External validation of the U-HIP prediction model for in-hospital mortality in geriatric hip fracture patients

Schuijt, H.J.; Smeeing, D.P.J.; Groenwold, R.H.H.; Velde, D. van der; Weaver, M.J.

Injury 53 (2022) 1144–1148

Contents lists available at ScienceDirect

Injury

journal homepage: www.elsevier.com/locate/injury

External validation of the U-HIP prediction model for in-hospital mortality in geriatric hip fracture patients

H.J. Schuijt[a,b,*], D.P.J. Smeeing[b], R.H.H. Groenwold[c,d], D. van der Velde[b], M.J. Weaver[a]

[a] Harvard Medical School Orthopedic Trauma Initiative, Brigham and Women's hospital, Boston, United States
[b] St. Antonius Hospital, Dpt. of Trauma Surgery, Utrecht, the Netherlands
[c] Leiden University Medical Center, Dpt. of Clinical Epidemiology, Leiden, the Netherlands
[d] Leiden University Medical Center, Dpt. of Biomedical Data Sciences, Leiden, the Netherlands

ARTICLE INFO

Article history:
Accepted 17 December 2021

Keywords:
Geriatric trauma
Hip fracture
Surgery
Prediction model
Decision-making

ABSTRACT

Introduction: Identification of high-risk hip fracture patients in an early stage is vital for guiding surgical management and shared decision making. To objective of this study was to perform an external international validation study of the U-HIP prediction model for in-hospital mortality in geriatric patients with a hip fracture undergoing surgery.

Materials and methods: In this retrospective cohort study, data were used from The American College of Surgeons National Surgical Quality Improvement Program. Patients aged 70 years or above undergoing hip fracture surgery were included. The discrimination (c-statistic) and calibration of the model were investigated.

Results: A total of 25,502 patients were included, of whom 618 (2.4%) died. The mean predicted probability of in-hospital mortality was 3.9% (range 0%-55%). The c-statistic of the model was 0.74 (95% CI 0.72–0.76), which was comparable to the c-statistic of 0.78 (95% CI 0.71–0.85) that was found in the development cohort. The calibration plot indicated that the model was slightly overfitted, with a calibration-in-the-large of 0.015 and a calibration slope of 0.780. Within the subgroup of patients aged between 70 and 85, however, the c-statistic was 0.78 (95% CI 0.75–0.81), with good calibration (calibration slope 0.934).

Discussion and conclusion: The U-HIP model for in-hospital mortality in geriatric hip fractures was externally validated in a large international cohort, and showed a good discrimination and fair calibration. This model is freely available online and can be used to predict the risk of mortality, identify high-risk patients and aid clinical decision making.

Introduction

The increase in the number of geriatric hip fracture patients is a global health concern. They constitute a fast-growing group of patients who are notorious for adverse outcomes [1]. Identification of high-risk patients in an early stage is vital for guiding surgical management and shared decision making. Prediction models can be used to predict the risk of clinical outcomes and help to identify high-risk patients [2].

A few prediction models have been developed to predict post-operative mortality among hip fracture patients, including two studies that investigated in-hospital mortality as an outcome [3–5]. However, in many prediction models, predictor values are dichotomized (even though this is strongly discouraged by experts in this field), and have not been sufficiently validated [2–4,6,7]. Additionally, these models showed a lack of fit and poor discrimination in previous studies [5]. Finally, both of these models have incorporated variables that are generally not known at the emergency department at the moment that the prediction is to be made (e.g. time to surgery), which is the most critical flaw of both these prediction models and severely limits their clinical usefulness [3–5]. Hence, there are no externally validated models predicting in-hospital mortality in this patient population that show a good predictive performance.

In a previous study, a prediction model (the U-HIP (Utrecht Hip) algorithm) was developed in 1014 hip fracture patients aged 85 years or older (median 90, IQR 87–93) in the Netherlands, with an in-hospital mortality of 4% ($n = 38$). After correction for optimism, this model showed good discrimination (c-statistic 0.77) at internal validation. Predictors in the model were age, sex, American Society of Anesthesiologists physical status classification system (ASA)

* Corresponding author.
E-mail address: h.j.schuijt@isala.nl (H.J. Schuijt).

https://doi.org/10.1016/j.injury.2021.12.028
0020-1383/

and hemoglobin serum levels (mmol/L) upon presentation at the ED [8]. The purpose of the current study was to perform a validation study to externally validate the U-HIP prediction model for in-hospital mortality in a North-American population of hip fracture patients aged 70 or above undergoing surgery [8,9]. The authors hypothesize that the model will show good discrimination (i.e. c-statistic $\geq$0.70) and calibration (i.e. calibration-in-the-large $<$0.02 and $>-$0.02, and calibration slope $>$0.75).

Materials and methods

This study was approved by the institutional review board and medical ethical committee and reported in accordance with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines of the enhancing the quality and transparency of health research network (EQUATOR) [10].

For this cohort study, data were collected from January 1st 2016 until December 31st 2018 by The American College of Surgeons National Surgical Quality Improvement Program (NSQIP) [11]. The NSQIP collects data on hip fracture surgery from over 150 hospitals across the world, although most of these hospitals are located in North-America. The inclusion criteria for the validation cohort were as follows: 1) patients aged 70 years or above, presenting to the emergency department with a hip fracture (OTA classification: 31-A or 31-B) [12] and; 2) undergoing hip fracture surgery for a nonpathological hip fracture. Patients with American Society of Anesthesiologists physical status classification system (ASA) V were excluded, because the development cohort did not include these patients and because patients with ASA status V are, by definition, moribund and thus accurate risk prediction are irrelevant [8,13]. The primary outcome for this study was in-hospital mortality. In this study, the authors decided to validate the model in patients aged 70 years or above, even though the development cohort consisted of patients 85 year or above [8]. The cut-off of 70 years was chosen because the vast majority of patients who experience in-hospital mortality after a hip fracture are aged 70 years or above. Because of expected heterogeneity in patients younger than 70 and the relatively low mortality risk in that group, a cut-off of 70 years was chosen. Hence, we focus on the patient population in whom the risk prediction is likely to be most relevant, while increasing the age range (compared to the development population) to improve clinical applicability and usefulness.

The following data were collected at baseline (i.e., at hospital admission): fracture type (i.e. femoral neck nondisplaced, femoral neck displaced, intertrochanteric, or subtrochanteric), age, sex, ASA classification [13] (I to IV), a previous diagnosis of dementia in medical history, and hemoglobin levels (mmol/L) at presentation at the emergency department. Since hemoglobin levels are not collected by NSQIP, hemoglobin levels in mmol/L (Hb) were calculated by converting hematocrit (Ht) using the following formula: Hb (mmol/L) = (Ht/2.941) $\times$ 0.6206 [14].

It is recommended that a sufficient sample size is used to externally validate existing prediction models, to facilitate possible recalibration of a prediction model [15]. We aimed to have 400 events, and given an expected in-hospital mortality of 4% (based on the study in which the prediction model was developed), the minimum required sample size for this study was estimated to be at least 10,000 hip fracture patients aged 70 years or above.

For all baseline characteristics, nominal variables were described with numbers and percentages and survivors and deceased patients were compared with a Chi-square test. Descriptive statistics were used to report numeric variables. Normality was determined for continuous variables by examining the boxplots and histograms. Normally distributed data were tested using a Students paired *t*-test and presented as a mean $\pm$ standard deviation (SD). Non-normally distributed data were tested with a Mann-Whitney *U test* and presented as a median with an interquartile range (IQR).

Missing data were analyzed. A total of 142 cases had one or more missing values in predictor variables needed for the algorithm ($<$0.1% of all data points) and were not included for the validation of the model ($<$0.1% of all patients). The authors chose not to impute missing data but instead do a complete case analysis for validation of the model. This resulted in a minimal loss of data.

Discrimination of the model was measured with the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, a measure commonly referred to as the c-statistic, including a 95% confidence interval (CI). Calibration of the model was examined by means of a calibration plot, which plots the predicted probabilities (based on the model) versus the observed risk of the outcome [16–18]. Calibration was quantified by determining the calibration slope of the calibration curve and determining calibration-in-the-large defined as the difference of the mean predicted probability and observed risk of the outcome, which is a measure for predictions being systematically too low or too high. For a perfect model, calibration-in-the large equals 0 and the calibration slope equals 1 [17,18].

If a low predictive accuracy was found during this validation study (i.e. c-statistic $<$0.70, calibration-in-the-large $>$0.02 or $<-$0.02, and/or calibration slope $<$0.75), the model was to be updated or recalibrated by either; 1) intercept recalibration, 2) recalibrating all predictors simultaneously, or 3) adding a new predictor variable (i.e. the presence of dementia at baseline) [19]. However, predictive accuracy was not found to be low and hence the model was not updated nor recalibrated.

The authors performed a subgroup analysis for the performance of the model in the group of patients aged 70–85 years (domain validation) and the group of patient aged 85 or older (original domain). The threshold for significance was set at 0.05. All analyses were conducted using SPSS Statistics (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.), except for the calibration curve analysis, which was conducted using R statistical package for Windows version 3.6 (R foundation, 2019, Vienna, Austria).

Results

Baseline characteristics

A total of 31,751 geriatric hip fracture patients were considered for this study. After exclusion of 6249 patients, 25,502 patients were included in this study (Fig. 1). The overall in-hospital mortality was 2.4% (Table 1). Patients who experienced in-hospital mortality were older at baseline (median 88, IQR 83–90) than survivors (median 85, IQR 79–90, $p<0.01$). Survivors were more often female (72%) than deceased patients (57%, $p<0.01$). Patients who experienced in-hospital mortality more often had a diagnosis of dementia (45%) than survivors (33%, $p<0.01$), and a higher ASA status (median 4, IQR 3–4) than survivors (median 3, IQR 3–3, $p<0.01$). Patients who died in-hospital also had lower levels of serum hemoglobin at presentation ($p<0.01$).

Performance of the model: discrimination and calibration

The mean predicted probability of in-hospital mortality was 3.9% (range 0%–55%) and the observed risk of in-hospital mortality was 2.4%. Calibration-in-the-large was 0.015. Graphical evaluation of the calibration plot showed that the model is fairly well-calibrated in the validation cohort, with a calibration slope of 0.780 (Fig. 2). The c-statistic of the model was 0.74 (95% CI 0.72–0.76).

A subgroup analysis was done for 11,617 patients who were aged between 70 and 85 years to investigate the discrimination

Table 1
Characteristics of geriatric hip fracture patients included in the external validation study of the U-HIP model.

Variable	Total (n = 25,502)	Missing	In-hospital mortality Survivors (n = 24,884)	Deceased (n = 618)	p-value
Age; median, IQR)	85 (80–90)	0	85 (79–90)	88 (83–90)	<0.01
Female sex; n (%)	18,248 (72)	0	17,894 (72)	354 (57)	<0.01
Dementia; n (%)	8567 (34)	0	8288 (33)	279 (45)	<0.01
ASA classification; n (%)		41 (0)			<0.01
ASA I	70 (0)		70 (0)	0 (0)	
ASA II	3458 (14)		3440 (14)	18 (3)	
ASA III	16,052 (63)		15,796 (64)	256 (42)	
ASA IV	5881 (23)		5539 (22)	342 (56)	
Serum hemoglobin (mmol/L) at presentation; mean (SD)	7.3 (1.1)	101 (0)	7.3 (1.1)	6.9 (1.2)	<0.01
Type of hip fracture; n (%)		433 (2)			0.25
Femoral neck, nondisplaced	2115 (9)		2108 (9)	47 (8)	
Femoral neck, displaced	7512 (30)		7349 (30)	163 (27)	
Intertrochanteric	13,987 (56)		13,631 (56)	356 (59)	
Subtrochanteric	1415 (6)		1376 (6)	39 (6)	

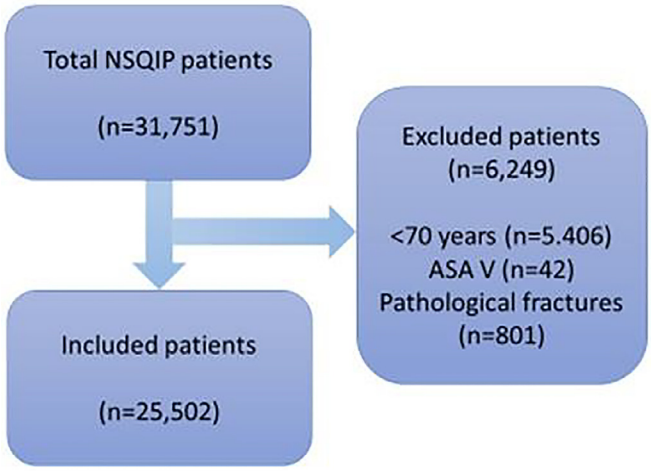

Fig. 1. Selection of patients in external validation study of the U-HIP model.

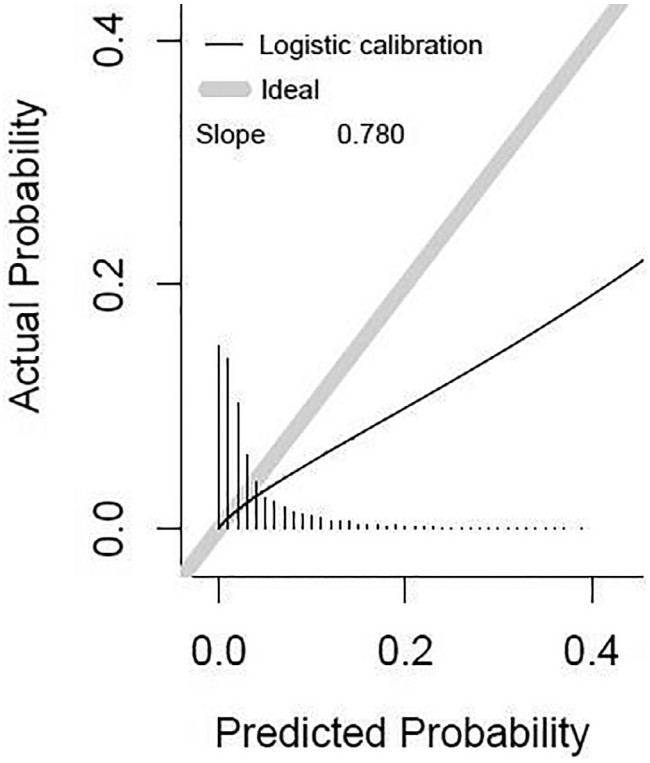

Fig. 2. Calibration plot of observed mortality risk versus predicted risk based on the U-HIP model.

and calibration of the model for this group specifically, because patients in this age category were not included in the development cohort. In this age subgroup, 197 patients died (1.7%), while the mean predicted probability of mortality was 2.2%. Calibration-in-the-large was 0.005. The calibration was good, with a calibration slope of 0.934 (Fig. 3). The c-statistics was 0.78 (95% CI 0.75–0.81).

There were 13,885 patients in the age group of patients aged 85 years or above. In this subgroup, 421 (3.1%) patients died in-hospital, while the mean predicted probability was 5.4%. Calibration-in-the-large was 0.022. The calibration was moderate in this group, with a calibration slope of 0.743 (Supplemental Figure 1). The c-statistic was 0.70 (95% CI 0.68–0.73).

Discussion

The aim of this study was to validate a previously developed prediction model (U-HIP) for in-hospital mortality in geriatric hip fracture patients [8]. In this large retrospective cohort study using NSQIP data, the model showed a fairly good predictive performance.

Development cohort versus validation cohort

The baseline characteristics of this external validation cohort were comparable to the development cohort in terms of sex, ASA classification, diagnosis of dementia and hemoglobin levels [8]. There were several differences between the development cohort and this external validation cohort. First, in the development cohort, no truncation for age was used for nonagenarians and centennials, whereas in the development cohort, no truncation was used. Second, this validation cohort included patients who received total hip arthroplasty as a treatment for their hip fracture, while the development cohort did not. It was not possible to conduct a subgroup analysis for this group. It is possible that there are differences in in-hospital mortality and model performance for total hip arthroplasty patients, considering that patients who are eligible for total hip arthroplasty are often in a better medical condition. Third, in this validation cohort, periprosthetic fractures were included. It was not possible to conduct a subgroup analysis for periprosthetic fractures because they do not have an identifier variable in the NSQIP dataset. Fourth, in this study, the model was validated for patients aged 70 years or above, not just 85 years or above as was the case in the development cohort. Fifth, this external geographical validation study was conducted in a large international

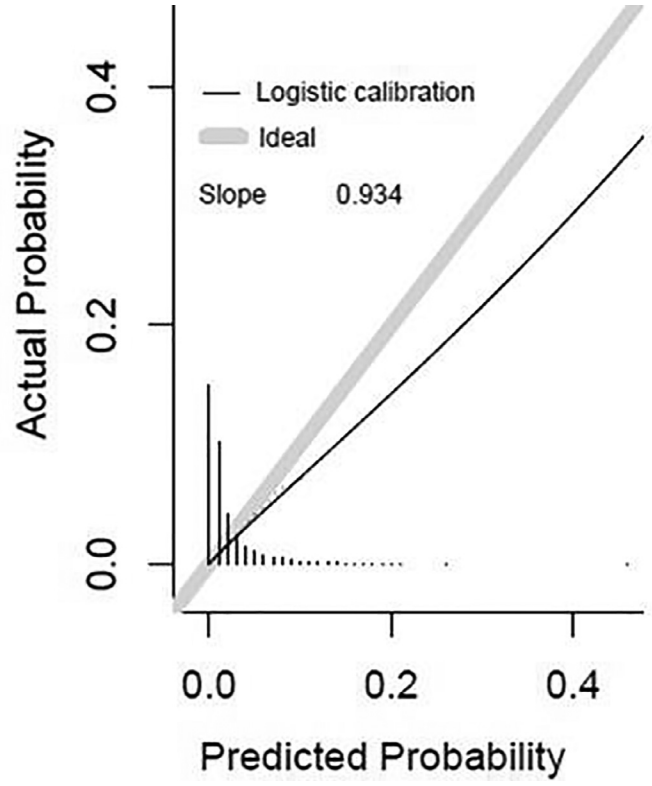

Fig. 3. Calibration plot of observed mortality risk versus predicted risk based on the U-HIP model in the subgroup of patients aged 70–84.

cohort of patients recruited from over 150 hospitals (most of them in Northern-America), whereas the development cohort included geriatric hip fracture patients from six Dutch hospitals.

Performance of the model: discrimination and calibration

The model presented here has a good discrimination and fair calibration. Both metrics are important in prediction modeling. The discrimination of the model was good, with a c-statistic of 0.74, which is comparable to the c-statistic of 0.78 that was found in the development cohort [8]. However, a good discrimination (i.e. separating people who experience a certain outcome from people who do not) alone does not make a good prognostic prediction model. For example, a model may show an excellent discrimination between patients who experience an outcome and patients who do not, but if predicted risk is substantially under- or overestimates of the actual risk of the outcome, the model is usually not suitable for supporting clinical decisions.

Calibration in prediction modeling is defined as the agreement between the observed risk and the predicted risk. There are currently no prediction models for geriatric hip fractures that are well calibrated [3,4,7,20,21]. Calibration is important in prognostic settings, because the magnitude of the predicted risk (or, in the absence of a formal prediction model, the estimated risk) is what drives medical management of our patients [17]. As can be seen in the calibration plots, the calibration is fairly good for the total population and very good for patients aged between 70 and 85 years, and somewhat lower for patients aged 85 years or above. The model tends to slightly overestimate the risk of in-hospital mortality, especially for patients with a higher risk of dying (Fig. 2, Fig. 3). Extremes are always hardest to predict, and it is not uncommon for prediction models to overestimate the higher deciles of the calibration plot [9]. This need not be a problem, given that there are very few patients that fall into these extreme categories, as can be seen in the spikes in Figs. 2 and 3, that show patient distribution according to their predicted probabilities. More importantly, it is unlikely that these overestimations will lead to incorrect medical decision making and thus are of no clinical consequence.

Strengths and limitations

The model presented here is well-calibrated and shows a good discrimination. The advantage of this model in comparison to other models is its predictive accuracy, that it is both internally and externally (even internationally) validated in large cohorts, that it offers exact risk prediction instead of risk stratification, and that it only uses predictor variables that are known at the time the prediction is to be made.

This study has a few limitations. First and foremost, in the NSQIP data, age was truncated at 90 years, which means that patients who are older than 90 years (e.g. 99 years) are entered into the database as being 90 years old. The reason behind this is that a very old age is regarded by the NSQIP as a possible patient identifier, and it is likely that this had led to an underestimation of the discrimination for patients aged 91 years or above. It is likely that many subjects who experienced the outcome were in the age category of 90 years or above, which explains the difference in calibration found between the total population and subgroup analysis of the population aged 75–84 years in this study. The performance of the model in patients aged 90 years or above would probably have been much better if their exact age had been used to develop the model.

Second, in-hospital mortality is frequently used as an outcome in geriatric traumatology, but length of stay and discharge policies can be different between centers, which may impact the performance of a model predicting in-hospital mortality, notably when moribund patients are discharged to other facilities such as hospice care. In our external validation setting, where data of 150 different centers was combined, the impact of this appears to be small.

Clinical application and future perspectives

The U-HIP model is available online for free as a web-based calculator at (https://www.evidencio.com/models/show/2268). Physicians can enter the patient characteristics, and the patients' individual risk of in-hospital mortality is automatically calculated and returned on screen. Alternatively, the model could be programmed into electronic medical records to calculate the mortality risk for each individual patient. Prediction models are useful tools that can be used to complement medical decision making, but not substitute it. The authors recommend a holistic approach for every geriatric hip fracture patient, preferably with geriatric co-management. This model can help guide clinical decision making for these patients, and palliative care should regularly be considered for patients with a very high risk of in-hospital mortality. Additionally, the authors encourage colleagues around the world to perform validation studies for this model in different settings and populations to further investigate model performance.

Conclusion

In this study, a previously developed model for in-hospital mortality in geriatric hip fracture patients was externally validated in a large North-American cohort. The model showed a good discrimination and fair calibration, with good calibration in the subgroup of patients aged 70–85 years. This model is available online as a web-based calculator, and can be used to predict the risk of mortality, identify high-risk patients and thus help guide clinical decision making.

Conflict of interest

No funding or financial support was received for this study the might pose a conflict of interest. Each author certifies that he or she has no commercial associations (e.g., consultancies, stock ownership, equity interest, patent/licensing arrangements, etc.) that might pose a conflict of interest in connection with the submitted article.

Acknowledgements

We would like to thank and acknowledge H. Kelder, MD, PhD (St. Antonius Hospital, department of epidemiology, The Netherlands) for his assistance with the construction of the calibration curves. Additionally, we would like to thank and acknowledge the American College of Surgeons National Surgical Quality Improvement Program for providing data for this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.injury.2021.12.028.

References

[1] Bhandari M, Swiontkowski M. Management of acute hip fracture. N Engl J Med 2017;377:2053–62. doi:10.1056/NEJMcp1611090.
[2] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98:683–90. doi:10.1136/heartjnl-2011-301246.
[3] Jiang HX, Majumdar SR, Dick DA, Moreau M, Raso J, Otto DD, et al. Development and initial validation of a risk score for predicting in-hospital and 1-year mortality in patients with hip fractures. J Bone Miner Res 2005;20:494–500. doi:10.1359/JBMR.041133.
[4] Endo A, Baer HJ, Nagao M, Weaver MJ. Prediction model of in-hospital mortality after hip fracture surgery. J Orthop Trauma 2018;32:34–8. doi:10.1097/BOT.0000000000001026.
[5] Karres J, Heesakkers NA, Ultee JM, Vrouenraets BC. Predicting 30-day mortality following hip fracture surgery: evaluation of six risk prediction models. Injury 2015;46:371–7. doi:10.1016/j.injury.2014.11.004.
[6] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127–41. doi:10.1002/sim.2331.
[7] Nijmeijer WS, Folbert EC, Vermeer M, Slaets JP, Hegeman JH. Prediction of early mortality following hip fracture surgery in frail elderly: the Almelo Hip Fracture Score (AHFS). Injury 2016;47:2138–43. doi:10.1016/j.injury.2016.07.022.
[8] Schuijt HJ, Smeeing DPJ, Würdemann FS, Hegeman JH, Geraghty OC, Houwert RM, et al. Development and internal validation of a prediction model for in-hospital mortality in geriatric hip fracture patients. J Orthop Trauma 2020. doi:10.1097/BOT.0000000000001851.
[9] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691–8. doi:10.1136/heartjnl-2011-301247.
[10] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73. doi:10.7326/M14-0698.
[11] Fink AS, Campbell DAJ, Mentzer RMJ, Henderson WG, Daley J, Bannister J, et al. The National Surgical Quality Improvement Program in non-veterans administration hospitals: initial demonstration of feasibility. Ann Surg 2002;236:344. doi:10.1097/00000658-200209000-00011.
[12] Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and dislocation classification compendium—2018. J Orthop Trauma 2018;32:S1–10.
[13] Owens MDWD, Felts MDJA, Spitznagel PhDEL. ASA physical status classifications: a study of consistency of ratings. Anesthesiol J Am Soc Anesthesiol 1978;49:239–43.
[14] Wennecke G. Hematocrit-A review of different analytical methods. Radiom Med ApS; 2004. https://acutecaretesting.org/en/articles/hematocrit–a-review-of-different-analytical-methods.
[15] Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23:2567–86. doi:10.1002/sim.1844.
[16] Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. J Biomed Inform 2017;76:9–18. doi:10.1016/j.jbi.2017.10.008.
[17] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31. doi:10.1093/eurheartj/ehu207.
[18] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38. doi:10.1097/EDE.0b013e3181c30fb2.
[19] Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. J Thorac Dis 2019;11:S574–84. doi:10.21037/jtd.2019.01.25.
[20] Maxwell MJ, Moran CG, Moppett IK. Development and validation of a preoperative scoring system to predict 30 day mortality in patients undergoing hip fracture surgery. Br J Anaesth 2008;101:511–17. doi:10.1093/bja/aen236.
[21] Karres J, Kieviet N, Eerenberg JP, Vrouenraets BC. Predicting early mortality after hip fracture surgery: the hip fracture estimator of mortality Amsterdam. J Orthop Trauma 2018;32:27–33. doi:10.1097/BOT.0000000000001025.