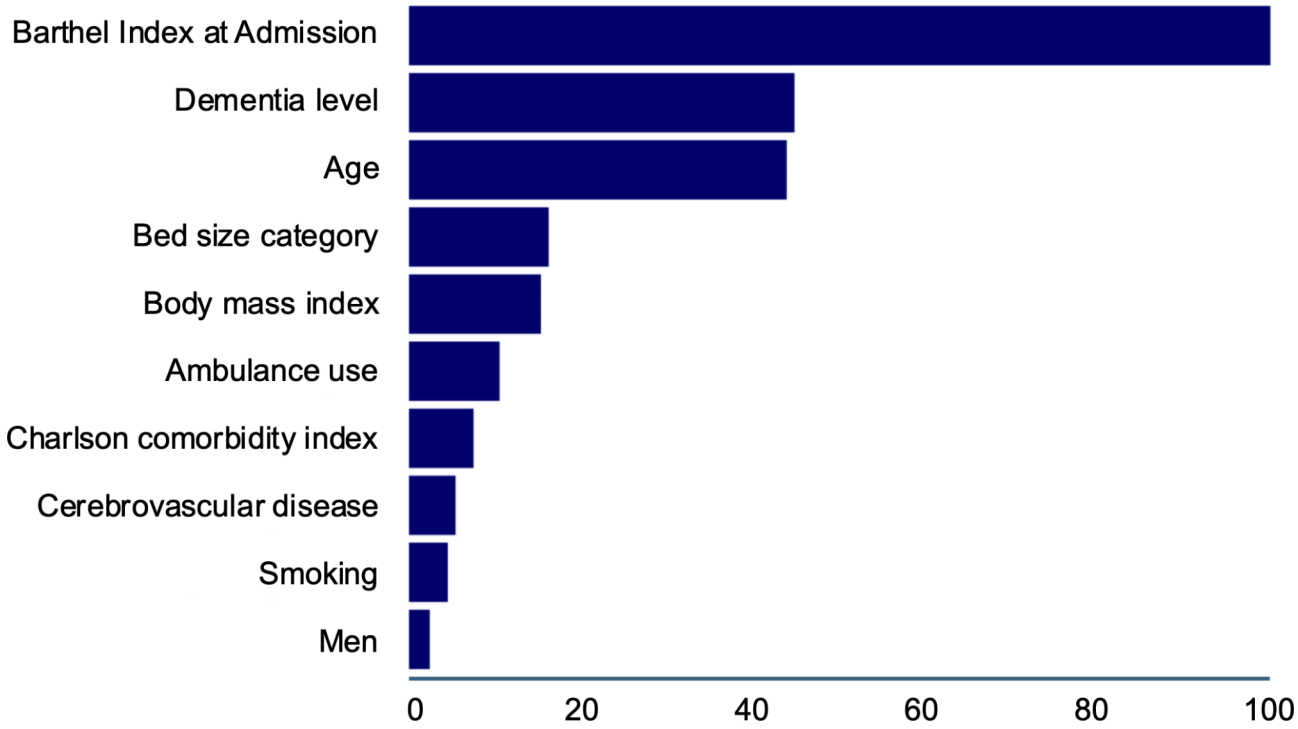

Fig. 4. Variable importance of the 10 most important predictors in the gradient-boosting decision tree model.

Previous validation studies have reported high specificity and moderate sensitivity for the recorded diagnoses and high specificity and sensitivity for the recorded procedures[28,29].

Study population

We identified all patients aged 65 years or older who were admitted for OVF during the period from April 2014 to December 2021. In cases where an individual patient had multiple admissions, each was treated as a separate admission, as the frequency of hospitalization for OVF was considered an indicator of osteoporosis severity. Exclusion criteria included the presence of concurrent fractures and transfer from another medical facility. Given that the aim of the study was to predict new-onset functional impairment after admission for OVF, we excluded patients who were functionally impaired prior to admission for OVF, namely those admitted from nursing homes, receiving home medical care prior to admission, with severe dementia requiring assistance with ADLs, and diagnosed with disuse syndrome prior to admission[14]. We also excluded patients who died within two days of the index admission, those with a hospital stay of less than three days or more than sixty days, those who underwent surgery for OVF during the hospitalization, and those with missing data on Barthel Index at the time of discharge. The definitions of the inclusion and exclusion criteria are given in detail (see Supplementary Table 1 online).

Outcomes

The outcome of interest was defined as functional impairment, characterized by a Barthel Index ≤ 60 at hospital discharge[14,30]. This threshold was decided on the basis of previous studies in which a Barthel Index ≤ 60 was considered equivalent to physical disability[31,32]. Details of the Barthel Index have been provided elsewhere[31,32]. Briefly, the Barthel Index quantifies the performance of ten basic ADLs (bowel, bladder, grooming, toileting, feeding, transferring, mobility, dressing, stair climbing, and bathing), with a total score between 0 and 100[33]. If patients died in hospital after the first two days of the index admission, a Barthel Index of 0 was assigned[32]. We also assessed in-hospital mortality, length of hospital stay, and discharge destination for exploratory purposes.

Predictors

Based on previous studies and clinical expertise[3,5,12,13], potential predictors were selected from data collected within the first two days of admission for OVF. These predictors included age, sex, body mass index (BMI), smoking index, Barthel Index at admission, ambulance use, level of dementia, presence of comorbidities, pharmacological treatment within two days of admission for OVF, use of orthoses, receipt of rehabilitation within two days of admission for OVF, number of admissions in the year prior to the index admission, number of OVF admissions prior to the index admission, and number of hospital beds. Detailed definitions of these predictors are given (see Supplementary Table 2 online).

Missing data

BMI, smoking index, and Barthel Index at admission had missing values in the dataset. We imputed the missing values for these variables using K-nearest neighbor imputation[34].

Statistical analysis

Training and test cohorts

Patients were divided into two cohorts: those admitted between April 2014 and December 2020 served as the training dataset to develop the prediction models, while those admitted between January and December 2021 formed the test dataset to evaluate the performance of the established models[35,36]. We did not calculate the required sample size and used all available data because a larger sample size leads to the development of more robust models[37].

Development of prediction models in the training cohort

We chose the logistic regression (LR) model as reference to evaluate the predictive ability of machine learning models. Three machine learning models were used to develop the prediction models in our study: random forest (RF), gradient-boosting decision tree (GBDT), and deep neural network (DNN). Details of these machine learning models are described elsewhere[14,38]. In brief, the RF approach creates a forest of decision trees. Each tree is built using a random subset of the data, and the final prediction is based on the consensus of all trees. The GBDT is similar to the RF in that this model sequences multiple decision trees. With each new tree, however, the model focuses on correcting the mistakes made by the previous trees. So instead of building a forest of trees all at once (as in RF), it builds and refines them one at a time. The DNN is made up of multiple interconnected neurons organized in layers, inspired by the way our brain works.

Continuous values were scaled to range between 0 and 1. Categorical variables with more than two categories were converted to multiple binary variables (one-hot encoding) for the machine learning models.

These machine learning models were developed using the training dataset. During this development, we adjusted hyperparameters to achieve the best performance. For each model, we tried five random levels of hyperparameter during the training step. Once trained on a randomly sampled 80% of the training dataset, these models were validated on the remaining 20% of the training dataset, which was 20% of our original training dataset. This process was repeated 10 times (a method called repeated random sub-sampling cross-validation)[39]. The model that achieved the largest area under the curve (AUC) during training was selected as the best model. The use of cross-validation and hyperparameter tuning for internal validation is considered a robust method for evaluating models before testing them on an external validation dataset and maximizes the potential performance of the models[39].

Evaluating the predictive performance of each model in the test dataset

The performance of each model was assessed using the AUC, predictive measures (i.e., sensitivity, specificity, positive predictive value, and negative predictive value), and the calibration curve on the test dataset.[40] 95% confidence intervals (CIs) for the AUC were obtained using a bootstrapping method. We compared the AUC between the models using Delong's test.

To understand how each predictor contributed to the performance of our best model (the one with the highest AUC), we evaluated their importance. These measures were adjusted so that the predictor with the highest importance was always marked as 100.

All statistical analyses were conducted using SAS version 9.4 (SAS Institute), R 4.2.1 (R foundation), and Python 3.11.0 (Python Software Foundation). The Python libraries we used were pandas, numpy, scikit-learn, scipy, xgboost, keras, tensorflow, and matplotlib.

Data availability

The data that support the findings of this study are available from JMDC Inc. but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the corresponding author (Koji Kawakami, kawakami.koji.4e@kyoto-u.ac.jp) upon reasonable request and with permission of JMDC Inc.

Received: 30 August 2024; Accepted: 4 December 2024
Published online: 28 December 2024

References

1. Klotzbuecher, C. M., Ross, P. D., Landsman, P. B., Abbott, T. A., Berger, M. & 3rd & Patients with prior fractures have an increased risk of future fractures: a summary of the literature and statistical synthesis. *J. Bone Miner. Res.* **15**, 721–739 (2000).
2. Caitriona, C. et al. Management of hospitalised osteoporotic vertebral fractures. *Arch. Osteoporos.* **15**, 14 (2020).
3. Funayama, T. et al. Therapeutic effects of conservative treatment with 2-week bed rest for osteoporotic vertebral fractures: a prospective cohort study. *J. Bone Jt. Surg. Am.* **104**, 1785–1795 (2022).
4. Lips, P. & van Schoor, N. M. Quality of life in patients with osteoporosis. *Osteoporos. Int.* **16**, 447–455 (2005).
5. Hoshino, M. et al. Impact of initial conservative treatment interventions on the outcomes of patients with osteoporotic vertebral fractures. *Spine* **38**, E641–E648 (2013).
6. Gold, L. S. et al. Mortality among older adults with osteoporotic vertebral fracture. *Osteoporos. Int.* **34**, 1561–1575 (2023).
7. Pron, G., Hwang, M., Smith, R., Cheung, A. & Murphy, K. Cost-effectiveness studies of vertebral augmentation for osteoporotic vertebral fractures: a systematic review. *Spine J.* **22**, 1356–1371 (2022).
8. Kobayashi, K., Ando, K., Nishida, Y., Ishiguro, N. & Imagama, S. Epidemiological trends in spine surgery over 10 years in a multicenter database. *Eur. Spine J.* **27**, 1698–1703 (2018).

9. Bigdon, S. F. et al. Epidemiologic analysis of 8000 acute vertebral fractures: evolution of treatment and complications at 10-year follow-up. *J. Orthop. Surg. Res.* **17**, 270 (2022).
10. Minamide, A. et al. Early versus delayed kyphoplasty for thoracolumbar osteoporotic vertebral fractures: the effect of timing on clinical and radiographic outcomes and subsequent compression fractures. *Clin. Neurol. Neurosurg.* **173**, 176–181 (2018).
11. Takahashi, S. et al. Differences in short-term clinical and radiological outcomes depending on timing of balloon kyphoplasty for painful osteoporotic vertebral fracture. *J. Orthop. Sci.* **23**, 51–56 (2018).
12. Inose, H. et al. Factors affecting the quality of life in the chronic phase of thoracolumbar osteoporotic vertebral fracture managed conservatively with a brace. *Spine J.* **23**, 425–432 (2023).
13. Matsumoto, T. et al. Prognostic factors for reduction of activities of daily living following osteoporotic vertebral fractures. *Spine* **37**, 1115–1121 (2012).
14. Ohbe, H., Goto, T., Nakamura, K., Matsui, H. & Yasunaga, H. Development and validation of early prediction models for new-onset functional impairment at hospital discharge of ICU admission. *Intensive Care Med.* **48**, 679–689 (2022).
15. Raita, Y. et al. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit. Care.* **23**, 64 (2019).
16. Takahashi, S. et al. Risk factors for cognitive decline following osteoporotic vertebral fractures: a multicenter cohort study. *J. Orthop. Sci.* **22**, 834–839 (2017).
17. Tsujio, T. et al. Characteristic radiographic or magnetic resonance images of fresh osteoporotic vertebral fractures predicting potential risk for nonunion: a prospective multicenter study. *Spine* **36**, 1229–1235 (2011).
18. Mahmoudi, E. et al. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ* **369**, m958 (2020).
19. Kawaguchi, S. et al. Symptomatic relevance of intravertebral cleft in patients with osteoporotic vertebral fracture. *J. Neurosurg. Spine.* **13**, 267–275 (2010).
20. Murata, K. et al. The factors related to the poor ADL in the patients with osteoporotic vertebral fracture after instrumentation surgery. *Eur. Spine J.* **29**, 1597–1605 (2020).
21. Oichi, T. et al. Can Elective spine surgery be performed safely among nonagenarians? Analysis of a National Inpatient database in Japan. *Spine* **44**, E273–E281 (2019).
22. Uemura, Y. et al. Prognostic impact of the preservation of activities of daily living on post-discharge outcomes in patients with acute heart failure. *Circ. J.* **82**, 2793–2799 (2018).
23. Sato, M. et al. Decreased activities of daily living at discharge predict mortality and readmission in elderly patients after cardiac and aortic surgery: a retrospective cohort study. *Medicine* **100**, e26819 (2021).
24. Moons, K. G. M. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* **162**, W1–73 (2015).
25. Nagai, K. et al. Data resource profile: JMDC claims databases sourced from Medical Institutions. *J. Gen. Fam. Med.* **21**, 211–218 (2020).
26. Hayashida, K., Murakami, G., Matsuda, S. & Fushimi, K. History and profile of diagnosis procedure combination (DPC): development of a real data collection system for acute inpatient care in Japan. *J. Epidemiol.* **31**, 1–11 (2021).
27. Masuda, S. et al. Incidence of surgical site infection following lateral lumbar interbody fusion compared with posterior/transforaminal lumbar interbody fusion: a propensity score-weighted study. *Spine* **48**, 901–907 (2023).
28. Yamana, H. et al. Validity of diagnoses, procedures, and laboratory data in Japanese administrative data. *J. Epidemiol.* **27**, 476–482 (2017).
29. Konishi, T. et al. Validity of operative information in Japanese administrative data: a chart review-based analysis of 1221 cases at a single institution. *Surg. Today.* **52**, 1484–1490 (2022).
30. Ng, J. P. H., Ho, S. W. L., Yam, M. G. J. & Tan, T. L. Functional outcomes of patients with schizophrenia after hip fracture surgery: a 1-year follow-up from an institutional hip fracture registry. *J. Bone Joint Surg. Am.* **103**, 786–794 (2021).
31. Sulter, G., Steen, C. & De Keyser, J. Use of the Barthel index and modified Rankin scale in acute stroke trials. *Stroke* **30**, 1538–1541 (1999).
32. Uyttenboogaart, M., Stewart, R. E., Vroomen, P. C. A. J., De Keyser, J. & Luijckx, G. J. Optimizing cutoff scores for the Barthel index and the modified Rankin scale for defining outcome in acute stroke trials. *Stroke* **36**, 1984–1987 (2005).
33. Mahoney, F. I. & Barthel, D. W. Functional evaluation: the barthel index. *Md. State Med. J.* **14**, 61–65 (1965).
34. Emmanuel, T. et al. A survey on missing data in machine learning. *J. Big Data.* **8**, 140 (2021).
35. Vollmer, S. et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* **368**, l6927 (2020).
36. Kamran, F. et al. Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. *BMJ* **376**, e068576 (2022).
37. Riley, R. D. et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* **368**, m441 (2020).
38. Sachiko, O. & Tadahiro, G. Introduction to supervised machine learning in clinical epidemiology. *Ann. Clin. Epidemiol.* **4**, 63–71 (2022).
39. Steyerberg, E. W. Validation in prediction research: the waste by data splitting. *J. Clin. Epidemiol.* **103**, 131–133 (2018).
40. Alba, A. C. et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* **318**, 1377–1384 (2017).

Acknowledgements

We thank Dr. Guy Harris DO of Dmed (https://www.dmed.co.jp) for his support with the writing of the manu script.

Author contributions

Conceptualization, S.Masuda. and T.F.; methodology, S.Masuda, T.F., S.I.; formal analysis and investigation, S.M. and T.F.; writing—original draft preparation, S.Masuda. and T.F.; writing—review and editing, B.O., K.M., T.S., T.S., S.H., K.S., M.S., S.Matsuda., K.K.; resources, K.K.; supervision, S.Matsuda., K.K.; All authors have read and agreed to the published version of the manuscript.

Competing interests

Each author certifies that Toshiki Fukasawa has been employed by the Department of Digital Health and Epidemiology with support from Eisai Co., Ltd. and Kyowa Kirin Co., Ltd.; and has received consulting fees from Real World Data Co., Ltd. and has or may receive payments or benefits, in any one year, an amount less than USD 10,000, from EPS Corporation and Research Institute of Healthcare Data Science (RIHDS) and Koji Kawakami has received research support funding from: Eisai Co., Ltd., Kyowa Kirin Co., Ltd., OMRON Corporation, and Toppan Inc.; consulting fees from Advanced Medical Care Inc., JMDC Inc., and Shin Nippon

Biomedical Laboratories Ltd.; executive compensation from Cancer Intelligence Care Systems, Inc.; and honoraria from Chugai Pharmaceutical Co., Ltd., and Pharma Business Academy. All other authors declare no conflicts of interest associated with this manuscript.

Ethics declarations

This study has been approved by the Ethics Committee of Kyoto University Graduate School and Faculty of Medicine (No. R4191).

Each author certifies that Toshiki Fukasawa has been employed by the Department of Digital Health and Epidemiology with support from Eisai Co., Ltd. and Kyowa Kirin Co., Ltd.; and has received consulting fees from Real World Data Co., Ltd. and has or may receive payments or benefits, in any one year, an amount less than USD 10,000, from EPS Corporation and Research Institute of Healthcare Data Science (RIHDS) and Koji Kawakami has received research support funding from: Eisai Co., Ltd., Kyowa Kirin Co., Ltd., OMRON Corporation, and Toppan Inc.; consulting fees from Advanced Medical Care Inc., JMDC Inc., and Shin Nippon Biomedical Laboratories Ltd.; executive compensation from Cancer Intelligence Care Systems, Inc.; and honoraria from Chugai Pharmaceutical Co., Ltd., and Pharma Business Academy. All other authors declare no conflicts of interest associated with this manuscript.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-82359-x.

Correspondence and requests for materials should be addressed to K.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.