

Generalizability of a prediction model for walking ability at discharge in older adults after hip fracture surgery: a multicenter study

Yasushi Kurobe

Fujimi-Kogen Hospital, Fujimi-Kogen Medical Center,

Keisuke Nakamura

keipons55@yahoo.co.jp

Shinshu University

Naoko Ushiyama

Fujimi-Kogen Hospital, Fujimi-Kogen Medical Center,

Kimito Momose

Shinshu University

Article

Keywords: Hip fractures, Rehabilitation, Clinical Prediction models, Prognosis, Ambulation

Posted Date: August 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10428287/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 6th, 2026.
See the published version at <https://doi.org/10.1038/s41598-026-73542-3>.

Abstract

This study aimed to evaluate the generalizability of a clinical prediction model for walking ability at discharge from acute-care hospitals after hip surgery in older patients. This retrospective multicenter study was conducted at acute-care hospitals. A total of 2,548 older patients registered in a hip fracture database involving 19 hospitals who were able to walk independently before injury and underwent surgery were included. To evaluate generalizability, we conducted internal–external cross-validation (IECV) across the participating hospitals. The performance of the prediction model was evaluated using the area under the curve (AUC), calibration intercepts, and calibration slopes. The generalizability of the model was assessed using 95% prediction intervals (PIs). Among the 2,014 patients included in the final analysis, 38.2% regained independent walking ability at discharge. Across IECV, the AUC was 0.81 (95% PI: 0.62–0.92). Calibration slopes were 0.96 (95% PI: 0.34–1.58) and calibration intercepts were 0.11 (95% PI: –1.56 to 1.77). Given the marked heterogeneity observed across clusters, our findings indicate that the generalizability of the prediction model for walking ability at discharge after hip fracture surgery in older adults is limited, highlighting the impact of between-hospital heterogeneity on model performance.

1. Introduction

Hip fractures are common among older adults, and the number of hip fracture cases is expected to reach 1.8 million by 2050 globally.[1] Several patients undergoing rehabilitation after hip fracture surgery fail to regain their pre-fracture ambulatory function.[2, 3] This decline in mobility reduces independence in daily activities and imposes a substantial burden on patients, families, and caregivers.[4, 5] Thus, early and accurate prediction of functional outcomes is essential for supporting clinicians, patients, and families in setting realistic rehabilitation goals, making shared decisions, individualizing interventions, prioritizing treatments, facilitating discharge planning, and monitoring progress.[4, 6, 7]

Clinical prediction models are statistical tools used to estimate current or future health outcomes through statistical modeling.[6] Quantifying the predictive accuracy of prediction models in different samples or settings is recommended to assess the generalizability of clinical prediction models.[7–9] Traditional external validation, which evaluates a model using a single independent dataset, may not fully capture the variability across clinical environments. To overcome this limitation, internal–external cross-validation (IECV) has been proposed as a robust method for assessing the generalizability of prediction models in medical research. One of the strengths of this approach is that it enables repeated assessment of model generalizability using multicenter data.[10, 11]

Several clinical prediction models have been developed to predict walking ability after hip fracture surgery in acute-care settings.[12, 13] Nonetheless, only a few studies have evaluated the generalizability of these models in older adults with hip fractures, and externally validated models available for clinical use remain limited.[12, 13] Therefore, in this study, we aimed to evaluate the generalizability of a clinical prediction model for walking ability at discharge from an acute-care hospital after hip surgery in older

adults. Given the limited evidence from external validation in previous studies, developing and evaluating highly generalizable models that can be applied across diverse clinical settings is required.

2. Methods

2.1 Study design and participants

This study was a retrospective analysis of data derived from Japan's Nagano Hip Fracture Database, which contains information on patients' prehospitalization characteristics, in-hospital treatments, and rehabilitation outcomes at discharge following surgery. As of May 2024, 19 hospitals have participated in the database.[14]

The inclusion criteria were as follows: (1) age ≥ 60 years, (2) diagnosis of femoral neck or intertrochanteric fracture, (3) having undergone surgery at an acute-care hospital, and (4) independent indoor ambulation before injury (with or without walking aids). The exclusion criterion was a cluster of < 40 patients.[15]

2.2 Definitions of variables

Patients' background information included age, sex, body mass index, medical history (respiratory, cardiovascular, and neurological diseases), cognitive function before injury, pre-fracture residence, fracture type, surgical procedure, and rehabilitation factors (e.g., postoperative days until the start of ambulation). Pre-injury mobility was defined based on the type of assistive device used during indoor ambulation. Preoperative waiting time was defined as the number of days from hospital admission to surgery; surgery performed within 2 days was classified as early surgery, whereas surgery performed after > 2 days was considered delayed surgery.[16] Postoperative ambulation timing was defined as the number of days from surgery to the first walking session; patients who initiated ambulation within 2 days postoperatively were categorized as early walkers, whereas those who initiated ambulation after > 2 days were classified as delayed walkers.[14]

2.3 Ethical considerations

This study was conducted in accordance with the principles embodied in the Declaration of Helsinki and adhered to the Transparent Reporting of Multivariable Prediction Models Developed or Validated Using Clustered Data (TRIPOD-Cluster) statement [17]. All procedures were approved by the Ethics Committee of the School of Medicine, Shinshu University (protocol number: 4541; approval date: November 12, 2019). The requirement for individual informed consent was waived by the ethics committee because this observational study used anonymized data obtained during routine clinical practice. Information about the study was disclosed at each participating site, and participants were provided with the opportunity to opt out.

2.4 Sample size calculation

Patients included in the Nagano Hip Fracture Database were not originally recruited to develop a prediction model for gait recovery. Thus, the sample size for this study was determined based on the number of eligible patients in the database. We considered the sample size and number of clusters to ensure a reliable estimation of between-cluster heterogeneity. Meta-analyses are recommended to include ten or more clusters when each cluster has a small-to-medium (40–400) sample size to achieve reliable heterogeneity estimates.[15] In this study, 11 clusters met the small-to-medium sample size criterion, providing an adequate number of clusters for analysis.

2.5 Outcomes

The primary outcome of this study was walking ability at discharge from an acute-care hospital. Independent walking ability was defined as a Functional Independence Measure (FIM) locomotion score of ≥ 5 , as assessed during walking evaluation, whereas dependent walking ability was defined as an FIM score of ≤ 4 or death at discharge.[14, 18] Physical therapists conducted the assessments.

2.6 Predictors

The following potential predictors were selected based on previous studies: age, fracture type, indoor walking ability before injury, cognitive function, and residence. Candidate predictors measurable at baseline and expected to be assessed preoperatively were considered. Indoor walking ability before injury was defined according to the type of walking aids used (T-cane, walker, or without walking aids) during indoor ambulation. Cognitive function was assessed using the Degree of Daily Life Independence Score for People with Dementia (DDLIS-PD), a seven-level scale commonly used for evaluating dementia in Japan.[19] In this study, cognitive impairment was defined as DDLIS-PD grade II or higher, indicating partial independence and some difficulty.

A data manager initially assessed the submitted data to identify potential outliers and missing values. Any identified issues were communicated to the contributing institutions for verification and correction, and a revised dataset was subsequently resubmitted by these institutions. After confirming the integrity of the data, the cleaned dataset was finalized and used for subsequent analyses. All patient characteristics and study variables were defined at the time of data collection, and uniform criteria and definitions were applied across all participating sites.

In accordance with current recommendations,[17] the risk of bias during both the development and validation phases of the prediction model was assessed using the Risk Of Bias ASsessment Tool for Artificial Intelligence (PROBAST + AI).[20] PROBAST + AI evaluates the risk of bias across four domains: participants, predictors, outcomes, and analyses. Two reviewers independently conducted the assessment, and discrepancies were resolved through discussion. Given the multicenter nature of the dataset, special attention was paid to the consistency of data collection procedures and potential heterogeneity across sites. During the IECV, any discrepancies in the definitions or measurements of key variables were systematically assessed and reported compared to those used in the development dataset.

2.7 Handling of missing data

The proportion of missing data in the overall dataset and within each cluster is presented in Table 1. Missing values were imputed using the missForest method, which is a random forest–based imputation technique.[21]

Table 1
Baseline characteristics of the study cohort

Age (years), mean (SD)	Overall (n = 2014)	Missing
	85.3 (8.4)	0 (0.0)
Sex, female, n (%)	1561 (77.5)	0 (0.0)
BMI (kg/m ²), mean (SD)	20.4 (3.5)	17 (0.8)
Pre-fracture residence, n (%)		0 (0.0)
Own home	1643 (81.6)	
Nursing facility	371 (18.4)	
Prefracture walking ability, n (%)		0 (0.0)
Without aids	1161 (57.6)	
With cane	390 (19.4)	
With walker	463 (23.0)	
Cognitive impairment, n (%)	906 (45.0)	13 (0.6)
Comorbidity, n (%)		
Respiratory disease	187 (9.3)	1 (0.0)
Cardiovascular disease	494 (24.5)	1 (0.0)
Neurological disease	282 (14.0)	1 (0.0)
Fracture type, n (%)		0 (0.0)
Neck	1041 (51.7)	
Intertrochanter	973 (48.3)	
Surgical procedure, n (%)		0 (0.0)
Osteosynthesis	1328 (65.9)	
Hip hemiarthroplasty	686 (34.1)	
Abbreviation: BMI, body mass index		

2.8 Statistical analysis

Patient characteristics, postoperative rehabilitation data, and candidate predictors were summarized for the entire cohort and separately for each hospital. Continuous variables were reported as means with standard deviations (SDs), whereas categorical variables were expressed as frequencies with percentages.

The IECV approach was used to evaluate between-hospital heterogeneity and assess the generalizability of the prediction model. A logistic regression model that included all candidate predictors was used. In each iteration, one of the 10 hospitals was excluded from the development dataset, and the model was developed using data from the remaining hospitals and then validated on the omitted hospital. In developing the prediction model, the intercept and regression coefficients were assumed to be fixed across all clusters. The functional form of the association with age was treated as a continuous variable, and the outcome was assessed using scatter plots prior to model development. Logarithmic transformation was applied to age to better represent the relationship with the outcome. All other variables were considered categorical. A random-effects meta-analysis was conducted to calculate the summary estimates of performance metrics across hospitals, and heterogeneity was assessed using forest plots. Model performance was evaluated based on the area under the receiver operating characteristic curve (AUC, also referred to as the c-statistic), calibration slope, and calibration intercept, which are commonly used to assess the discrimination and calibration of prediction models. Summary estimates of the performance measures were reported as point estimates with 95% confidence intervals (CIs). Additionally, 95% prediction intervals (PIs) were calculated to examine the relevance of between-hospital heterogeneity. The 95% CI reflects the precision of the average model performance across all hospitals, whereas the 95% PI accounts for the between-hospital heterogeneity and indicates the expected performance in a new hospital. If the average performance and generalizability were deemed acceptable, a global model was developed by estimating the predictor coefficients using data from all hospitals. In this final step, no clusters were excluded, thereby minimizing the variability in the coefficient estimates.

The case-mix similarity of baseline risk was examined to interpret the sources of heterogeneity across hospitals. Case-mix similarity between clusters was summarized using the mean and SD of the linear predictor (LP) values. The mean of the LP reflects the absolute risk within each hospital, whereas the SD of the LP reflects heterogeneity in the case mix.[22]

In addition, a sensitivity analysis was conducted to assess the impact of outcome timing. Specifically, we developed and validated an alternative model using the walking ability at 1 week postoperatively as the outcome measure.

All statistical analyses were performed using R software version 4.4.3 (R Foundation for Statistical Computing, Vienna, Austria) using “metamisc,”[23] “missforest,”[21] and “metafor”[24] packages.

3. Results

A total of 2,548 eligible patients across 19 centers in Japan were included in the Nagano Hip Fracture Database between December 2019 and May 2024. Overall, 16 clusters comprising 2,086 patients met the inclusion criteria. Five clusters with < 40 patients (72 patients in total) were excluded. The final analysis included 11 clusters and 2,014 patients (Fig. 1). The baseline characteristics of the study participants are presented in Table 1 and Supplementary Table S1. The mean length of hospital stay was 29.9 days. At discharge, 769 (38.2%) patients were able to walk independently. At 1 week postoperatively, 375 (18.6%) patients regained independent walking ability (Table 2 and Supplementary Table 2).

Table 2
Characteristics of interventions and clinical outcomes in the study cohort

LOS after surgery (days), mean (SD)	Overall (n = 2014)	Missing
	26.1 (23.3)	39 (1.9)
Days from admission to surgery, n (%)		2 (0.1)
Early surgery	823 (40.9)	
Delayed surgery	1189 (59.0)	
Days to first walking in rehabilitation after surgery, n (%)		185 (9.2)
Early walk	656 (32.6)	
Delayed walk	1173 (58.2)	
Postoperative weight bearing status, n (%)		2 (0.1)
Full weight-bearing	1897 (94.2)	
Restricted weight-bearing	115 (5.7)	
Locomotion ability at discharge, n (%)		44 (2.2)
Able to walk	769 (38.2)	
Unable to walk	1201 (59.6)	
Locomotion ability at 1 week after surgery, n (%)		78 (3.9)
FIM ≥ 5	375 (18.6)	
FIM ≤ 4	1561 (77.5)	
Abbreviation: LOS, length of stay; FIM, Functional Independence Measure		

During the model development phase, the risk of bias was judged to be low in the “participants” and “predictors” domains; however, it was rated high in the “outcomes” domain because of concerns that the

outcome assessors might have had access to predictor data at the time of assessment. The risk of bias in this analysis was low. Applicability concerns were low across all domains. Consequently, the development phase was judged to have a high risk of bias overall, with a low concern for applicability. A similar pattern was observed during the model validation phase. The risk of bias remained low for the “participants” and “predictors” domains, whereas the “outcomes” domain was again rated as high risk for the same concerns regarding assessor access to predictor information. The “analysis” domain was rated as low risk, and applicability concerns were consistently low across all domains. Therefore, the overall judgment of the validation phase mirrored that of the development phase, with a high risk of bias and low concern for applicability (Supplementary Table 3).

Forest plots showing the model discrimination (AUC) and calibration metrics (calibration slope and intercept) from the IECV of the prognostic model for walking status at discharge in the development cohort are presented in Fig. 2. The random-effects meta-analysis yielded a summary estimate of 0.81 (95% CI: 0.76–0.86; 95% PI: 0.62–0.92) for the AUC. The summary estimates were 0.96 (95% CI: 0.75–1.18; 95% PI: 0.34–1.58) for the calibration slope and 0.11 (95% CI: –0.39 to 0.60; 95% PI: –1.56 to 1.77) for the calibration intercept.

Similarly, forest plots showing model discrimination (AUC) and calibration metrics (calibration slope and calibration intercept) from the IECV of the prognostic model predicting walking ability at 1 week postoperatively in the development cohort are presented in Fig. 3. The summary estimate for the AUC was 0.84 (95% CI: 0.81–0.87; 95% PI: 0.78–0.89). The summary estimates for the calibration slope and intercept were 1.00 (95% CI: 0.88–1.12; 95% PI: 0.88–1.12) and –0.13 (95% CI: –0.47 to 0.21; 95% PI: –1.18 to 0.92), respectively. The mean and SD of LP distribution were similar across all clusters (Fig. 4).

Owing to the substantial heterogeneity across sites and the poor generalizability observed during the IECV, we did not develop or present a global model.

4. Discussion

We assessed the generalizability of a clinical prediction model for ambulatory independence at discharge using data from 2,014 patients who underwent surgery for hip fractures. The wide 95% PIs for discrimination and calibration across clusters observed in the IECV indicated the model’s limited generalizability in predicting walking ability at discharge after hip fracture surgery.

One of the strengths of this study is the use of a large multicenter dataset that reflects real-world clinical practices across diverse hospital settings. The application of the IECV framework enabled a rigorous evaluation of the model’s generalizability and assessment of the performance heterogeneity across clusters. To the best of our knowledge, few previous studies have explicitly accounted for clustering in the model validation of postoperative mobility outcomes, highlighting the unique contribution of this study to this field.

This study examined the generalizability of a prediction model for independent walking after hip fracture surgery using the IECV framework. The results revealed substantial heterogeneity in the model performance across clusters in terms of discrimination (AUC) and calibration (slope and intercept). This finding contrasts with those of many previous studies that did not account for the clustered structure of data.[25–27] Consequently, the prediction models developed in these studies may have underestimated the impact of inter-hospital variability on model performance. Our findings suggest that although the model performs well in certain settings, its predictive accuracy varies across hospitals, likely due to differences in patient characteristics, care processes, and timing of outcome assessment. The observed heterogeneity highlights the importance of evaluating prediction models using cluster-aware validation methods such as IECV, especially when models are intended for application across diverse healthcare settings.

When interpreting the model validation results, distinguishing between the different aspects of model generalizability, such as reproducibility and transportability, is essential.[28] In this study, the mean and SD of LP were consistent across hospitals (Fig. 4), indicating similar distributions of predicted probabilities of achieving independent walking and supporting the reproducibility of the model across institutions. In contrast, the timing of gait assessment at discharge varied among hospitals (Supplementary Table S2). This variation, known as follow-up period transportability,[28] likely affects the transportability of the model. The wide 95% PIs indicate substantial heterogeneity in model performance, likely driven by inter-hospital differences in discharge timing. The likelihood of ambulatory independence has been reported by previous studies to increase over time after hip fracture surgery.[29, 30] Therefore, even with similar patient characteristics, patients in hospitals with a shorter length of stay may be discharged before regaining independent ambulation, whereas those in hospitals with a longer length of stay may achieve mobility prior to discharge. This discrepancy in discharge timing likely contributed to the limited generalizability of the model. Supporting this interpretation, a sensitivity analysis using walking ability at one week postoperatively, standardized across all hospitals, demonstrated reduced heterogeneity in the discrimination and calibration metrics. These findings suggest that anchoring outcome assessment to a fixed postoperative time point rather than discharge may enhance the generalizability of prediction models by reducing the variability introduced by institutional practices.

This study has some limitations. First, although several candidate predictors were selected based on the previous literature, data constraints limited the inclusion of all variables considered in the literature. Consequently, the influence of some predictors (physical health status[31] and nutritional status[32]) might not have been fully evaluated. Nonetheless, this study incorporated key factors known to be associated with postoperative functional outcomes after hip fracture, including age, cognitive impairment, pre-injury functional status, and pre-fracture ambulatory ability.[12, 13, 33, 34] Therefore, the impact of the excluded variables on the predictive accuracy is likely limited. Second, the differences in postoperative care were not considered. Specifically, information regarding the content and quality of postoperative rehabilitation and care was unavailable; therefore, it was not incorporated into the prediction model. Although postoperative interventions may have a considerable impact on patient outcomes, we aimed to develop a model based on the information available at the time of prediction to

support early clinical decision-making. Thus, although excluding postoperative care variables is a limitation, it may enhance the applicability of the model to real-world clinical settings.

5. Conclusions

This study evaluated the generalizability of the prediction model using the IECV. Model performance varied across institutions, and consistent predictive accuracy was not observed, indicating limitations in the model's generalizability. Additionally, the analysis focused solely on the occurrence of outcomes without considering time-to-event data, and the temporal aspects of predictive performance were not assessed. Future studies should incorporate survival analysis methods to capture time-dependent risks, thereby enhancing the clinical utility of the prediction model.

Declarations

Disclosures

Competing interests

The authors declare no competing interests

Author Contribution

Y.K. conceptualized the study, developed the methodology, performed the formal analysis, and wrote the original draft of the manuscript. N.K. supervised the study, administered the project, curated the data, and reviewed and edited the manuscript. N.U. contributed to reviewing and editing the manuscript. K.M. supervised the study, administered the project, and reviewed and edited the manuscript. All authors contributed to the interpretation of the data, critically reviewed the manuscript, and approved the final version.

Acknowledgement

The authors thank all the rehabilitation staff and cooperating hospitals for their contributions to data collection.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

1. Sing, C. W. et al. Global epidemiology of hip fractures: secular trends in incidence rate, post-fracture treatment, and all-cause mortality. *J. Bone Min. Res.* **38**, 1064–1075 (2023).
2. Dyer, S. M. et al. A critical review of the long-term disability outcomes following hip fracture. *BMC Geriatr.* **16**, 158 (2016).
3. van Balen, R. et al. Hip fracture in elderly patients: outcomes for function, quality of life, and type of residence. *Clin. Orthop. Relat. Res.* **390**, 232–243 (2001).
4. Asif, M., Cadel, L., Kuluski, K., Overall, A. C. & Guilcher, S. J. T. Patient and caregiver experiences on care transitions for adults with a hip fracture: a scoping review. *Disabil. Rehabil.* **42**, 3549–3558 (2020).
5. Amarilla-Donoso, F. J. et al. Quality of life in elderly people after a hip fracture: a prospective study. *Health Qual. Life Outcomes.* **18**, 71 (2020).
6. Tousignant-Laflamme, Y. et al. Mastering prognostic tools: an opportunity to enhance personalized care and to optimize clinical outcomes in physical therapy. *Phys. Ther.* **102**, pzac023 (2022).
7. Steyerberg, E. W. et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* **10**, e1001381 (2013).
8. Efthimiou, O. et al. Developing clinical prediction models: a step-by-step guide. *BMJ* **386**, e078276 (2024).
9. Collins, G. S. et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* **384**, e074819 (2024).
10. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
11. Takada, T. et al. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. *J. Clin. Epidemiol.* **137**, 83–91 (2021).
12. Kaizu, Y. et al. Clinical prediction models for nonmortality outcomes in older adults with hip fractures: a systematic review. *J. Gerontol. Biol. Sci. Med. Sci.* **78**, 2363–2370 (2023).
13. Walsh, M. E. et al. Systematic review of multivariable prognostic models for outcomes at least 30 days after hip fracture finds 18 mortality models but no nonmortality models warranting validation. *J. Clin. Epidemiol.* **173**, 111439 (2024).
14. Nakamura, K. et al. Impact of early postoperative ambulation on gait recovery after hip fracture surgery: a multicenter cohort study. *Sci. Rep.* **15**, 12893 (2025).
15. Langan, D. et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res. Synth. Methods.* **10**, 83–98 (2019).
16. Cohn, M. R. et al. Factors associated with early functional outcome after hip fracture surgery. *Geriatr. Orthop. Surg. Rehabil.* **7**, 3–8 (2016).
17. Debray, T. P. A. et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. *BMJ* **380**, e071018 (2023).

18. Fu, G. et al. Rapid preoperative predicting tools for 1-year mortality and walking ability of Asian elderly femoral neck fracture patients who planned for hip arthroplasty. *J. Orthop. Surg. Res.* **16**, 455 (2021).
19. Nakanishi, K. et al. New score including daily life independence levels with dementia is associated with the onset of deep vein thrombosis in frail older adults. *Geriatr. Gerontol. Int.* **20**, 414–421 (2020).
20. Moons, K. G. M. et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025).
21. Stekhoven, D. J. & Bühlmann, P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
22. Debray, T. P. A. et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J. Clin. Epidemiol.* **68**, 279–289 (2015).
23. de Jong, V. M. T., Moons, K. G. M., Eijkemans, M. J. C., Riley, R. D. & Debray, T. P. A. Developing more generalizable prediction models from pooled studies and large clustered data sets. *Stat. Med.* **40**, 3533–3559 (2021).
24. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).
25. Hirose, J. et al. Prediction of postoperative ambulatory status 1 year after hip fracture surgery. *Arch. Phys. Med. Rehabil.* **91**, 67–72 (2010).
26. Doherty, W. J. et al. Prediction of postoperative outcomes following hip fracture surgery: independent validation and recalibration of the nottingham hip fracture score. *J. Am. Med. Dir. Assoc.* **22**, 663–9e2 (2021).
27. de González, C. et al. Predictive model of gait recovery at one month after hip fracture from a national cohort of 25,607 patients: the hip fracture prognosis (HF-Prognosis) tool. *Int. J. Environ. Res. Public. Health.* **18**, 3809 (2021).
28. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Ann. Intern. Med.* **130**, 515–524 (1999).
29. Vochteloo, A. J. H. et al. More than half of hip fracture patients do not regain mobility in the first postoperative year. *Geriatr. Gerontol. Int.* **13**, 334–341 (2013).
30. Magaziner, J. et al. Recovery from hip fracture in eight areas of function. *J. Gerontol. Biol. Sci. Med. Sci.* **55**, M498–507 (2000).
31. Chen, L. H. et al. The relationship between preoperative American Society of Anesthesiologists Physical Status Classification scores and functional recovery following hip-fracture surgery. *BMC Musculoskelet. Disord.* **18**, 410 (2017).
32. Malafarina, V. et al. Nutritional status and nutritional treatment are related to outcomes and mortality in older adults with hip fracture. *Nutrients* **10**, 555 (2018).

33. Kristensen, M. T. Factors affecting functional prognosis of patients with hip fracture. *Eur. J. Phys. Rehabil Med.* **47**, 257–264 (2011).
34. McGilton, K. S. et al. Factors influencing outcomes of older adults after undergoing rehabilitation for hip fracture. *J. Am. Geriatr. Soc.* **64**, 1601–1609 (2016).

Figures

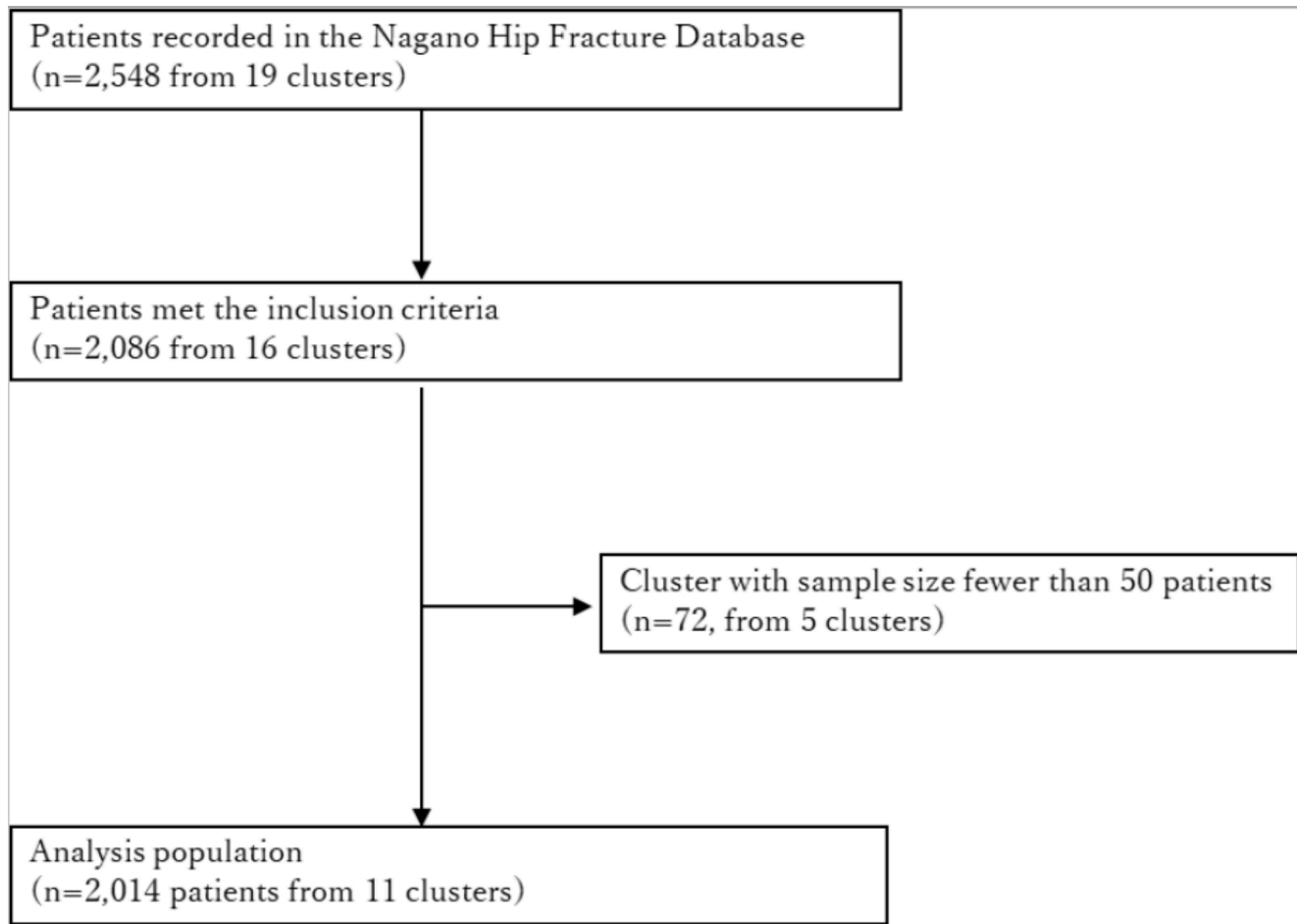


Figure 1

Flow diagram of patient and hospital selection.

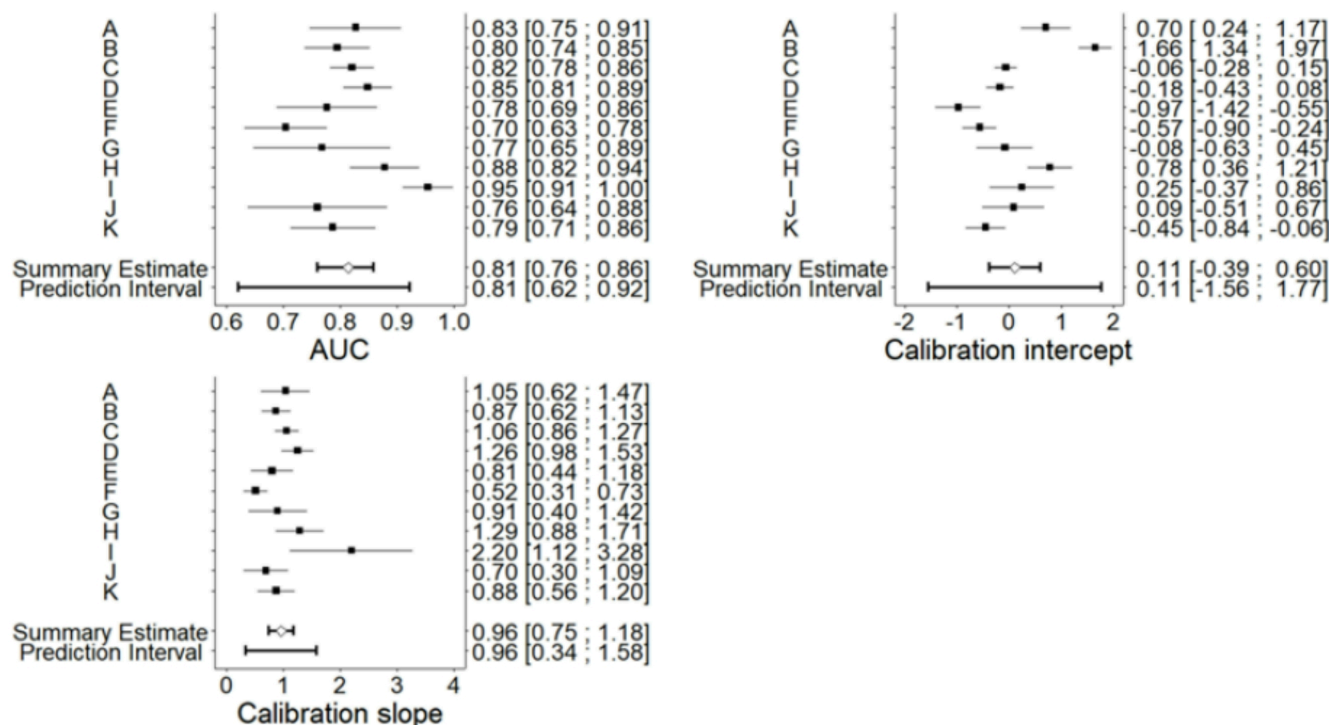


Figure 2

Summary estimates and prediction intervals (PIs) of model performance at discharge across hospitals. Forest plots showing the discrimination (AUC), calibration intercept, and calibration slope of the prediction model across the external validation datasets (A–K). Point estimates (square markers) and 95% confidence intervals (horizontal lines) are presented for each study. The diamond indicates the summary estimate across studies, whereas the horizontal bar below it represents the 95% PI. The AUC reflects model discrimination, whereas the calibration intercept and slope evaluate the agreement between the predicted and observed outcomes. A calibration intercept of 0 and a slope of 1 indicate perfect calibration.

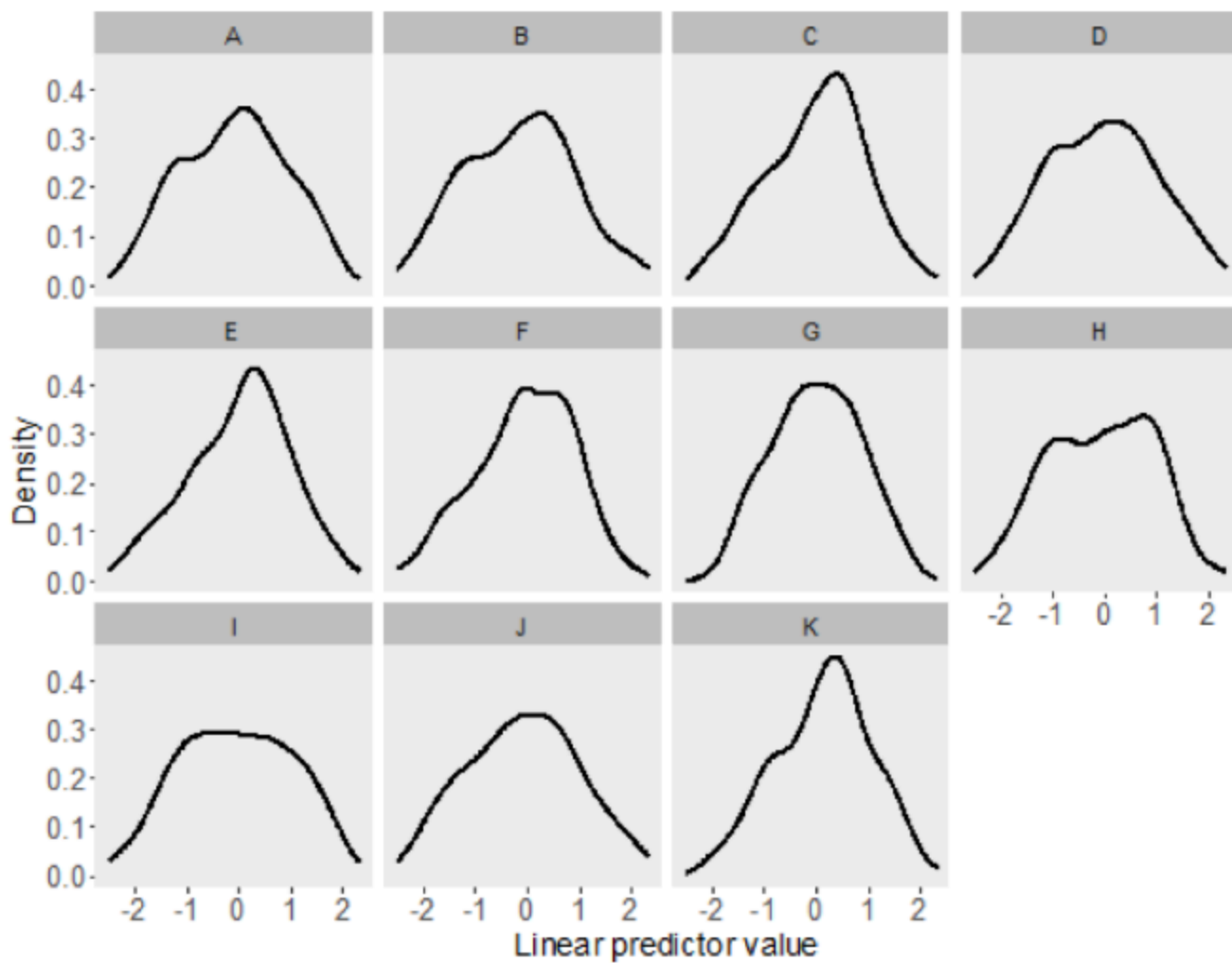


Figure 3

Distribution of linear predictor (LP) values by hospital. Density plots of the LP values are shown for each of the 11 hospitals (Panels A–K). The LP was derived from the prediction model using selected baseline predictors. Each curve represents the distribution of LP of independent walking at discharge for patients in a given hospital. Differences in distribution reflect variations in patient case mix across hospitals.

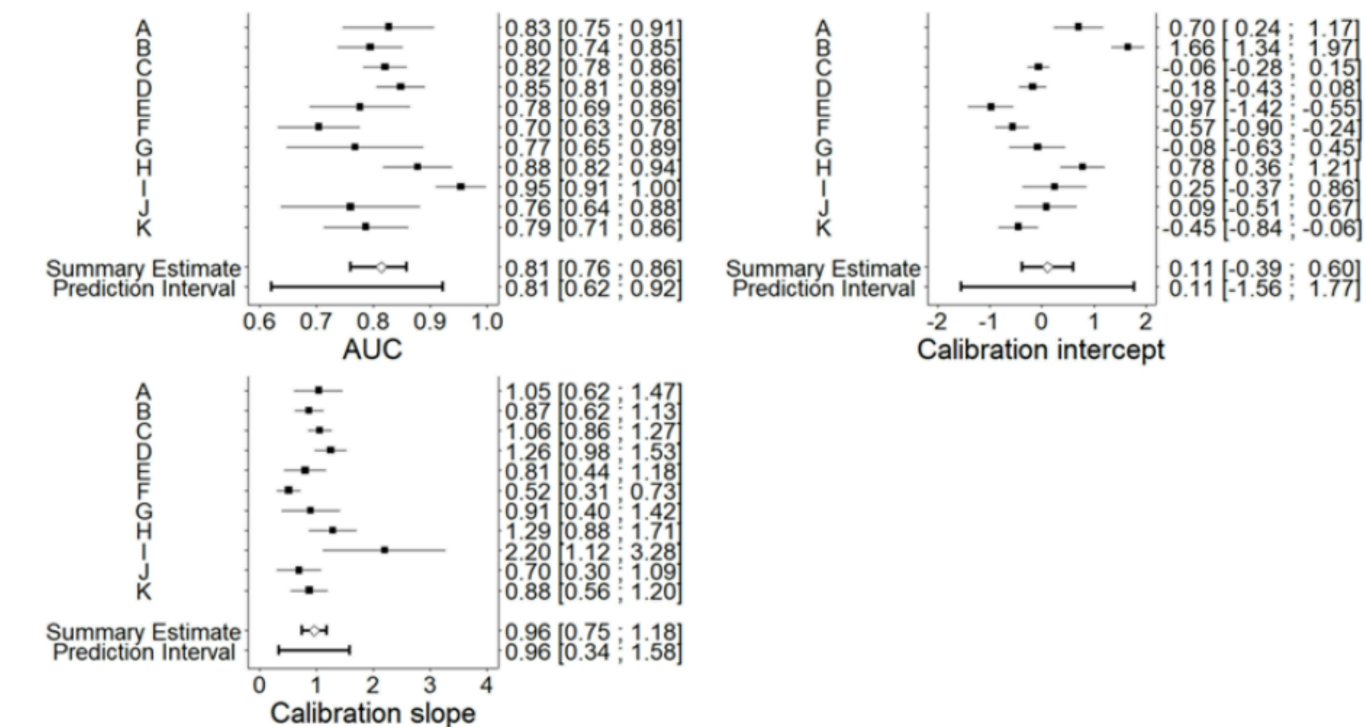


Figure 4

Summary estimates and prediction intervals (PIs) of model performance at 1 week after surgery across hospitals. Forest plots show the discrimination (AUC), calibration intercept, and calibration slope of the prediction model across the external validation datasets (A–K). Point estimates (square markers) and 95% confidence intervals (horizontal lines) are presented for each study. The diamond indicates the summary estimate across studies, whereas the horizontal bar below it represents the 95% PI. The AUC reflects the model discrimination, whereas the calibration intercept and slope evaluate the agreement between the predicted and observed outcomes. A calibration intercept of 0 and a slope of 1 indicates perfect calibration.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.pdf](#)