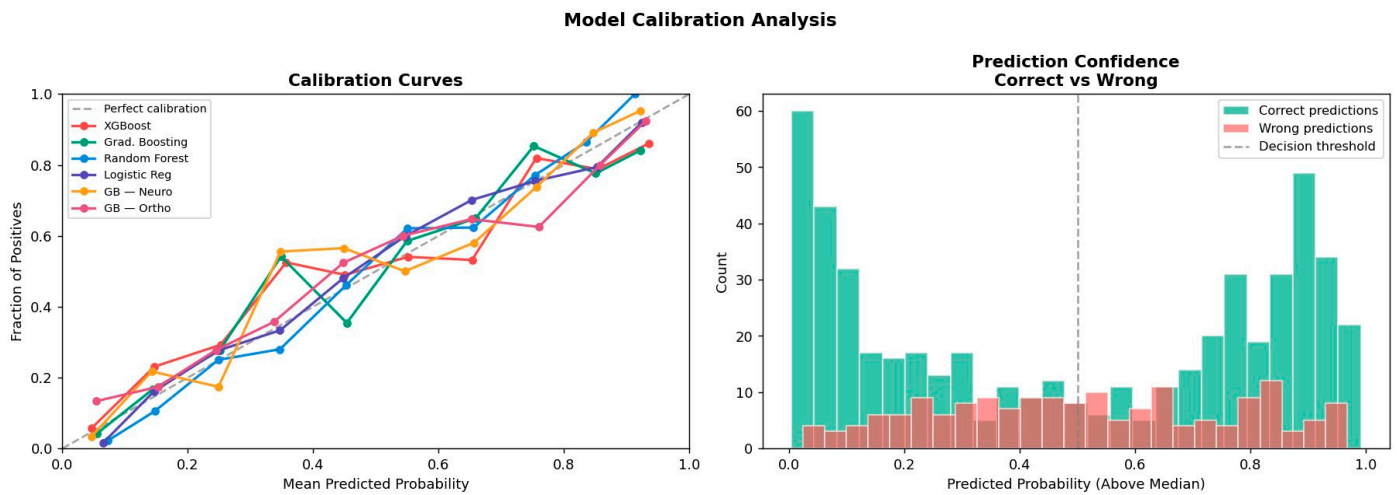


Supplementary Material

This supplementary file contains four additional figures that provide extended detail on model calibration, cluster-specific SHAP analysis, and individual patient explainability. These figures complement the main manuscript without being essential to the primary conclusions.

Supplementary Figure 1 — Model Calibration Analysis

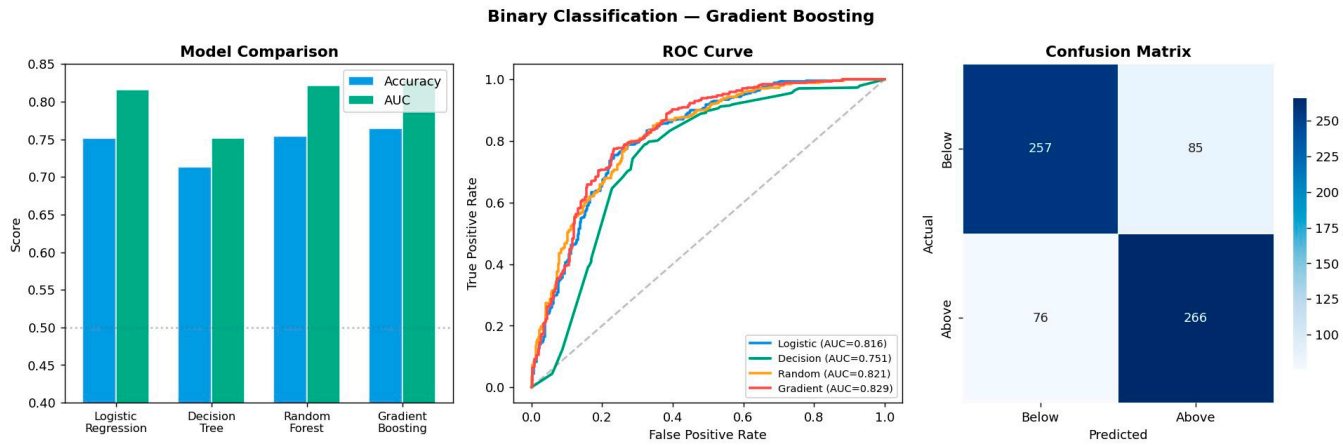
Calibration curves for all six model configurations (XGBoost, Gradient Boosting, Random Forest, Logistic Regression, GB-Neurological, GB-Orthopedic) are shown alongside a reference diagonal representing perfect calibration. All models demonstrated acceptable calibration in the 0.4–0.8 probability range, with some overconfidence observed at the extremes. The right panel presents the distribution of predicted probabilities for correctly classified vs. misclassified patients, showing that errors concentrate near the 0.50 decision threshold — confirming that the model's uncertainty zone (probability 0.40–0.60) is the primary source of misclassification.



Supplementary Figure S1. Model Calibration Analysis. Left: Calibration curves for all model configurations vs. the perfect calibration diagonal. Models follow the diagonal reasonably well, with slight overconfidence in the 0.3–0.5 range. Right: Distribution of predicted probabilities for correct (green) vs. incorrect (red) predictions — errors concentrate near the 0.50 decision threshold, while high-confidence predictions (probability < 0.20 or > 0.80) are predominantly correct.

Supplementary Figure 2 — Baseline Model Comparison (All Patients)

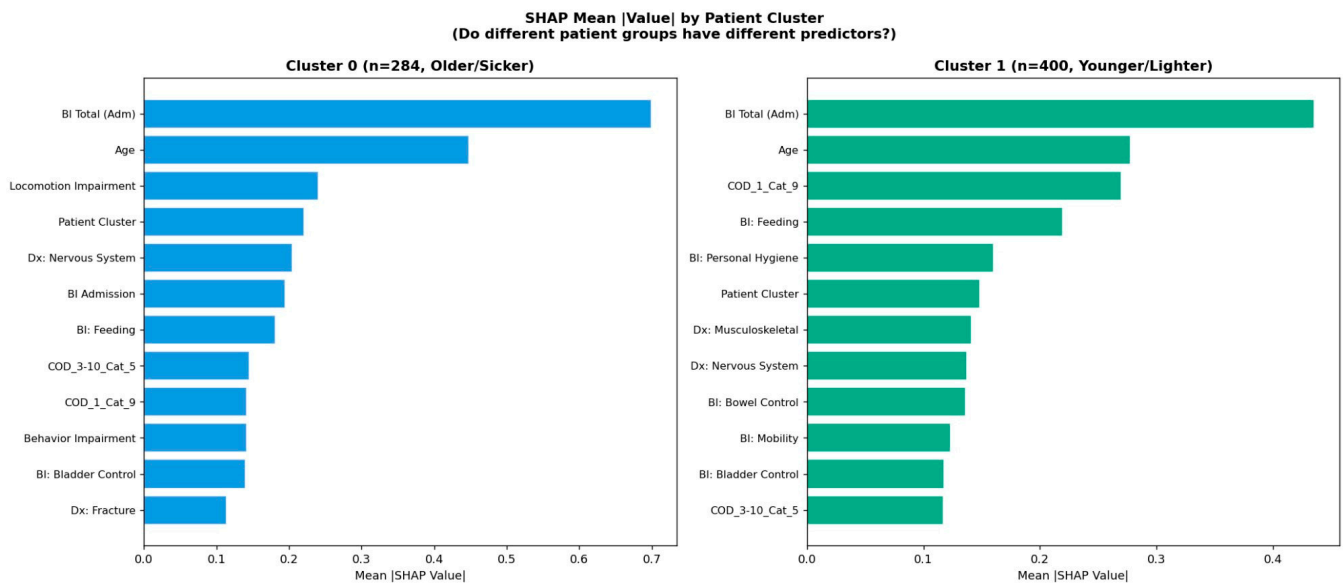
This figure provides a focused view of model performance on the All Patients cohort prior to the comprehensive group-stratified comparison presented in Figure 4 of the main manuscript. The left panel compares Test Accuracy and AUC-ROC across all nine models, illustrating the performance plateau — where tree-based ensemble models (XGBoost, LightGBM, Gradient Boosting, Voting, Stacking) converge within a narrow AUC band (0.816–0.833). The confusion matrix confirms near-balanced false positive and false negative classification errors.



Supplementary Figure S2. Baseline Model Comparison — All Patients Cohort. Left: Test Accuracy and AUC-ROC for all nine models, demonstrating the performance plateau among tree-based ensembles (AUC range: 0.816–0.833). Center: ROC curves showing close clustering of all models except Decision Tree. Right: Confusion matrix for Gradient Boosting (best test accuracy at 76.5%) showing near-balanced FP ($n=85$) and FN ($n=86$) classification errors.

Supplementary Figure 3 — SHAP Feature Importance by Patient Cluster

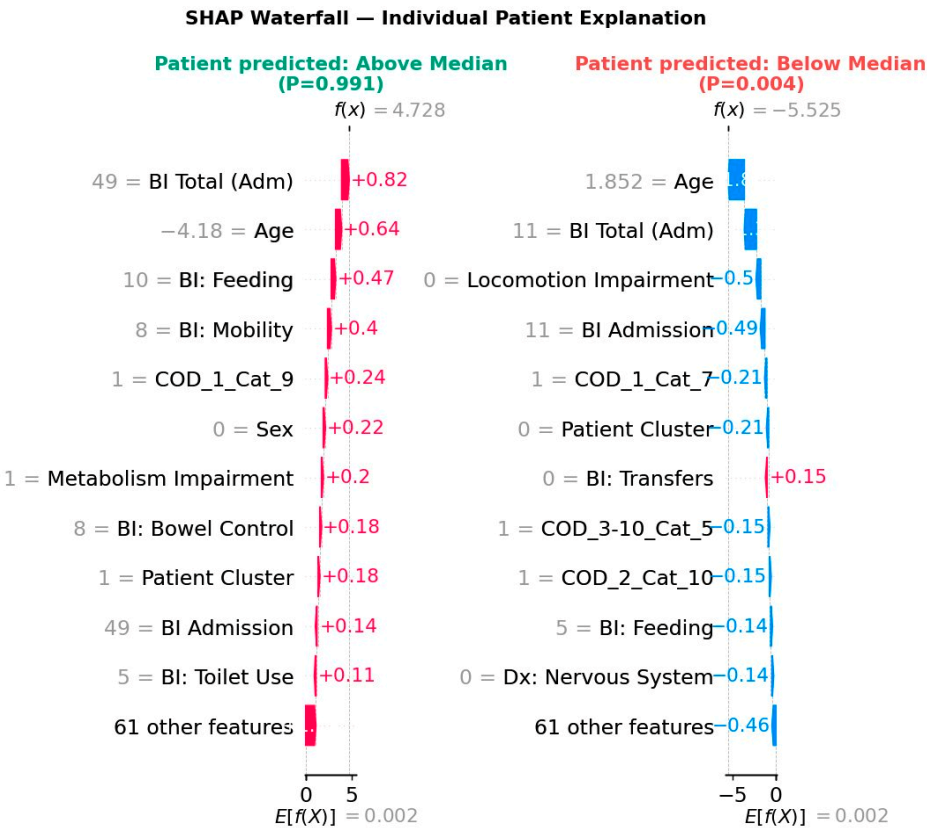
This figure extends the global SHAP analysis (Figure 6, main manuscript) by stratifying feature importance within each patient phenotype identified by K-Means clustering. In Cluster 0 (High-Burden, n=1,135), BI Total at admission dominated with mean |SHAP| = 0.700 — nearly twice the second-ranked predictor — indicating that initial functional status is the primary determinant of recovery in severely impaired older patients. In Cluster 1 (Moderate-Burden, n=1,600), predictive importance was distributed more broadly across features, suggesting that recovery in less severely impaired patients is governed by a more complex interaction of functional, diagnostic, and behavioral factors.



Supplementary Figure S3. SHAP Mean |Value| by Patient Cluster. *Left: Cluster 0 (High-Burden, n=1,135 full dataset; n=284 test patients) — BI Total at admission dominates at mean |SHAP| = 0.70, indicating that functional entry status is the primary recovery determinant in this phenotype. Right: Cluster 1 (Moderate-Burden, n=1,600 full dataset; n=400 test patients) — importance distributed across multiple features (top |SHAP| = 0.44), reflecting greater clinical heterogeneity. Patient Cluster feature appears in both panels, confirming consistent self-assignment.*

Supplementary Figure 4 — Individual Patient Explainability (Waterfall Plots)

Waterfall plots illustrate how SHAP values combine to produce predictions for two representative test patients at opposite ends of the predicted probability spectrum. These examples demonstrate the clinical coherence of individual-level explanations: the model's reasoning aligns with established clinical knowledge about rehabilitation prognosis. For the high-improvement patient, functional independence at admission, younger age, and Cluster 1 membership collectively drove the prediction well above the model's baseline expectation. For the low-improvement patient, compounding risk factors (older age, low admission BI, locomotion impairment, and Cluster 0 membership) produced a strongly negative prediction that proved accurate.



Supplementary Figure S4. SHAP Waterfall Plots — Individual Patient Explanations. Left: High-improvement patient (predicted probability $P=0.991$; actual outcome: Above Median) showing positive SHAP contributions from high BI Total at admission (+0.82), younger age (+0.64), independent feeding (+0.47), preserved mobility (+0.40), and Cluster 1 membership (+0.18). Right: Low-improvement patient ($P=0.004$; actual outcome: Below Median) showing negative contributions from older age (-0.80), low BI Total (-0.49), locomotion impairment (-0.50), and Cluster 0 membership (-0.21). The model's expected value $E[f(X)] = 0.002$ reflects the balanced class distribution.