Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comprehensive Geriatric Assessment–Augmented Risk Stratification for Postoperative Delirium After Spine Surgery: A Cross-Surgery Benchmarking and Model-Updating Study

Seung Jun Ryu
Daejeon Eulji Medical Center, Eulji University College of Medicine
Ji Yeon Baek
Asan Medical Center, University of Ulsan College of Medicine
Ji Hyeon Kim
Daejeon Eulji Medical Center, Eulji University College of Medicine
Jung Woo Hur
The Catholic University of Korea
Woo Keun Kwon
Korea University Guro HospitalKorea University Guro Hospital, Korea University College of Medicine
Bong Ju Moon
Gangnam Severance HospitalGangnam Severance Hospital, Yonsei University College of Medicine
Jae Keun Oh
Hallym University Sacred Heart HospitalHallym University Sacred Heart Hospital, Hallym University College of Medicine
Il Sup Kim
St. Vincent's Hospital, College of Medicine, The Catholic University of Korea
Chung Kee Chough
Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea
Sun Woo Jang
Gangneung Asan Hospital, University of Ulsan College of Medicine
Jin Hoon Park
spinejhpark@naver.com

Asan Medical Center, University of Ulsan College of Medicine

Research Article

Keywords: Postoperative delirium, spine surgery, comprehensive geriatric assessment, frailty, model updating, transportability, geriatric nutritional risk index, prediction model

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10428412/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Postoperative delirium (POD) after spine surgery in older adults reflects both baseline geriatric vulnerability and perioperative stress. Although externally derived electronic health record models may provide useful benchmarks, their direct transportability across surgical domains is uncertain. We evaluated whether comprehensive geriatric assessment (CGA)–augmented local model updating improves POD risk stratification in older adults undergoing spine surgery.

Methods

This retrospective benchmarking and model-updating study used an in-house spine-surgery cohort of 162 older adults, including 16 POD events, and a publicly available external hip-fracture cohort of 245 patients, including 30 POD events. A harmonized predictor set included demographics, comorbidities, nutritional and laboratory reserve, operative burden, and postoperative context. Dataset shift was assessed using standardized mean differences and a cohort-membership model, followed by overlap weighting. Predictor transportability was examined using cohort-specific univariable log odds ratios. Prediction models were evaluated in staged fashion: external-to-spine transport, recalibration, local core updating, CGA augmentation, and a simplified bedside score. Discrimination, calibration, Brier score, and decision-curve analysis were assessed.

Results

POD occurred in 30 of 245 patients in the external hip-fracture cohort and 16 of 162 patients in the in-house spine cohort. Among POD cases, the spine cohort was younger but had longer anesthesia or surgery duration. Before weighting, the cohorts showed substantial case-mix imbalance, with a maximum absolute standardized mean difference of 1.72; overlap weighting reduced the maximum residual imbalance to 0.16. The cohort-membership model showed near-separation between cohorts (AUROC 0.98). The transported external model showed modest discrimination in the spine cohort (AUROC 0.62), and recalibration did not improve discrimination (AUROC 0.62). Local updating improved AUROC to 0.88, and CGA augmentation further improved AUROC to 0.94. CGA vulnerability showed a graded association with POD incidence across quartiles: 0/41, 1/40, 6/40, and 9/41. A simplified frozen bedside score retained good discrimination (AUROC 0.86). At a score threshold of ≥ 3, sensitivity was 94%, specificity 70%, positive predictive value 25%, and negative predictive value 99%.

Conclusions

Risk domains from an external hip-fracture cohort showed partial conceptual transportability to older spine-surgery patients, but local updating with measured CGA variables substantially improved risk stratification. A CGA-augmented bedside score may provide a practical rule-out tool for targeting delirium-prevention resources, although prospective multicenter validation is needed.

INTRODUCTION

Postoperative delirium (POD) is a clinically important complication in older adults undergoing spine surgery. In this setting, POD can interfere with postoperative care, early mobilization, and rehabilitation, and has been associated with longer hospitalization, higher resource use, and worse clinical outcomes in older surgical patients (Fineberg et al., 2013; Raats et al., 2015). Reported POD incidence after spine surgery varies across cohorts, reflecting differences in patient age, baseline vulnerability, surgical complexity, and delirium surveillance intensity (Zhang et al., 2020).

POD after spine surgery is best understood as a multidomain syndrome rather than the consequence of a single perioperative exposure. Prior spine-surgery studies have identified both baseline vulnerability factors, including older age and impaired preoperative cognitive function, and perioperative stressors, including longer operative duration and transfusion, as clinically relevant risk domains (Kang et al., 2020). More recent work also suggests that neurovascular vulnerability may contribute to POD risk: in elderly patients undergoing elective spine surgery, increased internal carotid artery pulsatility index predicted POD even after adjustment for age and surgical duration (Saar et al., 2024). Together, these findings support a framework in which POD reflects the interaction between reduced cognitive or physiological reserve and acute surgical, anesthetic, hemodynamic, and inflammatory burden.

Comprehensive geriatric assessment (CGA) provides a clinically intuitive approach for capturing multidimensional vulnerability before surgery. However, full CGA can be resource-intensive and difficult to implement routinely in busy perioperative settings (Raats et al., 2015). For elective spine surgery, a pragmatic bedside approach that incorporates brief cognitive screening, functional status, frailty-related markers, medication burden, nutritional reserve, and selected perioperative factors may therefore be more feasible than a full specialist-led CGA. Emerging neurovascular and immune-marker studies in elderly spine surgery patients further support the biological relevance of vascular, inflammatory, and neuronal-injury pathways, but these approaches remain exploratory and may be limited in routine clinical use by assay availability, technical requirements, and the need for external validation (Baek et al., 2023; Saar et al., 2024; Vogelgesang et al., 2026). Thus, specialized biomarker approaches and bedside clinical risk scores should be viewed as complementary rather than competing strategies.

A remaining challenge is whether POD risk domains observed in spine surgery reflect procedure-specific risk or a broader geriatric susceptibility to delirium shared across older orthopedic surgical populations. Hip-fracture cohorts differ from elective spine surgery cohorts in acuity, trauma burden, baseline function, and perioperative care pathways, and therefore should not be treated as direct external validation datasets for a spine-specific model. Nevertheless, hip-fracture data may provide an

informative external comparator for examining whether vulnerability domains such as cognitive impairment, frailty, renal function, nutritional reserve, transfusion, and operative stress show conceptual or directional consistency across older orthopedic populations (Zhao et al., 2021; Zhao, 2022).

Therefore, this study aimed to develop and evaluate an exploratory CGA-augmented POD risk stratification score for older adults undergoing spine surgery. Using an in-house spine-surgery cohort enriched with bedside cognitive assessment, instrumental activities of daily living, frailty measures, medication variables, laboratory markers, and perioperative surgical variables, we sought to identify clinically interpretable POD risk domains. We further examined whether these domains showed conceptual and directional transportability in a publicly available hip-fracture delirium dataset. We hypothesized that a low-cost CGA-augmented score would capture the combined contribution of baseline geriatric vulnerability and perioperative surgical stress, providing a pragmatic framework for identifying older spine-surgery patients at increased POD risk.

METHODS

Study design and data sources

This retrospective benchmarking and exploratory model-updating study evaluated POD risk prediction in an in-house spine-surgery cohort and examined the conceptual transportability of selected risk domains using a publicly available hip-fracture delirium dataset. The study was structured according to prediction-model reporting principles, with prespecified descriptions of data sources, outcome definition, predictor handling, model specification, calibration, internal validation, and clinical utility assessment (Collins et al., 2015).

The external comparator dataset was the public hip-fracture cohort reported by Zhao et al., which included patients aged ≥ 65 years undergoing hip-fracture repair under spinal or general anesthesia between June 1, 2018 and May 31, 2019 at Peking University People's Hospital. This dataset included 245 patients, of whom 30 developed POD. The target dataset was an in-house spine-surgery cohort of 162 patients with available POD outcome information, including 16 POD events. The in-house study was approved by the institutional review board of [institution name; IRB number], with [waiver/requirement] of informed consent because of the retrospective design.

Outcome definition

The primary outcome was POD during the index hospitalization. In the external hip-fracture cohort, POD was assessed by trained nurses using the Confusion Assessment Method every 12 hours during hospitalization (Zhao et al., 2021). In the in-house spine cohort, POD was ascertained by structured review of clinical documentation, including nursing records, physician notes, delirium-related diagnostic coding, rescue antipsychotic use, and psychiatric consultation. Variables that could occur after delirium onset were used only for outcome ascertainment and were not entered as primary predictors.

Predictor harmonization and CGA-derived vulnerability score

A common predictor set was harmonized across the external and in-house datasets before analysis. Harmonized variables included age, sex, body mass index, hypertension, diabetes mellitus, coronary artery disease, stroke or transient ischemic attack, albumin, creatinine, hematocrit, anesthesia duration, blood product use, intensive care unit admission, and a chart-derived frailty proxy. Laboratory values were converted to common clinical units when required. The chart-derived frailty proxy was calculated as the sum of five routinely available vulnerability indicators: age > 70 years, body mass index < 18.5 kg/m², hematocrit < 35%, albumin < 3.4 g/dL, and creatinine > 2.0 mg/dL. In the in-house spine cohort, additional geriatric variables were available, including Mini-Cog, instrumental activities of daily living disability, gait speed, grip strength, Short Physical Performance Battery, CHS frailty category, polypharmacy, benzodiazepine use, and depression. These variables were used to characterize measured geriatric vulnerability and to construct a compact CGA-derived vulnerability score.

The CGA-derived vulnerability score was constructed as a continuous composite index ranging theoretically from 0 to 1, with higher values indicating greater vulnerability. Each component was transformed onto a 0–1 vulnerability scale before aggregation. The cognitive component was calculated as (5 − Mini-Cog score) / 5. The functional component was calculated as the number of impaired instrumental activities of daily living items divided by 6. Slow gait was coded as 1 when gait speed was < 0.8 m/s and 0 otherwise. Low physical performance was coded as 1 when the Short Physical Performance Battery score was ≤ 9 and 0 otherwise. Frailty was represented by the available CHS frailty category indicator as recorded in the source dataset. Medication and neuropsychiatric components were coded as binary indicators for polypharmacy, defined as use of ≥ 5 prescription medications, depression, and benzodiazepine use. For each patient, the CGA-derived vulnerability score was calculated as the arithmetic mean of the available component scores: cognition, IADL disability, slow gait, low SPPB, CHS frailty category, polypharmacy, depression, and benzodiazepine use. Thus, the denominator was the number of non-missing components available for that patient. Missing component values were not imputed before score construction. In the final n = 162 analytic cohort, Mini-Cog, IADL disability, gait speed, SPPB, polypharmacy, depression, and benzodiazepine use were available for all patients, whereas CHS frailty category was missing in a subset of patients and was therefore excluded from the denominator for those individuals. Grip strength was retained for descriptive and correlation analyses of measured geriatric phenotypes, but raw grip strength was not directly included in the composite score because the corresponding CHS grip-weakness component was unavailable in the analytic dataset.

For descriptive risk enrichment in Fig. 4D, patients were ordered according to the CGA-derived vulnerability score and divided into four rank-based quartiles, with Q1 representing the lowest vulnerability and Q4 the highest vulnerability. Rank-based quartiles were used to preserve approximately equal group sizes in the presence of tied score values. These quartiles were used for descriptive comparison of observed POD incidence, whereas the continuous CGA-derived vulnerability score was used in the CGA-augmented prediction model.

Statistical analysis

Continuous variables were summarized using mean and standard deviation or median and interquartile range, as appropriate. Categorical variables were summarized as counts and percentages. Between-cohort differences were quantified using standardized mean differences rather than hypothesis tests, because the primary purpose was to describe covariate imbalance and dataset shift. For continuous variables, standardized mean differences were calculated using pooled standard deviations; for binary variables, pooled binomial variances were used.

To assess covariate shift between the hip-fracture and spine-surgery cohorts, a cohort-membership model was fitted using the harmonized predictor set, with cohort membership as the dependent variable. The predicted probability of belonging to the in-house spine cohort was used to quantify overlap between cohorts. Overlap weights were calculated as the probability of belonging to the opposite cohort: for hip-fracture patients, the weight was the predicted probability of belonging to the spine cohort; for spine-surgery patients, the weight was one minus that probability. Weighted standardized mean differences were then used to describe balance in the empirically comparable region of the two cohorts (Li et al., 2018).

Univariable predictor transportability was evaluated by estimating the association between each harmonized predictor and POD separately in each cohort. Continuous predictors were standardized within cohort and modeled per 1-standard-deviation increase. Binary predictors were summarized using odds ratios, with sparse-cell correction when needed. Directional agreement of predictor effects between cohorts was summarized using the Spearman correlation of univariable log odds ratios.

Prediction models were fitted using penalized logistic regression with L2 regularization. Missing predictor values were handled within the modeling pipeline using median imputation, followed by standardization of continuous predictors. Penalized regression was selected to reduce overfitting risk given the limited number of POD events in the in-house spine cohort.

Model evaluation followed a staged transport-and-updating framework (Binuya et al., 2022). First, an external core model was trained in the public hip-fracture cohort using harmonized predictors and applied to the spine cohort without refitting. Second, logistic recalibration was performed in the spine cohort by fitting a recalibration model with the logit-transformed external predicted probability as the only predictor, estimating a new intercept and calibration slope. Third, a local in-house core model was fitted using the harmonized predictors. Finally, a CGA-augmented local model was fitted by adding the CGA-derived vulnerability score to the harmonized predictor set.

Model performance in the in-house spine cohort was assessed using discrimination, calibration, overall prediction error, and clinical utility. Discrimination was quantified using the area under the receiver operating characteristic curve and the area under the precision-recall curve. Overall prediction error was quantified using the Brier score. Calibration was evaluated using calibration intercept and calibration slope from a logistic calibration model, with ideal values of 0 and 1, respectively. Calibration plots were

generated by grouping predicted probabilities into five quantile-based bins and comparing mean predicted risk with observed POD incidence.

Internal validation of the local core and CGA-augmented models was performed using bootstrap optimism correction with 80 resamples. In each bootstrap sample, the complete modeling pipeline was refitted, model performance was evaluated in both the bootstrap sample and the original cohort, and the average optimism was subtracted from apparent performance to obtain optimism-corrected AUROC, AUPRC, and Brier score.

Clinical utility was assessed using decision curve analysis across threshold probabilities from 0.03 to 0.40. Net benefit was calculated by weighting false-positive classifications according to the odds of the selected threshold probability, and model-based strategies were compared with treat-all and treat-none reference strategies (Vickers & Elkin, 2006). A simplified bedside score was constructed as an integer score by assigning one point for each of the following clinically available risk indicators: GNRI < 98, Mini-Cog ≤ 3, any IADL disability, polypharmacy, transfusion or blood product exposure, anesthesia/surgery duration ≥ 300 minutes, creatinine ≥ 1.2 mg/dL, and coronary artery disease. Patients were stratified into low-risk, intermediate-risk, and high-risk groups using prespecified score categories of 0–2, 3–4, and ≥ 5 points, respectively. A candidate rule-out threshold of ≥ 3 points was used to summarize sensitivity, specificity, positive predictive value, negative predictive value, and the confusion matrix.

RESULT

Study population and cross-surgery phenotype

The open hip-fracture benchmark cohort included 245 patients, of whom 30 developed postoperative delirium (POD; 12.2%). The in-house spine analytic cohort included 162 patients, of whom 16 developed POD (9.9%). The overall POD incidence did not differ significantly between cohorts (p = 0.46). The analytic framework compared an open external hip-fracture cohort with an in-house spine cohort using harmonized domains available in both datasets, while comprehensive geriatric assessment (CGA) and medication-related variables were available only in the in-house spine cohort. Among patients who developed POD, the in-house spine cohort was younger than the external hip-fracture cohort: 71.4 ± 5.6 versus 79.3 ± 7.6 years, p < 0.001. Female sex was less frequent among in-house spine POD cases, although this difference did not reach statistical significance: 7/16, 44% versus 21/30, 70%, p = 0.082. Body size differed substantially between POD cases in the two cohorts. In-house spine POD cases had higher weight, 68.3 ± 12.3 versus 57.7 ± 11.2 kg, p = 0.004, and higher BMI, 26.5 ± 3.8 versus 22.5 ± 3.8 kg/m², p < 0.001. Nutritional reserve also differed between POD cases. GNRI was higher in the in-house spine POD group than in the external hip-fracture POD group, 95.7 ± 3.4 versus 91.2 ± 8.2, p = 0.009. However, the proportion with GNRI < 98 did not differ significantly, 11/16, 69% versus 24/30, 83%, p = 0.279. Albumin was numerically higher in the in-house spine POD group, 3.7 ± 0.2 versus 3.5 ± 0.4 g/dL, p = 0.053. Comorbidity and perioperative stress profiles showed clear domain-specific differences. Stroke/TIA was more frequent among external hip-fracture POD cases than in-house spine POD cases,

12/30, 40% versus 1/16, 6%, p = 0.015. Transfusion or blood product exposure was also more frequent in the external POD group, 19/30, 63% versus 3/16, 19%, p = 0.004. In contrast, anesthesia or surgery duration was markedly longer in the in-house spine POD group, 334.8 ± 72.2 versus 197.3 ± 59.0 minutes, p < 0.001. ICU admission was numerically lower in the in-house spine POD group but did not differ significantly, 3/16, 19% versus 9/30, 30%, p = 0.408.

Dataset shift before benchmarking and effect of overlap weighting

Before benchmarking, the two cohorts showed marked case-mix imbalance across demographic, nutritional, hematologic, comorbidity, and operative-stress domains. The largest pre-weighting imbalances were observed for anesthesia duration, chart-derived vulnerability, age, hematocrit, GNRI, BMI, albumin, blood product exposure, sex, stroke/TIA, coronary artery disease, and ICU admission. The maximum absolute standardized mean difference was 1.72 before weighting. The cohort-membership model showed near-separation between the open hip-fracture and in-house spine cohorts, with an AUROC of 0.98, indicating substantial dataset shift before weighting.

Overlap weighting substantially reduced this imbalance and targeted the region of empirical overlap between the two surgical populations. After overlap weighting, most covariates moved close to zero standardized mean difference, and the maximum residual absolute standardized mean difference was reduced to 0.16. The largest remaining imbalance was anesthesia duration, but it was substantially attenuated compared with the unweighted comparison. This supported the use of weighted benchmarking as a transportability-oriented comparison rather than a naive direct cohort comparison.

Predictor transportability across surgical domains

The predictor transportability atlas demonstrated partial, rather than complete, transportability of POD predictors across surgical domains. Directional concordance between univariable predictor effects was modest (Spearman ρ = 0.37). Several predictors showed directionally concordant risk signals across the external hip-fracture and in-house spine cohorts, including ICU admission, blood product or transfusion exposure, lower GNRI or GNRI < 98, chart-derived vulnerability, lower albumin, and longer anesthesia or surgery duration.

However, the magnitude of predictor effects differed by surgical context. In the open hip-fracture cohort, perioperative and vascular-context variables such as blood product exposure, stroke/TIA, ICU admission, and anesthesia duration showed prominent associations with POD. In the in-house spine cohort, coronary artery disease, lower GNRI, chart-derived vulnerability, lower albumin, older age, and anesthesia duration showed stronger local signals. Coronary artery disease was particularly context-shifted, showing a stronger positive association in the in-house spine cohort than in the open hip-fracture cohort. These findings indicate that the broader risk domains were transportable, but individual predictor weights required local updating.

Bridging chart-derived vulnerability with measured CGA vulnerability

Within the in-house spine cohort, chart-derived vulnerability was higher among patients who developed POD than among those who did not, supporting its role as a pragmatic vulnerability proxy. The correlation heatmap showed that chart-derived vulnerability was related to measured geriatric domains, including CGA score, reversed Mini-Cog, IADL disability, slow gait, grip weakness, low SPPB, polypharmacy, benzodiazepine use, and depression. These findings suggest that the chart-derived signal captures part, but not all, of the broader measured geriatric vulnerability construct.

Measured CGA vulnerability showed a graded relationship with observed POD incidence across quartiles. POD occurred in 0/41 patients in Q1, 1/40 in Q2, 6/40 in Q3, and 9/41 in Q4, corresponding to observed POD incidences of 0%, 2.5%, 15.0%, and 22.0%, respectively. These quartile counts sum exactly to the N = 162 analytic cohort and account for all 16 POD events.

Local CGA updating improved model performance beyond transported external benchmarking. In the incremental updating panel, AUROC, AUPRC, and Brier score were 0.62, 0.16, and 0.12 for the transported external core model; 0.62, 0.16, and 0.09 after recalibration; 0.88, 0.51, and 0.07 for the in-house core model; 0.94, 0.64, and 0.06 for the in-house CGA-augmented model; and 0.86, 0.33, and 0.08 for the frozen bedside score. Bootstrap optimism-corrected AUROC was 0.75 for the in-house core model and 0.84 for the CGA-augmented model.

Model updating and discrimination

External-to-spine transport showed modest discrimination. The external core model achieved an AUROC of 0.62 when applied to the in-house spine cohort, and intercept recalibration did not materially change discrimination, with AUROC remaining 0.62. Local model updating improved discrimination, with the in-house core model achieving an AUROC of 0.88. Adding CGA-based geriatric variables further improved discrimination to an AUROC of 0.94. A simplified frozen bedside score retained good discrimination, with an AUROC of 0.86.

Calibration plots showed that the transported external model exhibited calibration drift when applied to the spine cohort. Recalibration improved average risk alignment but did not improve discrimination. In contrast, the in-house core and CGA-augmented models showed improved alignment between predicted and observed POD risk. Decision-curve analysis showed that the CGA-augmented model provided the most consistent net benefit across clinically relevant threshold probabilities, while the bedside score preserved practical utility for prevention-bundle targeting.

Bedside score risk stratification and rule-out performance

The bedside score stratified the N = 162 in-house spine cohort into clinically distinct risk groups using prespecified categories of 0–2, 3–4, and ≥ 5 points. Observed POD incidence was 1/103 in the low-risk group, 10/48 in the intermediate-risk group, and 5/11 in the high-risk group, corresponding to approximately 1.0%, 20.8%, and 45.5%, respectively. Thus, the bedside score identified a large low-risk

group with very low observed POD incidence while concentrating events in the intermediate- and high-risk strata.

Using a rule-out threshold of score ≥ 3, the confusion matrix showed 102 true low-risk non-POD patients, 44 high-risk non-POD patients, 1 low-risk POD patient, and 15 high-risk POD patients. This threshold yielded sensitivity 94%, specificity 70%, positive predictive value 25%, and negative predictive value 99%. In matrix-derived terms, 15 of 16 POD events were detected and 102 of 103 low-score patients were POD-free. These findings suggest that the bedside score is most useful as a high-sensitivity rule-out tool for identifying patients unlikely to develop POD, while patients above the threshold may be candidates for targeted delirium-prevention bundles.

Discussion

In this exploratory cross-surgery benchmarking study, we developed and evaluated a CGA-augmented postoperative delirium risk stratification framework for older adults undergoing spine surgery. The principal finding was that risk domains derived from an external hip-fracture cohort showed only partial transportability to the in-house spine cohort, whereas local updating with directly measured geriatric vulnerability substantially improved model performance and clinical usefulness. The external core model showed modest discrimination after transport to the spine cohort, while the locally updated model and the CGA-augmented model improved discrimination stepwise. These findings support the concept that POD in older spine-surgery patients is best understood as the convergence of baseline geriatric vulnerability and perioperative stress rather than as the consequence of a single demographic, comorbidity, or operative exposure.

Previous spine-surgery studies have identified multiple POD-related factors, including older age, comorbidity burden, impaired cognition, lower physiologic reserve, longer surgery, transfusion, and postoperative care intensity. However, many existing studies have relied primarily on conventional demographic, laboratory, or comorbidity variables, which may incompletely represent the multidimensional vulnerability of older surgical patients. Our findings extend this literature by demonstrating that measured CGA domains—cognition, IADL disability, physical performance, frailty-related function, medication burden, depression, and sedative exposure—add clinically meaningful information beyond chart-derived predictors. This is consistent with the broader spine-surgery literature showing that POD risk is multifactorial and that cognitive and functional reserve are central components of risk stratification (Luo et al., 2024; Susano et al., 2019).

A key contribution of this study is the bridge between routinely available chart-derived vulnerability and measured geriatric vulnerability. Chart-derived frailty was associated with POD and correlated with measured CGA constructs, but it did not fully capture the CGA signal. This distinction is clinically important because chart-derived variables are scalable and easy to extract, but they may miss vulnerabilities that are visible at the bedside yet poorly encoded in routine EMR data, such as subtle cognitive impairment, reduced instrumental function, gait slowing, polypharmacy, and benzodiazepine

exposure. The graded increase in POD incidence across CGA vulnerability quartiles, from 0/41 in the lowest quartile to 9/41 in the highest quartile, suggests that measured CGA captures a latent vulnerability axis that is directly relevant to POD risk. Prior prospective work in lumbar spine surgery similarly emphasized the importance of baseline cognitive status, particularly lower MMSE score, as a risk factor for POD (Pan et al., 2019).

The external hip-fracture benchmark analysis further clarified why simple model transport is insufficient. Hip-fracture and spine-surgery cohorts shared broad vulnerability domains, including age, nutritional reserve, renal function, transfusion exposure, and operative stress. However, the magnitude and relative importance of individual predictors differed substantially by surgical context. In the external hip-fracture cohort, vascular and acute-care factors such as stroke/TIA, blood product exposure, and ICU admission showed prominent associations with POD. In contrast, the in-house spine cohort showed stronger local signals for coronary artery disease, GNRI, chart-derived vulnerability, albumin, and anesthesia duration. These findings suggest that delirium risk domains may be conceptually transportable across older orthopedic populations, but predictor weights require recalibration or local updating before clinical use. This interpretation is consistent with geriatric surgery evidence showing that frailty-related risk is important but must be contextualized by procedure type, perioperative pathway, and local patient characteristics (Raats et al., 2015).

Nutritional reserve also emerged as a clinically interpretable shared domain. GNRI differed between POD cases in the external hip-fracture and in-house spine cohorts, and low GNRI showed directionally relevant signals across surgical settings. A recent study of elderly patients with degenerative lumbar disease reported that lower preoperative GNRI was independently associated with POD after TLIF, supporting GNRI as a low-cost marker of nutritional vulnerability in spine surgery. In our study, GNRI appeared most useful not as a standalone predictor, but as one component of a broader vulnerability-stress framework that also included cognition, function, medication exposure, and operative burden (Chen et al., 2024).

Our study also helps position CGA-based prediction relative to emerging biomarker approaches. Biomarker studies using urine exosomal proteins or complement pathway profiling may provide important mechanistic insight into neuronal injury, systemic inflammation, immune activation, and neurocognitive vulnerability. However, such assays remain exploratory, require specialized platforms, and are not yet routinely available for perioperative bedside decision-making. By contrast, CGA variables are inexpensive, clinically interpretable, and actionable before surgery. Therefore, the present findings do not argue against biomarker development; rather, they suggest that measured geriatric vulnerability may serve as a practical intermediate layer between fragmented EMR-based prediction and specialized molecular profiling (Baek et al., 2023; Vogelgesang et al., 2026).

The simplified frozen bedside score was designed to translate the CGA-augmented model into a practical clinical tool. Despite simplification, the score retained good discrimination and separated the in-house spine cohort into clinically meaningful risk strata. The low-risk group had only 1 POD event among 103 patients, whereas the intermediate- and high-risk groups showed substantially higher observed POD

incidence. At the score ≥ 3 threshold, the score achieved high sensitivity and a negative predictive value of 99%, suggesting that its most immediate use may be as a rule-out tool. In practice, this could help identify patients unlikely to require intensive prevention-bundle resources, while focusing CGA-informed interventions on patients above the threshold. This clinical framing is aligned with decision-curve methodology, which emphasizes whether a prediction model improves decisions across relevant threshold probabilities rather than discrimination alone (Vickers & Elkin, 2006).

Several limitations should be acknowledged. First, the in-house spine cohort was relatively small and included only 16 POD events. Although penalized modeling, compact score construction, and internal validation were used to reduce overfitting, the high apparent performance of the CGA-augmented model should be interpreted as hypothesis-generating rather than definitive. Second, POD ascertainment differed between datasets. The external hip-fracture cohort used structured CAM-based surveillance, whereas the in-house spine cohort relied on structured review of clinical documentation and outcome-adjacent evidence. This difference may introduce surveillance bias or outcome misclassification. Third, the external hip-fracture cohort should be interpreted as a benchmarking comparator, not a direct external validation cohort for a spine-specific prediction model. Hip-fracture surgery differs from elective spine surgery in acuity, trauma burden, baseline function, and perioperative care pathway. Finally, some anthropometric variables used for GNRI calculation required harmonization across in-house data sources, underscoring the need for prospective standardization of height, weight, albumin, and CGA measurements.

Despite these limitations, this study provides a clinically grounded framework for POD risk stratification in older adults undergoing spine surgery. The results suggest that CGA-augmented local updating can overcome some limitations of both transported external models and purely chart-derived EMR prediction. Future multicenter prospective studies should validate the bedside score, evaluate calibration across institutions, and test whether CGA-guided prevention bundles reduce POD incidence, postoperative complications, length of stay, and non-home discharge. Integration of structured CGA extraction into perioperative clinical decision support systems may ultimately allow older spine-surgery patients to receive risk-adapted preoperative optimization and postoperative monitoring.

In conclusion, a low-cost CGA-augmented bedside score may provide a pragmatic precision-medicine tool for POD risk stratification in older spine-surgery patients. The findings support a convergence model in which baseline geriatric vulnerability and perioperative stress jointly shape delirium risk. Rather than replacing biomarker or EMR-based approaches, measured CGA may connect biological plausibility, bedside interpretability, and actionable perioperative decision-making.

Declarations

Conflicts of interest

The authors declare no conflicts of interest.

Ethics approval

This study was approved by the Institutional Review Board of Asan Medical Center (IRB number: 2025-0013). The requirement for informed consent was obtained because of the retrospective study design.

Data availability statement

The external hip-fracture dataset is publicly available through Mendeley Data. The in-house spine-surgery data are available from the corresponding author upon reasonable request and subject to institutional and ethical restrictions.

Funding

The authors received no specific funding for this work.

Clinical trial number

Not applicable.

References

1. Baek W, Lee J, Jang Y, Kim J, Shin DA, Park H, Koo BN, Lee H. Assessment of risk factors for postoperative delirium in older adults who underwent spinal surgery and identifying associated biomarkers using exosomal protein. J Korean Acad Nurs. 2023;53(4):371–84.
2. Binuya MA, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: A systematic review. BMC Med Res Methodol. 2022;22:316.
3. Chen Q, Zhu C, Ai Y, Wang J, Ding H, Luo D, Li Z, Song Y, Feng G, Liu L. Preoperative geriatric nutritional risk index is useful factor for predicting postoperative delirium among elderly patients with degenerative lumbar diseases. Eur Spine J. 2024;33:1055–60.
4. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis. TRIPOD statement Annals Intern Med. 2015;162(1):55–63.
5. Fineberg SJ, Nandyala SV, Marquez-Lara A, Oglesby M, Patel AA, Singh K. Incidence and risk factors for postoperative delirium after lumbar spine surgery. Spine. 2013;38(20):1790–6.
6. Kang T, Park SY, Lee JH, Lee SH, Park JH, Kim SK, Suh SW. Incidence and risk factors of postoperative delirium after spinal surgery in older patients. Sci Rep. 2020;10:9232.
7. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Am Stat Assoc. 2018;113(521):390–400.
8. Luo M, Wang D, Shi Y, Yi Q, Wang Z, Zhou B, Yang G, Chen J, Liang C, Wang H, Zeng X, Yang Y, Tan R, Xie Y, Chen J, Tang S, Huang J, Mei Z, Xiao Z. Risk factors of postoperative delirium following spine

surgery: A meta-analysis of 50 cohort studies with 1.1 million participants. Heliyon. 2024;10:e24967.

9. Pan Z, Huang K, Huang W, Kim KH, Wu H, Yu Y, Kim KN, Yi S, Shin DA, Vora D, Gragnaniello C, Phan K, Tasiou A, Winder MJ, Koga H, Azimi P, Kang SY, Ha Y. The risk factors associated with delirium after lumbar spine surgery in elderly patients. Quant Imaging Med Surg. 2019;9(4):700–10.
10. Raats JW, van Eijsden WA, Crolla RMPH, Steyerberg EW, van der Laan L. Risk factors and outcomes for postoperative delirium after major surgery in elderly patients. PLoS ONE, 2015; 10(8).
11. Saar A, Müller J, Veser Y, Behr F, Rathmann E, Schroeder HWS, Flöel A, Müller JU, von Sarnowski B, Fleischmann R. Cerebrovascular pulsatility indicates preoperative subcortical cognitive impairment and an increased risk for postoperative delirium in elderly patients undergoing elective spine surgery. Front Med. 2024;11:1433380.
12. Susano MJ, Scheetz SD, Grasfield RH, Cheung D, Xu X, Kang JD, Smith TR, Lu Y, Groff MW, Chi JH, Crosby G, Culley DJ. Retrospective analysis of perioperative variables associated with postoperative delirium and other adverse outcomes in older patients after spine surgery. J Neurosurg Anesthesiol. 2019;31(4):385–91.
13. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
14. Vogelgesang A, Wolf H, Strack S, Flöel A, Schroeder HWS, Müller J, Müller JU, Fleischmann A, Fleischmann R, Pauly D, Ruhnau J. Complement activation as a predictor of postoperative delirium in elderly spine surgery patients. Int J Mol Sci. 2026;27(2):1077.
15. Zhang HJ, Ma XH, Ye JB, Liu CZ, Zhou ZY. Systematic review and meta-analysis of risk factor for postoperative delirium following spinal surgery. J Orthop Surg Res. 2020;15:509.
16. Zhao H. Machine learning delirium (Version 1) [Data set]. Mendeley Data. 2022.
17. Zhao H, You J, Peng Y, Feng Y. Machine learning algorithm using electronic chart-derived data to predict delirium after elderly hip fracture surgeries: A retrospective case-control study. Front Surg. 2021;8:634629.

Table

Table 1 is available in the supplementary files section

Figures

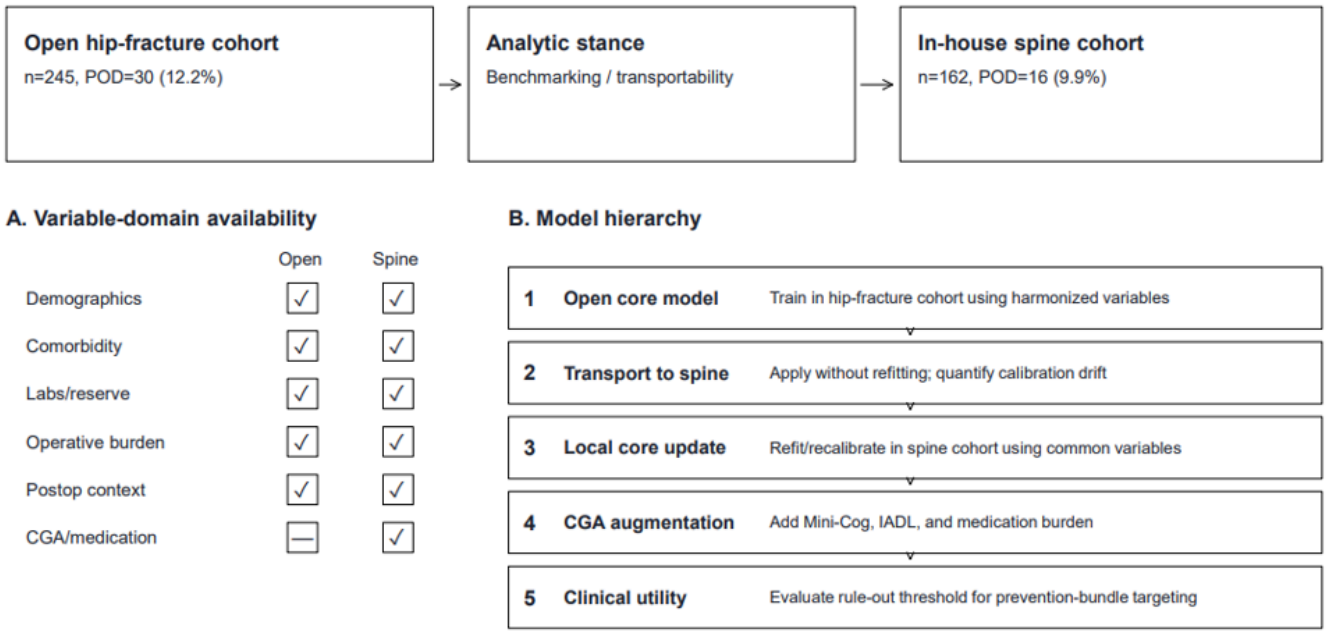


Figure 1

Open-cohort benchmarking framework. Study design for external benchmarking and local model updating. The external hip-fracture cohort included 245 patients with 30 POD events, and the in-house spine cohort included 162 patients with 16 POD events. Harmonized variables were used for cross-cohort transportability analysis, whereas CGA and medication variables were available only in the in-house cohort. Outcome-adjacent variables were excluded from primary prediction models to reduce temporal leakage.

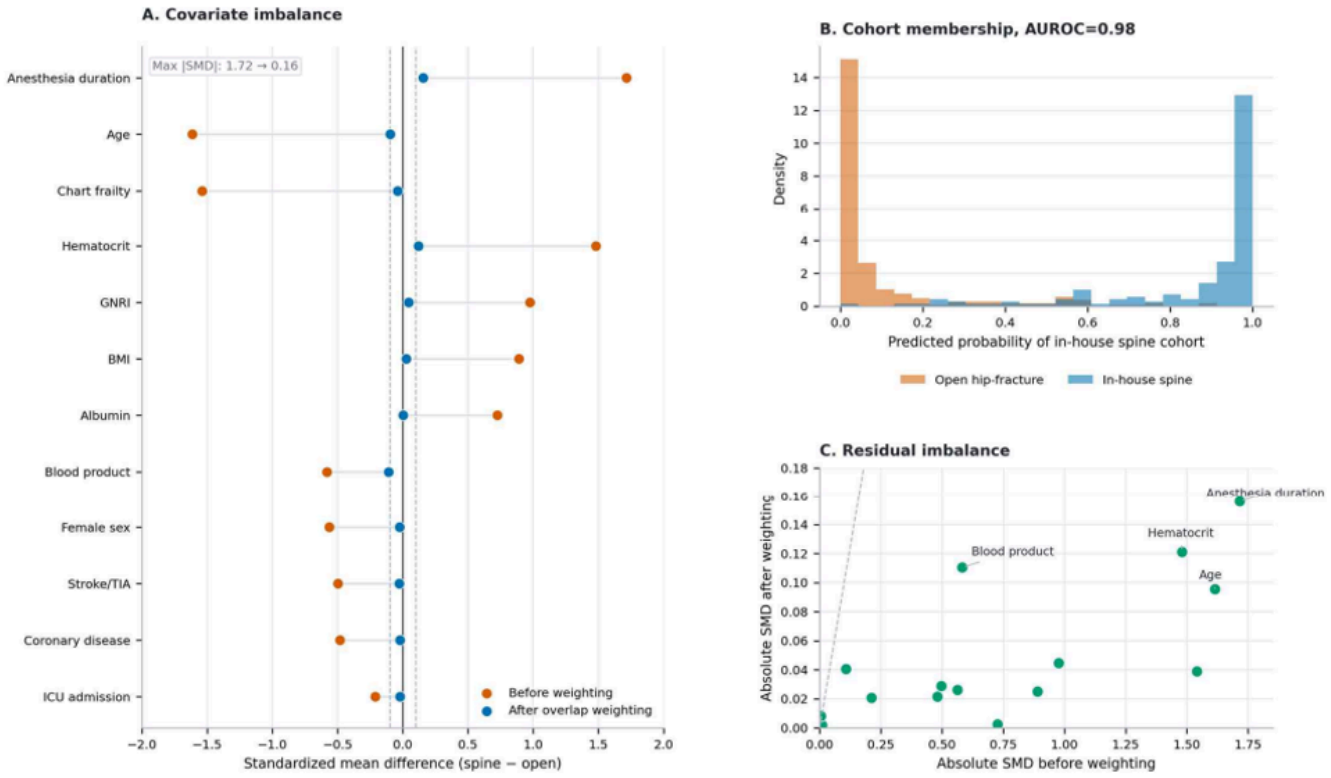


Figure 2

Dataset-shift fingerprint before cross-surgery benchmarking. Baseline imbalance between the external hip-fracture and in-house spine cohorts. A, Standardized mean differences before and after overlap weighting; the maximum absolute SMD decreased from 1.72 before weighting to 0.16 after overlap weighting. B, Cohort-membership model showing near-separation between cohorts (AUROC 0.98). C, Residual imbalance plot showing improved balance after overlap weighting. SMD, standardized mean difference.

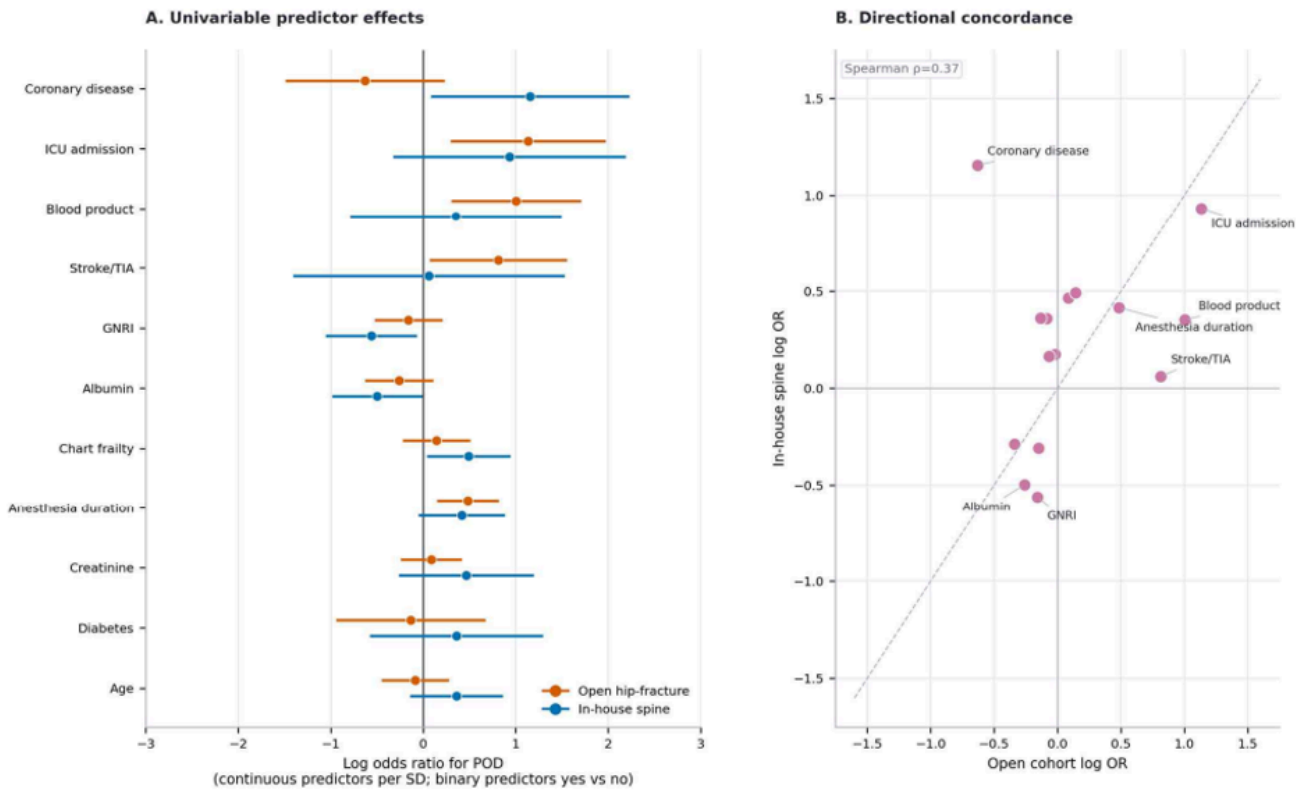


Figure 3

Predictor transportability atlas across surgical contexts. Comparison of common POD predictors across surgical domains. A, Univariable log odds ratios for POD in each cohort. B, Directional concordance between external hip-fracture and in-house spine predictor effects was modest (Spearman ρ=0.37). Shared domains showed partial transportability, but predictor magnitude differed by surgical context.

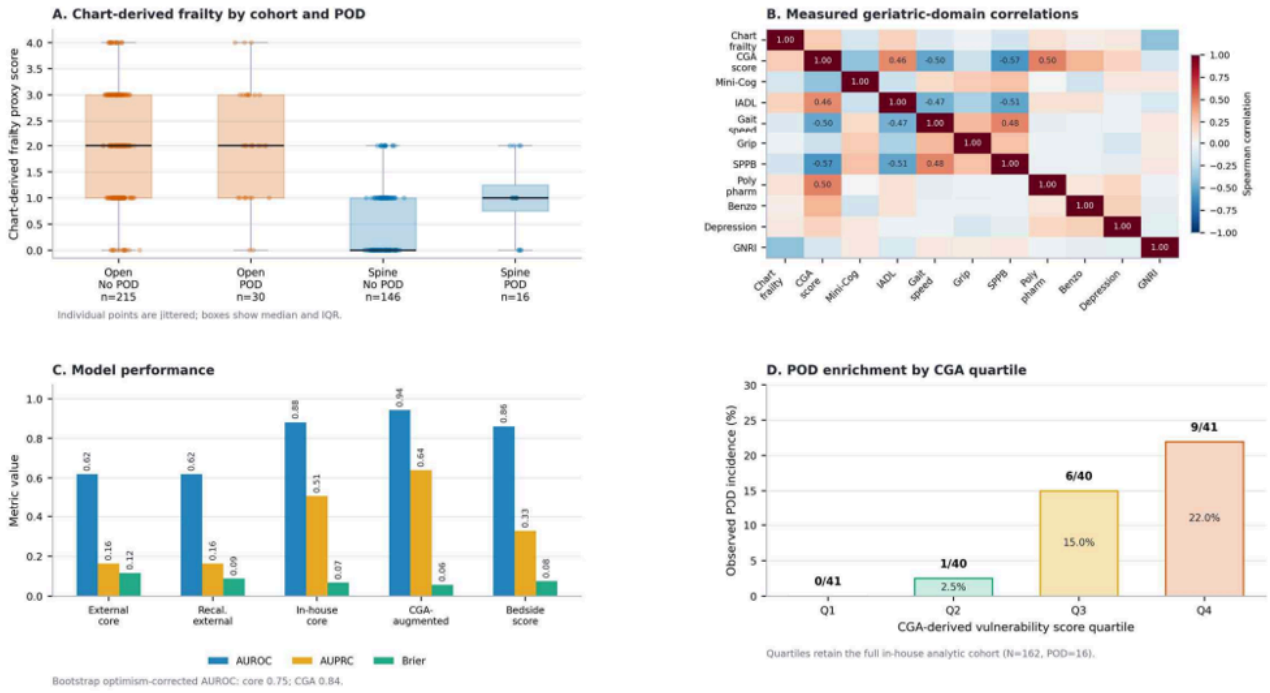


Figure 4

Bridging chart-derived frailty with measured geriatric vulnerability. Relationship between chart-derived vulnerability and measured CGA domains. A, Chart-derived frailty by cohort and POD status, including Open no POD n=215, Open POD n=30, Spine no POD n=146, and Spine POD n=16; individual patient values are jittered and boxes show median and interquartile range. B, Correlation between chart frailty, CGA, cognition, function, physical performance, medication burden, depression, and GNRI. C, Incremental model performance after local CGA updating: AUROC/AUPRC/Brier were 0.62/0.16/0.12 for the transported external core model, 0.62/0.16/0.09 after recalibration, 0.88/0.51/0.07 for the in-house core model, 0.94/0.64/0.06 for the CGA-augmented model, and 0.86/0.33/0.08 for the bedside score. Bootstrap optimism-corrected AUROC was 0.75 for the in-house core model and 0.84 for the CGA-augmented model. D, POD incidence increased across CGA vulnerability quartiles: 0/41, 1/40, 6/40, and 9/41.

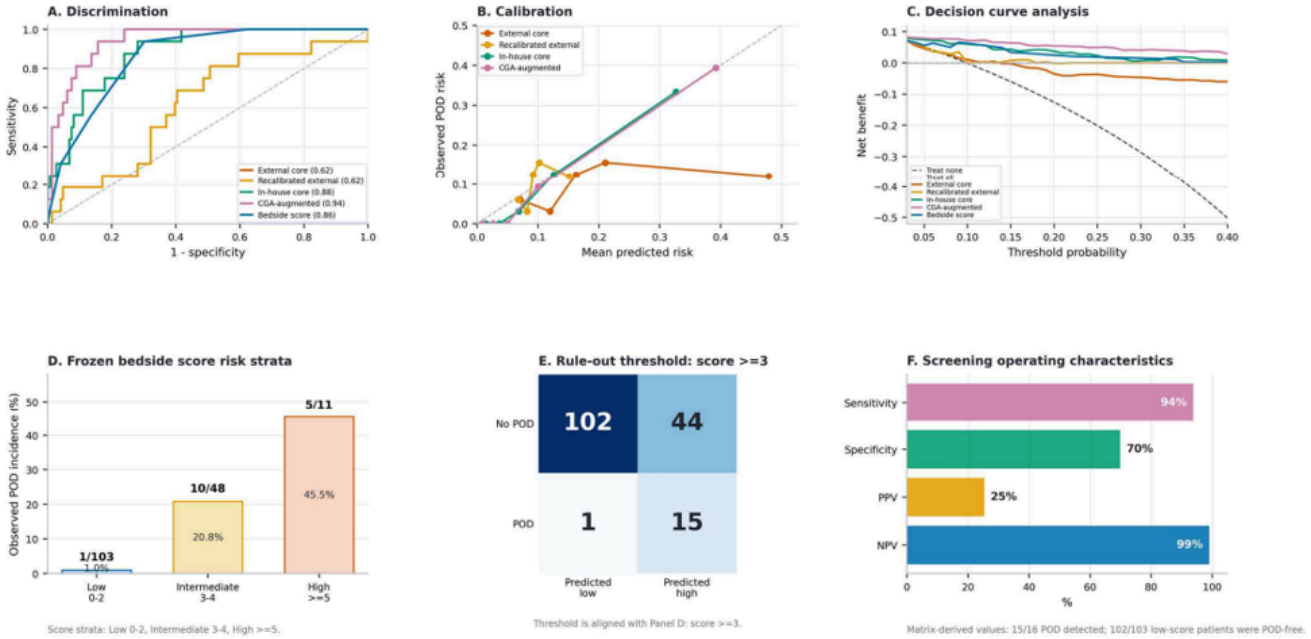


Figure 5

Model updating and clinical utility dashboard. Performance and clinical utility of transported, locally updated, CGA-augmented, and frozen bedside models. A, AUROC improved from external transport to local and CGA-augmented models: external core 0.62, recalibrated external 0.62, in-house core 0.88, CGA-augmented 0.94, and bedside score 0.86. B, Calibration drift and local recalibration. C, Decision-curve analysis for prevention-bundle targeting. D, Frozen bedside score risk strata showed POD rates of 1/103, 10/48, and 5/11 for low (0–2), intermediate (3–4), and high (≥5) score categories. E–F, A score ≥3 rule-out threshold yielded a confusion matrix of TN=102, FP=44, FN=1, and TP=15, corresponding to sensitivity 94%, specificity 70%, PPV 25%, and NPV 99%.

Supplementary Files

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

- Table1.jpg