HUMAN BIOLOGY
2024, VOL. 94, ISSUE 6
ORIGINAL ARTICLE

External Validation of a Prognostic Model for One-Year Survival After Fragility Hip Fracture: A Retrospective Cohort Study

Hairui Fu[1], Feixiong Li[1], Bin Liang[1], Dou Wu[2], Xuan Zhang[3]*

ABSTRACT

Purpose: To validate a prognostic model of mortality among patients with fragility hip fractures.

Methods: This was a retrospective cohort study. Patients with fragility hip fractures were consecutively admitted to the orthopaedics department from January 2016 to October 31, 2021. We evaluated the performance of a survival after fragility hip fracture (SFHF) model (including the full model and the simplified model) in the following ways: (1) Discrimination. We present the concordance (c) index of the model, including Harrell's c-index and Uno's c-value. Overall performance was measured with Nagelkerke R2 values. (2) Calibration. The calibration plot method was used to evaluate the calibration of the model. (3) Clinical value. Decision curve analysis (DCA) was used to determine whether the model had clinical value in the validation population.

Results: A total of 877 (out of 1132) eligible patients with fragility hip fractures (≥50 years) were included in this study. Among these patients, 47 patients were lost to follow-up. Among the patients who were successfully followed up, 87 died within 1 year after fracture. After simple imputation was applied to address missing values, the final effective sample size was 93 patients. The 1-year mortality rate after fracture was 10.6%. The Harrell's c-index values of the full and simple SFHF models were 0.764 (standard error, 0.024) and 0.763 (0.024), respectively. Uno's c-values were 0.765 (0.024) and 0.763 (0.024), respectively. The Nagelkerke R2 values were 0.144 and 0.144, respectively. The calibration plot revealed good calibration between the predicted and actual values of the model. DCA revealed that the model was clinically useful within a risk of death threshold range of 0.03-0.38.

Conclusion: Our study preliminarily confirmed that the SFHF model has good accuracy and generalizability in predicting the one-year survival rate of patients with fragility hip fractures and that it has good clinical value. This predictive model may be considered for use in clinical practice.

INTRODUCTION

Background and Objectives

For the purpose of scientifically stratifying the management of hip fractures in elderly individuals and simultaneously identifying high-risk patients to guide joint decision-making by doctors and patients about whether to operate, we previously developed a prognostic model to predict 1-year survival after fragility hip fracture (SFHF) Fu et al. (2021). This prognostic Cox proportional hazards model was developed on the basis of a retrospective cohort study that was conducted at a secondary care hospital in China. Among the 735 eligible patients with fragility hip fractures who were included in the study, 68 died within 1 year after fracture. The simplified version of the model includes 8 preoperative clinical indicators (age, albumin levels, sex, serum creatinine levels, malignancy, hypertension, ability to live independently, and cardiovascular and cerebrovascular diseases) that can be quickly determined within 48 hours after admission; the addition of whether to choose surgical treatment increases the number of predictor variables to 9. With the use of the nomogram derived from this model, a weighted score for each variable can be quickly obtained after patient admission. On the basis of the total score, the 1-year survival rate of the patient following surgery (surgery=1) or conservative treatment (surgery=0) can be simultaneously predicted before a treatment plan is formulated. The model shows good discriminative ability (Harrell's c statistic=0.814 (95% confidence interval (CI) 0.762–0.865)) and no significant overfitting (c=0.795, internal validation by 1000 bootstrap repetitions).

[1]Department of Orthopaedics, Fenyang Hospital Affiliated with Shanxi Medical University, Fenyang Hospital of Shanxi Province, No. 186 Shengli Street, Fenyang City, Lvliang City, Shanxi Province, China
[2]Department of Orthopaedics, The Third Hospital of Shanxi Medical University (Shanxi Bethune Hospital, Shanxi Academy of Medical Sciences, Tongji Shanxi Hospital), No. 99, Longcheng Street, Taiyuan City, Shanxi Province, China
[3]Central Office, Shanxi Provincial Centre for Disease Control and Prevention, No. 8 Xiaonanguan, Shuangta West Street, Taiyuan City, Shanxi Province, China

Address for Correspondence to: Hairui Fu, Department of Orthopaedics, Fenyang Hospital Affiliated with Shanxi Medical University, Fenyang Hospital of Shanxi Province, No. 186 Shengli Street, Fenyang City, Lvliang City, Shanxi Province, China. E-mail: hairuifu1983@163.com.

Keywords: External validation, Prognostic model, Fragility hip fracture, Survival

However, if the model is ultimately to be used for clinical decision-making, external validation is essential Altman et al. (2000), Moons et al. (2009), Royston et al. (2013), Ramspek et al. (2020). For this purpose, we designed and conducted another retrospective cohort study to examine the accuracy and generalizability of this model in a relatively related new population Ramspek et al. (2020), Justice et al. (1999).

In recent years, the development of predictive models that focus on the 1-year mortality (or 1-year survival) rates of fragility hip fracture patients after fracture has not been uncommon. The external validation of multiple similar models on the same dataset can provide potential model users with more information, such as information about which model seems to perform better. However, here, we did not perform such a validation study for a few reasons. 1. The purposes and contexts of model development differ, and our own developed models are best suited to our own local environment. Therefore, we mostly focused on the external validation of the performance of our model. This is probably the reason why more model validation research is conducted by developers themselves. 2. This study was specifically designed to validate our own model. Therefore, in terms of predictor definitions, outcome definitions, and data measurement and collection, the heterogeneity between these validation data and our own development data is naturally minimal compared with the development data of other models. In the presence of such systematic biases, results of objective model comparison studies cannot be obtained. This is a fact Ramspek et al. (2020), Trevisan et al. (2021). 3. A prospective cohort study design is more appropriate for a model comparison study, since certain predictors can be collected specifically for the target model. However, this was a retrospective cohort study, and variables such as “poor handgrip strength” (8), “not being able to drive” (9), and “difficulty preparing meals” (9) could not be collected. As a result, the validation of models containing these predictors could not be achieved. 4. Because of the different contexts in which similar models were developed, some of the predictors in these models are not meaningful in certain contexts. For example, “long-term care residence” Jiang et al. (2005) and “living in an institution” van de Ree et al. (2020) are meaningless in most areas of China because it is not common for elderly individuals to live in nursing facilities. In China, most elderly people live alone or with a spouse or children Cui et al. (2022), Bao et al. (2022), Nie et al. (2022), Zhang et al. (2006). Therefore, it does not make sense to validate such a model. According to a reliable survey that was conducted in 2018, the vast majority of elderly people in China (96.4%) lived with family members, while a small percentage lived alone (3.3%) or in nursing homes (0.3%) Cui et al. (2022), Center for Healthy Aging and Development Studies et al. (2020). Therefore, this study validates only our self-developed SFHF model.

MATERIALS AND METHODS

This study followed the 2015 TRIPOD recommendations Moons et al. (2015).

Data Source

We designed and conducted a single-centre, retrospective longitudinal cohort study. We predicted the survival of each patient in this cohort via the SFHF model and compared it with actual observations to determine whether the model is a good tool for predicting 1-year survival after fracture in these individuals. When necessary, the model was updated to better fit this population. This study included all patients with hip fractures who were consecutively admitted to the Orthopaedics Department of Shanxi Bethune Hospital in China from January 1, 2016, to October 31, 2021. Patient follow-up was then completed via telephone interviews from March 14 to 29, 2022. The preoperative individual characteristics of each patient were obtained from the patient's medical records. This study was approved by the Medical Ethics Committee of Shanxi Bethune Hospital. Informed consent was waived by the Medical Ethics Committee of Shanxi Bethune Hospital. All the methods in this study followed the Declaration of Helsinki.

Participants

Shanxi Bethune Hospital is located in Taiyuan, which is the capital city of Shanxi Province in Central China. Shanxi Bethune Hospital is a comprehensive teaching hospital and a tertiary regional referral hospital. As in the developmental study, we identified eligible participants according to the following criteria: 1. patients ≥ 50 years old; 2. patients with low-energy fractures (fractures that were caused by a patient falling from a standing height or lower), excluding patients with pathological, high-energy fractures; periprosthetic fractures after previous hip replacement surgery; and reoperation due to failure of internal fixation for hip fracture, regardless of whether the patient had a primary or secondary fracture; and 3. patients with hip fractures, including femoral neck, intertrochanteric and subtrochanteric fractures.

Unlike in the developmental study, eligible patients could also have other low-energy fractures, such as wrist fractures, femoral neck fractures, or vertebral compression fractures, resulting from falls that lead to hip fractures. For the same patient who was hospitalized twice for hip fractures on different sides during the study period, the most recent hospitalization data were selected.

Surgical treatment was considered for all patients after admission, and routine preoperative preparations were completed as soon as possible. If a patient suffered from obvious medical diseases that affected eligibility to undergo surgical treatment, surgery was postponed, and the relevant departments were requested to assist with diagnosis and treatment. Surgical treatment was administered when a patient's overall condition was stable. In the developmental study, participants underwent osteodistraction immediately after hospitalization unless it was determined that surgery would be performed soon. In this study population, osteodistraction was not performed unless it was clear that a patient could not undergo surgery in the short term or that surgery was not being considered.

Similar to the developmental study, different surgical approaches were implemented depending on the fracture site. For femoral neck fractures, cannulated screw fixation and total hip or hemihip replacement were selected according to different patient age ranges. Intramedullary nails were used in the surgical treatment of intertrochanteric and subtrochanteric fractures.

All eligible patients were included in this study. For individuals with missing data, we used advanced statistical methods to impute to avoid serious bias due to simple deletion.

Outcome

Similar to the developmental study, our event of interest was all-cause mortality within 1 year (365 days) after fracture. We first collected the demographic characteristics and preoperative clinical indicators of each patient, and follow-up was conducted by telephone interviews.

Similar to the developmental study, we also developed an interview strategy to increase the interview rate and reduce loss to follow-up Fu et al. (2021).

Predictors

Similar to the developmental study, we collected 24 individual patient indicators. Except for whether surgery was performed, the remaining 23 variables were all preoperative characteristics. These variables included demographic characteristics, such as age, sex and medical insurance; fracture-related characteristics, such as fracture site, fracture type, days from fracture to hospitalization, days from hospitalization to surgery, and length of stay (LOS); medical history information, such as ability to live independently (ALI), lung disease (LD), cardiovascular and cerebrovascular disease (CCD), kidney disease (KD), malignancy (MAL), hypertension (HYP), diabetes, and mean arterial pressure (MAP); laboratory-related factors, such as partial pressure of oxygen (PaO2), haemoglobin (Hb) , serum creatinine (SC), fasting blood sugar (BS), albumin (ALB), and total protein (TP) levels; and treatment-related factors, such as osteodistraction and surgery (SUR).

The definition and measurement time of each indicator can be found in the developmental study Fu et al. (2021). Here, we describe only the variables that differ from those that were used in developmental studies. In this study, we redefined the CCD characteristic. In the developmental study, CCD-positive patients included patients with previous diagnoses of myocardial infarction, cerebral infarction, cerebral haemorrhage, or extremity thrombosis or patients who were previously undiagnosed but were identified as having an infarct or thrombosis during the admission examination. In this study, to better capture patients at high risk of cardiovascular and cerebrovascular diseases, we also considered patients with coronary heart disease who were not diagnosed with myocardial infarction as being positive for CCD. Regarding LD, LD-positive patients included patients with previously diagnosed chronic bronchitis and chronic obstructive pulmonary disease at baseline.

Notably, because the prefracture ALI was included in the telephone interview, when the description of a patient's ALI at the time of medical history collection was different from the description in the telephone interview, the results of the telephone interview were used.

To ensure the reliability of data collection, we checked the original data again when necessary to avoid human error or to find a reasonable explanation for unreliable data. For example, we encountered extreme values when data cleaning (such as extremely high or low BS levels) or details that should have been the same but were inconsistent (such as inconsistency between the days from hospitalization to surgery and the number of people undergoing surgery).

Sample Size

There are few studies on the sample size that is needed for validation studies, and there are fewer studies on the same size that is needed for validation studies on models of survival data. Some empirical studies have shown that for the external validation of prognostic models, a minimum effective sample size of 100 is needed, and the ideal effective sample size is 200 or more Steyerberg et al. (2019), Collins et al. (2016), Van Calster et al. (2016).

According to this criterion, in order for this study to achieve an unbiased and accurate estimation of the performance of the prognostic model, the study population should have at least 100 deaths. Given the relatively fixed number of hip fracture patients who are admitted each year, the needed sample size can be achieved only by expanding the time frame of the study. However, a timeline that extends too far into the past leads to difficulties in follow-up, in turn leading to decreased accuracy of outcome information. Therefore, we had to strike a balance between ensuring the accuracy of the follow-up results and expanding the study time frame. Under this premise and referring to the finding that revealed an approximately 10% mortality rate within

1 year in the developmental study, we collected a dataset with a sample size of approximately 1000 to obtain 100 deaths.

Missing Data

We estimated missing values via advanced multivariate model imputation techniques. Although studies have shown that the use of multiple imputation (MI) methods yields data that approximate the true values, these methods are complex and sometimes unnecessary Harel et al. (2007), Janssen et al. (2010). Some studies have shown that when the loss to follow-up rate is less than 10%, there is no significant difference between MI methods and other simple estimation methods Barzi et al. (2004). Empirical studies have shown that nonstatistician-friendly single imputation (SI) results in model prediction studies are not significantly different from MI results Missing Values et al. (2022). Therefore, we chose SI to address missing data (using the mouse package in R). When implementing SI, all 14 variables that were entered into the model were referenced.

Statistical Analysis Methods

In the model developmental study Fu et al. (2021), we clearly reported the mathematical equation of the SFHF model and the baseline hazard at 1 year after low-energy hip fracture. Therefore, we can easily obtain the prognostic index (PI) of the model and use this index to calculate the one-year survival rate of each individual in the validation study. The Cox regression model is typically expressed as $h\ (t,\ X) = h_0(t)\exp\ (\beta_1X_1 + \beta_2X_2 +\ldots + \beta mXm)$, where $h\ (t,\ X)$ is the hazard function of an individual with covariate X at time t, $h_0(t)$ is the baseline hazard function, β' is the regression coefficient of m predictor variables X_1 to X_m, and $\beta_1X_1 + \beta_2X_2 +\ldots + \beta mXm$ is the linear part of the model formula, which is positively related to the hazard function $h\ (t,\ X)$. In other words, the greater the risk is, the larger the value of $\beta_1X_1 + \beta_2X_2 + \ldots + \beta mXm$. Therefore, the linear part of the model reflects an individual's prognosis, which is statistically referred to as the prognostic index (PI), that is, $PI = \beta_1X_1 + \beta_2X_2 +\ldots + \beta mXm$. By transforming the above formula, we obtain $PI = \ln\ [h(t,\ X)/h_0(t)]$, which means that an individual's PI is the logarithm of the relative risk compared to a hypothetical individual with a PI of zero Royston et al. (2013), McLernon et al.(2023), Steyerberg et al.(2019). The PI is a dimensionless quantity that is derived from various indices and formulas, making it difficult to define a fixed range of values or a unit of measurement. In general, a higher PI indicates a poorer patient prognosis, whereas a lower PI suggests a better outcome. However, the interpretation of the PI may vary depending on specific clinical circumstances and the indices that are used for its calculation, thus requiring careful analysis and judgement. If expressed in terms of the survival rate, the model can be written as $S\ (t,\ X) = S_0(t)^{\exp\ (\beta 1X1\ +\ \beta 2X2\ +\ldots +\ \beta mXm)} = S_0(t)^{\exp\ (PI)}$. Specifically, for SFHF, according to the model formula we established in the development research, its PI = 0.042*Age + 0.305*SEX + MLA - 0.272*ALI + 0.343*CCD - 1.573*SUR + 0.008*SC - 0.082*ALB + 0.171*HYP. The one-year survival rate after a patient experiences a fragility hip fracture can be calculated as $S\ (1) = 0.984^{\ \exp\ (PI)}$, and the absolute risk of death within one year can be calculated as

$1\text{-}S\ (1) = 1\text{-}S_0(1)^{\ \exp\ (PI)} = 1\text{-}0.984^{\ \exp\ (PI)}$. Although our developmental study provided a nomogram that as developed on the basis of a simple model, we declined to use the nomogram to calculate the survival of patients in the validation cohort because it was inefficient and prone to error.

In the model developmental study, we described a full model and a reduced model; therefore, we evaluated both models in this study.

Traditional Measures

We evaluated the performance of the SFHF model by examining its discrimination and calibration on an external validation dataset D'Agostino et al. (2003).

Specifically, for this study, discrimination refers to the ability of the model to distinguish patients with shorter survival times after hip fracture (predicting patients at high risk) from those with longer survival times (predicting patients at low risk) McLernon et al. (2023), Pencina et al. (2012). We used Harrell's c-index and Uno's c-value, the latter of which was more suitable for the censored survival data model Ramspek et al. (2020), McLernon et al. (2023), Pencina et al. (2012), Uno et al. (2011). The value of c ranged from 0.5 to 1, with 0.5 indicating that the model has no discriminative ability and 1 indicating perfect discrimination. Generally, a c value greater than 0.7 indicates that the discrimination of the model is acceptable.

Calibration refers to the degree of agreement or consistency between the 1-year survival rate that is predicted by the model and the actual 1-year survival rate of a patient; that is, calibration refers to the degree of prediction accuracy. Crowson et al. (2016) Some studies have compared the advantages and disadvantages of mean calibration, weak calibration, moderate calibration and strong calibration, and these studies have concluded that model development and validation research should focus on moderate calibration because it ensures that clinical decisions based on the model will not cause harm Van Calster et al. (2016). Therefore, we generated a calibration plot of the model to assess its calibration Crowson et al. (2016). The specific method involved calculating the 1-year survival rate of each individual in the validation data via the SFHF model. The cohort was divided into ten equal groups according to the magnitude of the survival rate, after which the average predicted survival rate (as the x-axis) and the actual survival rate (as the y-axis) of each group were determined; the bootstrapping procedure of 1000 repetitions was used to obtain corrected values, which were added to a calibration plot for intuitive comparison. If the model produced perfect predictions for each group of patients, the respective values would fall on the standard 45-degree straight line in the graph.

We also grouped the validation cohort according to prognostic index size and plotted Kaplan–Meier curves that were superimposed with Kaplan–Meier curves of the developmental cohort. The discrimination and calibration performance of the model can also be intuitively determined from this figure. If the survival curves for the risk groups in the validation cohort are separated to the same extent as the developmental data are, then the discriminative power of the model is preserved in validation; if the degree of separation between the sets of curves decreases, then the model is less discriminative and vice versa. If the sets of curves from the two studies are intertwined and follow the same trend, the model has good calibration. Otherwise, the calibration is poor.

We also calculated the overall performance measure, namely, the Nagelkerke R2, to determine the model's ability to explain variation in the results of the validation dataset Steyerberg et al. (2018). This value ranges from 0 to 1. The larger the value is, the greater the ability of the model to explain the variation and the higher the prediction accuracy of the model Steyerberg et al. (2010).

Utility Measures

The use of metrics such as model discrimination and calibration to assess the performance of a model is more of a statistical perspective that does not provide information about whether the model has clinical value Vickers et al. (2006), Van Calster et al. (2018), Vergouwe et al. (2002). The original intention of developing and validating prognostic models was to provide guidance for clinical decision-making. In this study, the original aim of the SFHF model was to identify patients at high risk of dying within 1 year after fragility fracture and to treat such patients conservatively in an attempt to reduce the mortality rate among these patients. Whether the application of the model can achieve this clinical goal with the traditional model performance measures that were mentioned above cannot answer this question. Decision curve analysis (DCA), which has emerged in recent years, can link a model to clinical consequences and answer the most basic and most important question of whether the use of a model can promote clinical development Vickers et al. (2006), Van Calster et al. (2018), Vickers et al. (2010). Therefore, we performed DCA to determine whether the SFHF model has clinical value in the validation population.

DCA can reveal within which probability threshold range a model is valuable and determine the magnitude of the net benefit (NE) Van Calster et al. (2018), Vickers et al. (2008). The threshold probability means that under the probability of risk, for example, the expected benefit of choosing an intervention is equal to the expected benefit of rejecting the intervention Vickers et al. (2006), Localio et al. (2012). A model is considered clinically valuable if it achieves a greater net benefit than does the default strategy (treat all or treat none) within a reasonable threshold. Among them, whether this threshold range is reasonable depends on how much risk an individual is willing to take on a certain intervention.

Importantly, if the validation results showed poor calibration in the SFHF model, then we simply revised the model to make it better suited to the new environment Steyerberg et al. (2019), Houwelingen et al. (2000). Otherwise, we did not update the model.

Risk Groups

Another purpose of establishing the SFHF model was to group patients with hip fractures according to their prognostic index to facilitate the clinically stratified management of these patients. The grouping standard was still inconclusive, and patients were mostly divided into 3 or 4 groups on the basis of their clinical needs Royston et al. (2013), Altman et al. (2009). In the developmental study, we divided patients into low-, intermediate- and high-risk groups, with equal numbers of patients in each group. In the validation study, however, we regrouped the developmental data in an attempt to maximize between-group differences and minimize within-group differences Ramspek et al. (2020), Altman et al. (2009). Our grouping method was as follows: according to the size of the prognostic index, in terms of percentiles, the cut-off points were 0.15, 0.5, and 0.9. This resulted in a worst prognosis group that included 10% of the total sample, and this group exhibited a 1-year mortality rate (9.25%) that was similar to that observed in the developmental study. In this study, we also divided the validation cohort into 4 groups according to the abovementioned grouping method on the basis of the prognostic index that was provided by the SFHF model in the validation population.

Development versus Validation

To help readers more clearly understand the differences in patient conditions between the two studies, we provided special explanations for the differences in setting, patient inclusion criteria, predictors, and outcome definitions and measurements between the validation and developmental studies.

The dataset that was used in the developmental study came from Fenyang Hospital, which is a comprehensive teaching hospital for secondary care, in Lvliang city, Shanxi Province, China (located in western central Shanxi). The majority of the patients who were admitted (approximately 80%) were from rural areas. The dataset that was used in the validation study came from Shanxi Bethune Hospital in China, which is a level 3 regional referral hospital located in Taiyuan, the provincial capital, that is affiliated with the comprehensive teaching hospital of Shanxi Medical University. Approximately 60% of the patients who were treated at this hospital were from rural areas. Therefore, the two hospitals are not only geographically different but also substantially different in level.

Among the predictors, we redefined CCD and LD (see above for details), expanded the range of diseases that were considered positive results, and ensured that disease definitions more truly reflected individual patient characteristics. All the variables were coded in the same way as in the developmental study. Extreme SC values were also winsorized (creatinine values > 99th percentile were contracted to 99th percentile values). We also made slight adjustments to the patient inclusion criteria. The developmental study excluded patients with hip fractures coexisting with other fractures. In the validation study, eligible patients were allowed to have fragility fractures at other sites if they were caused by the same trauma that caused the hip fracture, regardless of whether the fracture had been treated surgically. These combined fractures, including distal radius, proximal humerus, and vertebral compression fractures, are the most prone to fragility fractures. The reason for this adjustment was that this condition is not uncommon in the clinic.

There was no difference between the two studies in terms of the outcome of interest, which was all-cause mortality within 1 year of fracture.

In general, compared with the developmental study, this validation study differed considerably in terms of its research setting, differed slightly in terms of its inclusion criteria and definition of predictors, and did not differ in its target results.

RESULTS

Participants

This study ultimately included 877 eligible patients with fragility hip fractures, 47 of whom were lost to follow-up. Among the 830 patients who were successfully followed up, 87 died within 1 year (shown in Figure 1). The follow-up time ranged from 7 days to 2295 days, with a median follow-up time of 829 days. Because we focused on survival or death within 1 year (365 days) after fracture, follow-up times beyond 1 year were truncated to 365 days. The characteristics of the eligible participants are shown in Table 1.

Figure 1: Flowchart of patient screening process

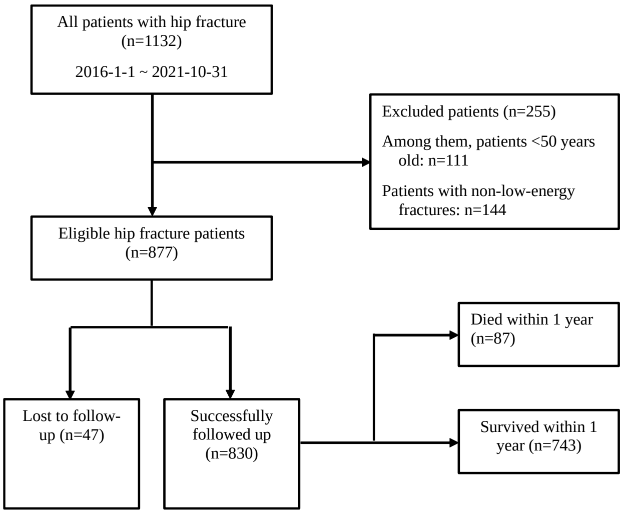

All patients with hip fracture
(n=1132)
2016-1-1 ~ 2021-10-31
Excluded patients (n=255)
Among them, patients <50 years old: n=111
Patients with non-low-energy fractures: n=144
Eligible hip fracture patients
(n=877)
Died within 1 year
(n=87)
Lost to follow-up (n=47)
Successfully followed up
(n=830)
Survived within 1 year (n=743)

Fig. 1. Flowchart of patient screening process

To help readers clearly understand the difference between the validation cohort and the developmental cohort, we present the characteristics of the two cohorts in a comparison table (shown in Table 2). Because the level of the validation research hospital was higher than that of the developmental research hospital, the patients who were admitted to the former had more complicated and serious conditions. This can be seen from the fact that the proportion of patients with secondary hip fractures and patients with malignancies was substantially greater in the validation study than in the developmental study.

Model Performance

To make it easier for readers to understand the differences in the predictions that were made by the SFHF model in the two cohorts, we compare the distributions of the PIs from the two studies. The PI is the weighted sum of the variables in the model, where the weights are the regression coefficients Royston et al. (2013), Ramspek et al. (2020), Moons et al. (2015). Comparisons of the PI distributions of the two studies are shown in Table 3 and Figure 2. There are no obvious outliers in the validation data, and the PI distribution in the validation study is wider than that in the developmental study.

Discrimination

Harrell's c-index, Uno's c-value and the Nagelkerke R2 of the SFHF model are shown in Table 4.

Calibration

The calibration plots of the full and simplified SFHF models when used on the validation data are shown in Figure 3 and Figure 4.

Table 1: Participant Characteristics

Characteristics	Missing Values, n (%)	Value
Sociodemographic characteristics		
Mean age (years)	0	76.3 (SD, 9.6) (range 50-97)
Male	0	298 (34.0%)
Medical insurance	0	
Employee medical insurance (EMI)		301 (34.3%)
Non-EMI		576 (65.7%)
Fracture-related		
Fracture site	0	
Femoral neck		447 (51.0%)
Intertrochanteric		414 (47.2%)
Intertrochanteric		16 (1.8%)
Fracture type	0	
Primary		696 (79.4%)
Secondary		181 (20.6%)
Fracture to admission (d)	0	4.0 (SD, 13.2) (range 0-216)
Admission to surgery (d)	0	4.6 (SD, 3.5) (range 0-38)
(n=769)		
Length of stay (LOS)	0	12.1 (SD, 9.1) (range 1-130)
Medical history		
Diabetes	2 (0.2%)	190 (21.7%)
Hypertension (HYP)	2 (0.2%)	440 (50.3%)
Malignancy (MLA)	0	62 (7.1%)
Kidney disease (KD)	0	18 (2.1%)
Lung disease (LD)	28 (3.2%)	191 (22.5%)
Ability to live independently (ALI) (no = 0, yes = 1)	0	110 (12.5%)
Cardiovascular and cerebrovascular disease (CCD)	1 (0.1%)	477 (54.5%)
Clinical indicators		
Blood sugar (BS) (mmol/L)	166 (18.9%)	6.7 (SD, 2.4) (range 3.2-25.1)
Serum creatinine (SC) (μmol/L) *	53 (6.0%)	79.3 (SD, 41.7) (range 26.5-378.3)

Haemoglobin (Hb) (g/L)	20 (2.3%)	116.6 (SD, 20.2) (range 47.0-218.0)
Albumin (ALB) (g/L)	43 (4.9%)	35.3 (SD, 4.4) (range 18.9-52.3)
Mean arterial pressure (MAP) (mmHg)	0	100.0 (SD, 14.3) (range 59.3-147.3)
Partial pressure of oxygen (PaO2) (mmHg)	289 (33.0%)	72.5 (SD, 20.3) (range 25.8-184.4)
Total protein (TP) (g/L)	103 (11.7%)	63.1 (SD, 6.6) (range 41.1-93.5)
Treatment		
Osteodistraction	1 (0.1%)	96 (11.0%)
Surgery	0	769 (87.7%)

Note. SD: standard deviation

* Value before SC was winsorized: 82.1 (SD, 64.9) (range 26.5-854.6).

Table 2: Comparison of Participant Characteristics between the Developmental and Validation Cohorts

Characteristic	Developmental Cohort(n=735)	External Validation Cohort (n=877)
Setting		
Mean age (years)	74.8 (SD, 9.5) (range 50–103)	76.3 (SD, 9.6) (range 50-97)
Male	279 (38.0%)	298 (34.0%)
EMI	101 (13.7%)	301 (34.3%)
Femoral neck	305 (41.5%)	447 (51.0%)
Intertrochanteric	413 (56.2%)	414 (47.2%)
Secondary fracture	46 (6.3%)	181 (20.6%)
Admission to surgery (d)	5.5 (SD, 3.3) (range 1–45)	4.6 (SD, 3.5) (range 0-38)
LOS	13.0 (SD, 6.4) (range 1–52)	12.1 (SD, 9.1) (range 1-130)
Predictors		
HYP	357 (48.6%)	440 (50.3%)
MLA	18 (2.4%)	62 (7.1%)
ALI	107 (14.6%)	110 (12.5%)
CCD	295 (40.1%)	477 (54.5%)
SC (μmol/L)	70.3 (SD, 23.8) (range 26.0–190.8)	79.3 (SD, 41.7) (range 26.5-378.3)
ALB(g/L)	37.9 (SD, 4.1) (range 21.8–48.2)	35.3 (SD, 4.4) (range 18.9-52.3)
Surgery	637 (86.7%)	769 (87.7%)
Outcome		
Death within 1 year	68	87
Lost to follow-up	11 (1.5%)	47 (5.4%)

Table 3: Comparison of PI distributions in the developmental and validation studies

Measurement	Developmental study		Validation study	
	FULL	LASSO	FULL	LASSO
Mean	-1.25	-1.241	-0.736	-0.74
Median	-1.293	-1.286	-0.835	-0.859
Range of values	-4.071, 2.563	-3.963, 2.566	-3.785, 3.890	-3.693, 3.841
Range	6.634	6.529	7.675	7.534
SD	1.211	1.194	1.278	1.253

Note. SD: standard deviation

Table 4: Discriminative performance measurements of the SFHF model

Measurement	Developmental			Internal Validation (B=1000)		External Validation		
	Harrell's c-index (SE)	Uno's c-value (SE)	R2	Harrell's c-index	R2	Harrell's c-index	Uno's c-value	R2
Full model	0.816 (0.025)	0.817 (0.025)	0.188	0.789	0.147	0.764 (0.024)	0.765 (0.024)	0.144
LASSO model	0.814 (0.026)	0.815 (0.025)	0.187	0.795	0.158	0.763 (0.024)	0.763 (0.024)	0.144

Figure 2: PI histograms for the developmental and validation datasets. The three vertical lines on the graph are the percentile cut-off points of risk grouping, 0.15, 0.5, and 0.9.

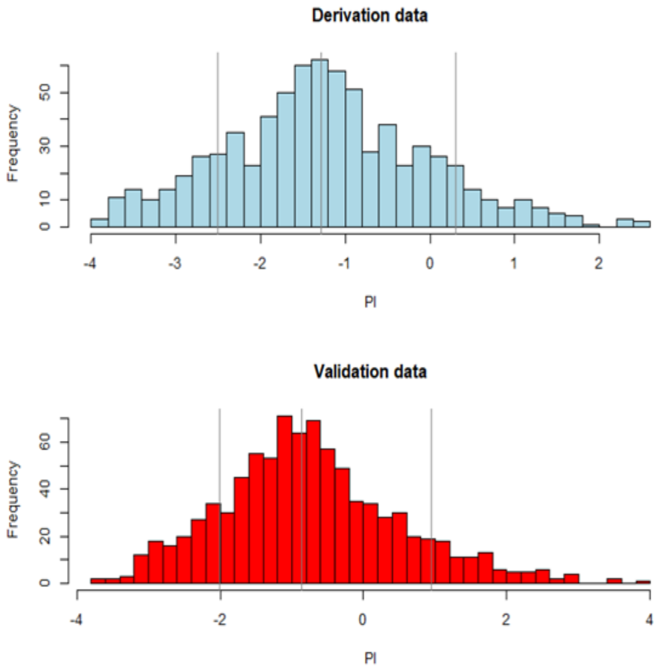


Figure 3: The calibration plot of the FULL model in the validation dataset

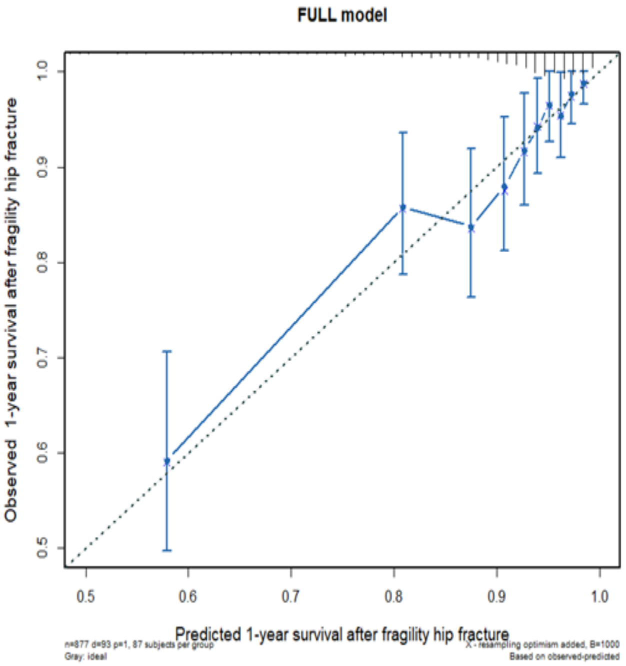
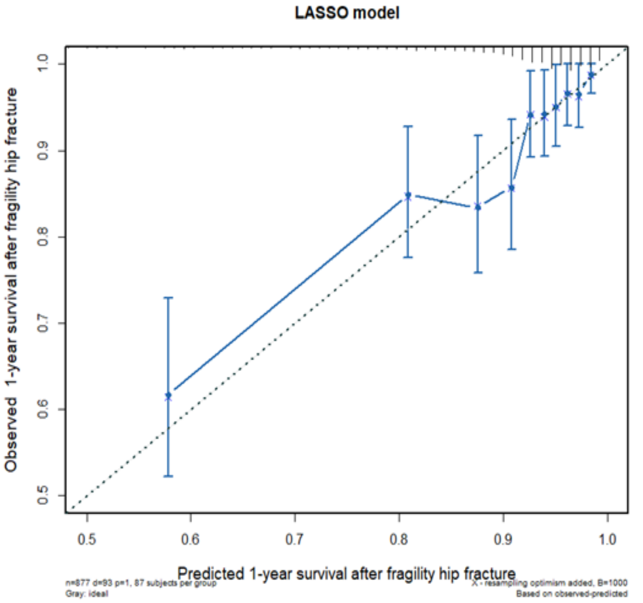


Figure 4: The calibration plot of the LASSO model in the validation dataset



As shown in Figure 3, most of the validation study populations were concentrated in the high survival direction. The difference in survival between the group with the lowest predicted survival rate (mean predicted survival rate = 0.579) and the group with the second lowest survival rate (0.808) was the most significant. The difference between groups decreased as the predicted survival rates increased. The differences among the 5 groups with the highest survival rates were essentially the same. The confidence intervals for the predicted risks of all the groups covered the ideal 45° line.

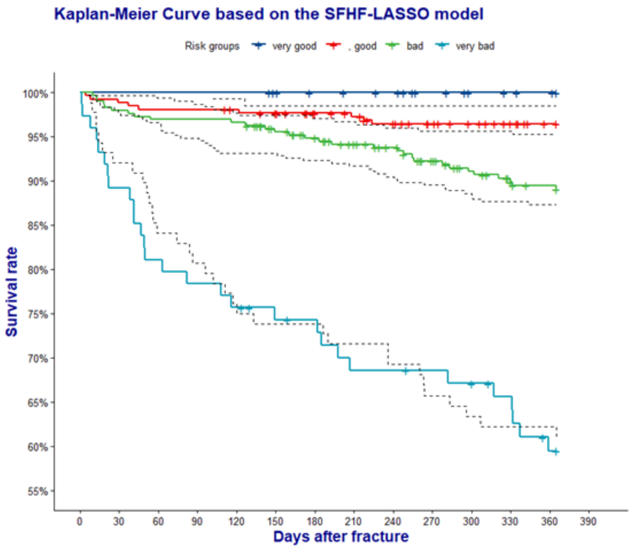
Figure 4 shows that the calibration map features of the least absolute shrinkage and selection operator (LASSO) model are basically the same as those of the full model.

Risk Groups

After obtaining the prognosis of each patient, we divided the population into four groups according to the aforementioned grouping method and plotted the Kaplan–Meier curves of each group. We overlapped the survival curves from the two studies for comparison (shown in Figure 5). The specific survival rates of each risk group in the two studies are shown in Table 5.

Figure 5 shows that, in general, the survival curves of each group in the validation study are in good agreement with those in the developmental study. Specifically, in the very good and bad groups, it seems that the validation curves were systematically lower than the developmental curves, but in the good and very bad groups, the curves of the two studies were almost identical. The degree of separation between the respective curves of the two studies was similar, and the degree of separation between the curves of the good group and the bad group in the validation study was even better than that in the developmental study.

Figure 5: Survival curves for each risk group in the developmental and validation studies



Clinical Usefulness

The results of DCA of the SFHF model in the validation cohort is shown in Figure 6.

Figure 6: Decision curve analysis of SFHF in the validation data

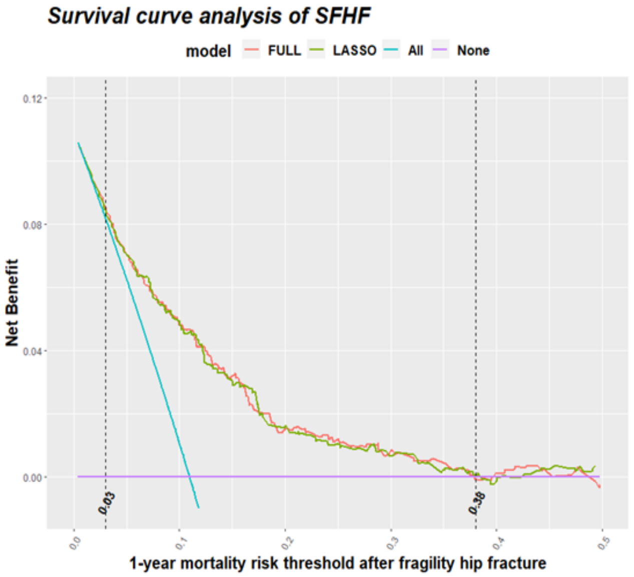


Figure 6 shows that when the risk threshold is in the range of 0.03~0.38, the net benefit of using the SFHF model for clinical decision-making is better than that of the default treat all or treat none strategies. When the value is between 0.38 and 0.50, the net benefit becomes unstable. Compared with the treat none strategy, the corresponding range of net benefit of using the SFHF model is 0.105 (risk threshold = 0; at this time, using the SFHF model is equivalent to adopting the treat-all

Table 5: Survival rates for the respective risk groups in the developmental and validation studies

Risk group	Developmental data				Validation data			
	N (risk)	N (mean)	Survival	SE	N (risk)	N (mean)	Survival	SE
Very good	95	0	1	0	111	2	0.985	0.011
Good	211	9	0.964	0.012	242	14	0.953	0.012
Bad	219	30	0.891	0.019	261	43	0.873	0.018
Very bad	38	29	0.594	0.059	51	34	0.61	0.052

strategy) ~ 0 (0.38). Compared with the treat all strategy, the net benefit of using the SFHF model ranges from 0 (0.03) to 0.046 (0.11; at this time, the net benefit of the treat all strategy is 0). The decision curves of the full model and the simple model are basically the same.

Model Updating

Given the good performance of the model in calibration plots and survival curve plots, we did not consider updating the SFHF model.

DISCUSSIONS

Limitations

Any study has certain limitations, and this research is no exception. First, owing to the inherent deficiencies in the retrospective study design, there was a lack of initiative in the collection of patients' individual characteristics. Some predictors were not available, which resulted in lost opportunities to evaluate some similar models. Moreover, the measurement of data cannot be standardized in advance, as in prospective studies; this affects the accuracy of the measurement of different variables to a certain extent, thereby affecting the reliability of the verification results. Second, the effective sample size of this study was 93 deaths. This number was close to the value of at least 100 valid events, which is currently considered the minimum requirement for external validation studies; however, it is still far from the ideal requirement (the ideal valid sample size is 200 or more). This affects the robustness of the findings to some degree. Third, compared with other outcomes, such as postoperative function and pain, all-cause mortality is relatively objective, so retrospective follow-up of death is reliable. However, for patients who have been deceased for 3 years or longer, it is difficult to determine the specific time of death via retrospective telephone follow-up, and the accuracy of this approach is still not satisfactory compared with the follow-up approaches of prospective studies. This also affects the reliability of the study. Fourth, this study has geographic/broad validation, and it is a moderate-intensity study Ramspek et al. (2020), Moons et al. (2012). However, a validation study that is performed by a developer inherently has a certain degree of bias compared with a study that is organized by others. Fifth, although the population of the validation study came from other institutions, the two institutions are geographically adjacent, and there is a small overlap in the patients who are admitted. This affects the evaluation of the transportability of the model. The generalization of the results of this study should be performed with great caution if the model is to be considered for use in the wider region of China. Sixth, the validation study revealed that the Nagelkerke R2 was 0.144, indicating that there is still much room for improvement in the ability to explain outcome variation. Such a result can be expected; after all, we used only preoperative characteristics that were easily and quickly available in a short timeframe after admission as predictors and excluded intraoperative and postoperative characteristics (such as intraoperative or postoperative complications) that account for more variation in mortality outcomes Li et al. (2021). Seventh, some patients' families and doctors may be more concerned about the risk of death in a shorter period after fracture, such as six months or three months after fracture, which can provide more sufficient reasons for the decision to avoid surgery. Risk predictions for these time points were less robust because of the smaller effective sample size in shorter periods. Therefore, the use of this model to predict the risk of death in a shorter period is not recommended.

Interpretation

As mentioned above, from the perspective of different settings and individual predictor definitions between the two studies, this study has geographic or broad validation, so the research results are moderately strong. Therefore, we focused on the transportability of the model rather than its reproducibility. According to the research results described above, it can be concluded that the performance of the simple LASSO model is essentially indistinguishable from that of the full model. Therefore, here, we explain only the results of the LASSO model. According to the comparison of the PI of the two studies, the SFHF model provided a similar amount of prognostic information on the two sets of data Altman et al. (2000). According to the grouping method of this study, the prognostic information value of the model in the developmental study was mortality in the highest risk group and mortality in the lowest risk group, that is, 0.392-0=0.392. In the validation study, this value was 0.386-0.015=0.371. The amount of prognostic information available in developmental studies was largely maintained in the validation data. From this perspective, the accuracy and generalizability of the model are confirmed. In terms of discrimination, although Harrell's c-index changed from the original value of greater than 0.8 (good) to 0.763 (acceptable), according to the numbers, the discrimination decreased. However, given the marked differences in the patient conditions between the two studies, we believe that the SFHF model has good discriminative power in validation. To further confirm the good discriminative power of the SFHF model, we also provided Uno's c statistic, which is more applicable to survival data with censored outcomes. A comparison of Uno's c-value with Harrell's c-index revealed that the discriminative power of the SFHF model was very stable in validation (see Table 4). In addition, good discrimination is also shown in Figure 5, where the survival curves of the high-risk population in the validation study were clearly separated from those of the other groups, confirming the good ability of the model to identify the high-risk population. Some scholars believe that the best indicator for assessing the usefulness of a model is its ability to separate different risk groups. If we follow this standard, we can say that the SFHF model has been successfully validated Altman et al. (2009).

From a calibration point of view, the predicted probability of the SFHF model is relatively accurate. The patients were divided into 10 groups according to the predicted risk probability obtained from the model, and compared with the actual observations of each group, there was no significant difference between the predicted and actual results. This is reflected by the fact that the predicted risk confidence intervals for each group covered the ideal standard line that indicates perfect prediction.

Similarly, the accuracy of the prediction can be visually assessed, as shown in Figure 5. In Figure 5, although the very good and bad groups seem to have slightly systematically high predictions, in the high-risk group, on which we are most focused, the survival curves of the two groups of data are intertwined with a high degree of agreement. Overall, the survival curves of the respective risk groups in the two studies were in good agreement, confirming good calibration of the SFHF model.

Research shows that accurate models are not necessarily Useful Steyerberg et al. (2010), Vickers et al. (2006), Vickers et al. (2008). Therefore, it is insufficient to conclude that a model has been successfully validated on the basis solely of statistical accuracy. In terms of clinical usefulness, the SFHF model has great practical value in externally validated populations. In the DCA, when the risk threshold was between 0.03 and 0.38, the net benefit of the model was better than that of the treat all or treat none strategies. When we predict a patient's 1-year risk of death to be low, i.e., less than 0.05, it is clear that doctors and patients will not hesitate to choose surgery when choosing a treatment approach. At this risk level, surgery clearly benefits most patients. Conversely, when the risk of death is obvious, i.e., greater than 0.5, doctors and patients will not hesitate to choose conservative treatment. When the risk lies in the middle, doctors and patients are confused. Our DCA results show that, within these risk thresholds, clinical decision-making based on our model is highly valuable. For example, if we set the 1-year risk of death threshold to 10%, if a patient's predicted risk of death exceeds 10%, conservative treatment is considered; otherwise, surgery is considered compared with the option of surgical treatment for all patients (treat none), NE =0.049. This means that the net result of making decisions on the basis of the SFHF model is that out of 100 patients, we would conservatively treat 4.9 patients who would die within 1 year, rather than treating all patients with surgery and having surgery-treated patients not die within 1 year. In contrast to the conservative treatment strategy for all patients (treat all), the model's NE=0.049-0.010=0.039, which was calculated via the net benefit formula 0.039×100/(0.1/0.9)=35.1. This means that, according to the predictive model, the use of conservative treatment decreased by 35.1% among patients who would not die within 1 year, whereas the number of patients who would die within 1 year of surgical treatment did not increase Vickers et al. (2006). Compared with the use of surgery for all patients as the default strategy, the difference in NE that was obtained by model-based decision-making was ΔNE=0.049-0=0.049, and the test trade-off was 1/0.049=20.4. If we are willing to apply the model to 20 patients to identify patients who will die within 1 year of fracture, then the model is valuable Van Calster et al. (2018). Unfortunately, we cannot provide a reasonable risk threshold here, as it varies from individual to individual and involves the need to reasonably evaluate all possible outcomes Van Calster et al. (2018).

Overall, the reproducibility and generalizability of the SFHF model were confirmed in this external study, both by traditional metrics and by clinical usefulness. At this point, we can conclude that the SFHF model has been initially successfully validated. Unlike the Li et al. (2021) and Endo et al. (2018) models, which use intraoperative or postoperative indicators as predictors, our model includes only preoperative individual characteristics that can be quickly and easily determined after admission. Additionally, unlike the model described by Ma et al. (2018), which requires multidisciplinary collaboration to obtain accurate predictors, the SFHF model does not require the assistance of multidisciplinary consultation, and the predictors can be independently determined by an orthopaedic surgeon 48 hours after admission. The outstanding features of convenience, speed, and accuracy provides the most fundamental guarantee that this model will enter clinical practice in the future. There is no complex scoring system for the model predictors, avoiding the use of the model as a predictor, as in Hjelholt et al. (2022), Trevisan et al. (2021), Li et al. (2021), and Söderqvist et al. (2009). Moreover, no interaction variable is introduced, making the model easy to understand and explain.

Implications

The initial verification of the SFHF model was successful, confirming that the strategy we chose for model development on the basis of a small sample is feasible. That is, according to the principles and purposes of modelling, combined with professional knowledge and previous research results, the number of input variables is controlled instead of using a data-driven variable selection strategy to avoid overfitting. Owing to the limited strength of this validation study, we urgently need to increase the validation strength. For example, nondevelopers should validate the model with a more heterogeneous, sufficient and relevant population from other regions of China (e.g., eastern and southern China) to further evaluate the generality and accuracy of the SFHF model. On the other hand, given the effectiveness of the preliminary validation, we may consider conducting model impact studies in our hospital to investigate whether the clinical application of the SFHF model can reduce 1-year mortality after hip fracture among elderly individuals compared with not using the model Moons et al. (2009), Reilly et al. (2006). In addition, in the study of prognostic factors, this model can also be used as guidance for multivariate adjustment analysis.

CONCLUSION

The results of the study indicate that, from a statistical point of view, the SFHF model has good discrimination and calibration; that is, the model has certain accuracy and generalizability. From a clinical point of view, the model has clinical value. Therefore, we can consider applying this model to the management of elderly patients with hip fractures and, accordingly, the formulation of treatment plans, with the expectation of reducing 1-year mortality in this population.

ABBREVIATIONS

SFHF: Survival after fragility hip fracture; DCA: Decision curve analysis; CI: Confidence interval; LOS: Length of stay; ALI: Ability to live independently; LD: Lung disease; CCD: Cardiovascular and cerebrovascular disease; KD: Kidney disease; MAL: Malignancy; HYP: Hypertension; MAP: Mean arterial pressure; PaO2: Partial pressure of oxygen; Hb: Haemoglobin; SC: Serum creatinine; BS: Blood sugar; ALB: Albumin; TP: Total protein; SUR: Surgery; MI: Multiple imputation; SI: Single imputation; PI: Prognostic index; LASSO: Least absolute shrinkage and selection operator; NE: Net benefit.

DECLARATIONS

Ethics Approval and Consent to Participate

This study was approved by the Medical Ethics Committee of Shanxi Bethune Hospital. All the steps were performed in accordance with the relevant guidelines and regulations. The requirement for informed consent was waived by the Medical Ethics Committee of Shanxi Bethune Hospital.

Consent for Publication

Not applicable.

Availability of Data and Materials

The datasets that were used in this study are available from the corresponding authors upon reasonable request, if required by the reader.

Competing Interests

The authors have no relevant financial or nonfinancial interests to disclose.

Funding

This study was funded by the Lvliang Municipal Science and Technology Bureau, a department of the Lvliang Municipal Government in Shanxi Province, China. The project is titled "A Study on the Correlation Between Time to Injury and Mortality in a Population of Patients with Fragility Hip Fractures," with project number 2023SHFZ14, and was approved on November 1, 2023. In support of promoting national health and medical advancement, this government agency solely provides financial support without imposing any specific restrictions on the research content.

Authors' Contributions

In this study, Hairui Fu and Feixiong Li were Co-First authors, Hairui Fu participated and was responsible for the entire research process. Feixiong Li was responsible for data collection, follow-up and verification and participated in the writing of the paper. Xuan Zhang participated in the verification and analysis of the data; Bin Liang provided expertise and study design guidance; and Dou Wu provided support for the study in terms of ethics approval, hospital department coordination, statistics, and clinical knowledge.

Acknowledgements

The authors thank teachers from the hospital's medical record room, information centre, and orthopaedic reference room for their help during the data collection process. Additionally, I would like to thank my wife, Xingying Li, for help during the phone interviews. Finally, I would like to thank my colleagues Enzhe Zhao and Liming He for their help during the information verification process.

REFERENCES

1.Fu H, Liang B, Qin W, et al. 2021. Development of a prognostic model for 1-year survival after fragile hip fracture in Chinese. J Orthop Surg Res. 16(1):695.

2.Altman DG, Royston P. 2000. What do we mean by validating a prognostic model? Stat Med. 19(4):453-73.

3.Moons KGM, Altman DG, Vergouwe Y, et al. 2009. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 338: b606.

4.Royston P, Altman DG. 2013. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 13:33.

5.Ramspek CL, Jager KJ, Dekker FW, et al. 2020. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 14(1):49-58.

6.Justice AC, Covinsky KE, Berlin JA. 1999. Assessing the generalizability of prognostic information. Ann Intern Med. 130(6):515-24.

7.Trevisan C, Gallinari G, Carbone A, et al. 2021. Efficiently stratifying mid-term death risk in femoral fractures in the elderly: introducing the ASAgeCoGeCC Score. Osteoporos Int. 32(10):2023-31.

8.Menéndez-Colino R, Gutiérrez Misis A, Alarcon T, et al. 2021. Development of a new comprehensive preoperative risk score for predicting 1-year mortality in patients with hip fracture: the HULP-HF score. Comparison with 3 other risk prediction models. Hip Int. 31(6):804-11.

9.Cenzer IS, Tang V, Boscardin WJ, et al. 2016. One-Year Mortality After Hip Fracture: Development and Validation of a Prognostic Index. J Am Geriatr Soc. 64(9):1863-8.

10.Jiang HX, Majumdar SR, Dick DA, et al. 2005. 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. 20(3):494-00.

11.van de Ree CL, Gosens T, van der Veen AH, et al. 2020. Development and validation of the Brabant Hip Fracture Score for 30-day and 1-year mortality. Hip Int. 30(3):354-62.

12.Cui L, Li J, Xie D, et al. 2022. Role of the Social Support and Health Status in Living Arrangement Preference of the Elderly in China-A Cross-Sectional Study. Front Public Health. 10:860974.

13.Bao J, Zhou L, Liu G, et al. 2022. Current state of care for the elderly in China in the context of an aging population. Biosci Trends. 16(2):107-18.

14.Nie J, Fan R, Wu Y, et al. 2022. By Internal Network or by External Network? -Study on the Social Network Mechanism of Reducing the Perception of Old-Age Support Risks of Rural Elders in China. Int J Environ Res Public Health. 19(22):15289.

15.Zhang Y, Goza FW. 2006. Who will care for the elderly in China? Journal of Aging Studies. 20(2):151–64.

16.Center For Healthy Aging and Development Studies. 2020. The Chinese Longitudinal Healthy Longevity Survey (CLHLS)-Longitudinal Data（1998-2018）. Peking University Open Research Data

17.Moons KGM, Altman DG, Reitsma JB, et al. 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 162(1): W1-73.

18.Steyerberg EW. 2019. Patterns of External Validity. In: Steyerberg EW, editor. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Statistics for Biology and Health. 367–97.

19.Collins GS, Ogundimu EO, Altman DG. 2016. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 35(2):214-26.

20.Van Calster B, Nieboer D, Vergouwe Y, et al. 2016. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology. J Clin Epidemiol. 74:167-76.

21.Harel O, Zhou XH. 2007. Multiple imputation: review of theory, implementation and software. Statistics in Medicine. Stat Med. 26(16):3057-77.

22.Janssen KJM, Donders ART, Harrell FE Jr, et al. 2010. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol. 63(7):721-7.

23.Barzi F, Woodward M. 2004. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. Am J Epidemiol. 160(1):34-45.

24.Missing Values. 2022.

25.McLernon DJ, Giardiello D, Calster BV, et al. 2023. Assessing Performance and Clinical Usefulness in Prediction Models with Survival Outcomes: Practical

Guidance for Cox Proportional Hazards Models. Ann Intern Med. 176(1):105-14.

26.Steyerberg EW. 2019. Statistical Models for Prediction. In: Clinical Prediction Models. Statistics for Biology and Health. 59–93.

27.D'Agostino RB, Nam BH. 2003. Evaluation of the Performance of Survival Analysis Models: Discrimination and Calibration Measures. In: Handbook of Statistics. 23:1–25.

28.Pencina MJ, D'Agostino RB Sr, Song L. 2012. Quantifying discrimination of Framingham risk functions with different survival C statistics. Stat Med. 31(15):1543-53.

29.Uno H, Cai T, Pencina MJ, et al. 2011. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 30(10):1105-17.

30.Crowson CS, Atkinson EJ, Therneau TM. 2016. Assessing calibration of prognostic risk scores. Stat Methods Med Res. 25(4):1692-706.

31.Steyerberg EW. 2019. Evaluation of performance. In: Clinical prediction models: A practical approach to development, validation, and updating. 277–308.

32.Steyerberg EW, Vickers AJ, Cook NR, et al. 2010. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 21(1):128-38.

33.Vickers AJ, Elkin EB. 2006. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 26(6):565-74.

34.Van Calster B, Wynants L, Verbeek JFM, et al. 2018. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. Eur Urol. 74(6):796-04.

35.Vergouwe Y, Steyerberg EW, Eijkemans MJC, et al. 2002. Validity of prognostic models: when is a model clinically useful? Semin Urol Oncol. 20(2):96-107.

36.Vickers AJ, Cronin AM. 2010.Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. Semin Oncol. 37(1):31-8.

37.Vickers AJ. 2008. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. Am Stat. 62(4):314-20.

38.Localio AR, Goodman S. 2012. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. Ann Intern Med. 157(4):294-5.

39.Steyerberg EW. 2019. Updating for a New Setting. In: Steyerberg EW, editor. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Statistics for Biology and Health.399–429.

40.Houwelingen V. 2000. Validation, calibration, revision and combination of prognostic survival models.

41.Altman DG. 2009. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest. 27(3):235-43.

42.Moons KGM, Kengne AP, Grobbee DE, et al. 2012. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 98(9):691-8.

43.Li Y, Chen M, Lv H, et al. 2021. A novel machine-learning algorithm for predicting mortality risk after hip fracture surgery. Injury. 52(6):1487-93.

44.Vickers AJ. 2008. Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. Am Stat. 62(4):314-20.

45.Endo A, Baer HJ, Nagao M, et al. 2018. Prediction Model of In-Hospital Mortality After Hip Fracture Surgery. J Orthop Trauma. 32(1):34-38.

46.Ma M, Lu H, Zhang P, et al. 2018. Formulating a preoperative risk scoring system for elderly patients with hip fracture. Chinese Journal of Orthopaedic Trauma.1031–7.

47.Hjelholt TJ, Johnsen SP, Brynningsen PK, et al. 2022. Development and validation of a model for predicting mortality in patients with hip fracture. Age Ageing. 51(1): afab233.

48.Söderqvist A, Ekström W, Ponzer S, et al. 2009. Prediction of mortality in elderly patients with hip fractures: a two-year prospective study of 1,944 patients. Gerontology. 55(5):496-04.

49.Reilly BM, Evans AT. 2006. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 144(3):201-9.