Supplementary Information

Preoperative assessment of 180-day outcomes after first tibial transverse transport for severe diabetic foot ulcers

This Supplementary Information contains Supplementary Methods, Tables 1–28, Figures 1–9, and Supplementary References. Tables 27–28 and Figures 8–9 provide detailed internal and external performance comparisons supporting the main-text displays.

Contents

Supplementary Methods....3
Supplementary Table 1 Candidate information and cohort characteristics....11
Supplementary Table 2 Death-related conditions and event timing....13
Supplementary Table 3 IBG definition and threshold provenance....14
Supplementary Table 4 Models, scoring procedures, and development choices....16
Supplementary Table 5 Repeated performance, calibration and stability....16
Supplementary Table 6 Clinical grading benchmarks....18
Supplementary Table 7 Incremental prediction beyond the clinical reference....20
Supplementary Table 8 IBG versus graded CRP comparisons....21
Supplementary Table 9 Calibration, availability and stability....22
Supplementary Table 10 Outcome gradient and category assignment....23
Supplementary Table 11 Component ranking and detection....25
Supplementary Table 12 Clinical foot-temperature contribution....26
Supplementary Table 13 Model structure and alternative predictors....27
Supplementary Table 14 Heart failure, PCT and outcome sensitivities....28
Supplementary Table 15 Same-centre temporal assessment....29
Supplementary Table 16 Imputation pooling sensitivity....30
Supplementary Table 17 Model complexity and variable selection....31
Supplementary Table 18 Additional cardiorenal review scenarios....33
Supplementary Table 19 Locked scoring specifications....35
Supplementary Table 20 External outcomes and input availability....38
Supplementary Table 21 Locked external performance and paired comparisons....38
Supplementary Table 22 Cumulative external calibration....40
Supplementary Table 23 Supplementary baseline-risk updating....41
Supplementary Table 24 External cardiorenal review scenarios....41
Supplementary Table 25 External sensitivity analyses....43
Supplementary Table 26 Locked CRP and updating parameters....43
Supplementary Table 27 Internal predictive performance comparisons....45
Supplementary Table 28 Matched locked external model performance....46
Supplementary Fig. 1 Probability alignment and variability....46
Supplementary Fig. 2 Foot-temperature removal....47
Supplementary Fig. 3 Temporal assessment....48
Supplementary Fig. 4 Cardiorenal review scenarios....49
Supplementary Fig. 5 Class-specific calibration....50
Supplementary Fig. 6 External cohorts and input availability....51
Supplementary Fig. 7 External probability error and review scenarios....52
Supplementary Fig. 8 Internal discrimination and calibration....53
Supplementary Fig. 9 Clinical grading and predictor contributions....54
Supplementary References....55

Supplementary Methods

Patient outcomes and recorded clinical background

Development cohort ascertainment. The Medical Records Department of The First Affiliated Hospital of Guangxi Medical University supplied original records for 965 departmental patients treated with tibial transverse transport (TTT) or tibial periosteal distraction (TPD) during the study period. Excluding 540 patients for absence of diabetes, receipt of TPD, or Wagner grade 1–2 diabetic foot ulcers (DFUs) left 425 eligible first-TTT patients with Wagner grade 3–5 DFUs. Two patients with indeterminate 180-day outcomes were excluded, leaving 423 patients: 319 with healed wounds, 61 with persistently unhealed wounds, and 43 with severe outcomes. Each patient contributed their first TTT procedure.

Prediction time and measurement selection. Prediction was anchored after the decision to undertake first TTT, when the latest preoperative information was available but before surgery. For each laboratory test, examination, or clinical assessment, we selected the latest completed result available before first TTT, using the original measurement date. Demographics and medical history were extracted from preoperative records. The foot-temperature variable used the latest preoperative palpation assessment. Two ward physicians agreed on the final four-level grade, which was retained in the database. The target was prognosis among patients who had been selected for and received first TTT within the observed multimodal care pathway. The intended application is additional preoperative cardiorenal review and perioperative planning, alongside routine care and clinically indicated consultations.[1]

Treatment context. TTT formed part of multimodal limb-salvage care, including infection control, debridement and wound care, offloading, and metabolic management. Revascularisation was guided by vascular assessment and clinical indications. Treatment was individualised by the clinical team. Preoperative records documented anti-infective treatment in 310/423 patients (73.3%), debridement in 179/423 (42.3%), and revascularisation in 70/423 (16.5%) (Supplementary Table 1). Infection–inflammation measurements represented the latest available preoperative state within this pathway.[2–4]

Outcome window and interpretation. Complete epithelialisation of all index wounds had to occur by day 180 after the first TTT procedure and be maintained for at least 1 week. For wounds epithelialised near the end of the 180-day window, the nearest subsequent follow-up record could confirm persistence for at least 1 week. This confirmation did not extend the 180-day window for ipsilateral below-knee amputation or all-cause death. Healed patients also had to remain alive and free of ipsilateral below-knee amputation through day 180. Patients who had not met the healing criterion but were alive without ipsilateral below-knee amputation were classified as persistently unhealed. Amputation or death within 180 days determined the severe category, including after wound closure. Patients with both amputation and death were counted once in the severe composite and included in both component-event analyses. Deaths remained in the outcome denominator. New or recurrent ulceration after first confirmed healing did not reclassify patients as persistently unhealed; minor amputation was classified according to index-wound healing. The healing category therefore captures first sustained index-wound closure with 180-day survival and freedom from ipsilateral below-knee amputation. The final verification of the eight previously outstanding records was completed by Shikai Lin on October 1, 2026, using telephone follow-up or outpatient review; their frozen day-180 categories were retained.

A separate sensitivity analysis excluded three healed patients whose healing was confirmed by telephone alone. Saved out-of-fold (OOF) predictions from the earlier five-fold evaluation were retained for the remaining 420 patients; models were not refitted (Supplementary Table 14).

The supplementary workbook was linked to the original 425 candidate records using a unique key constructed from 11 stable source fields. After harmonising numeric strings, all records matched uniquely. The original two exclusions, 423-patient analysis cohort, outcomes (319/61/43), and five-fold assignments were retained. Site, ischaemia, neuropathy, bacterial infection, area, and depth (SINBAD) components and totals and University of Texas (UT) grade and stage were complete for all 423 patients. Component totals and the infection/ischaemia definitions of UT stage were checked. All patients had Wagner grade 3–5 disease and recorded UT grade 3, denoting wounds extending to bone or joint. The SINBAD neuropathy domain was retrospectively coded using documented clinical diagnoses of type 2 diabetic peripheral neuropathy. Following review of the original clinical records, infection coding was corrected in 122 included patients: UT stage A to B in 90 and C to D in 32, with a one-point increase in each corresponding SINBAD total. The other five SINBAD components, cohort membership, outcomes, and model predictors were unchanged. SINBAD totals were 3–6; UT stages were B in 315 and D in 108 patients. The corrected totals and stages were used to refit both benchmark evaluations.

Source and timing. Clinical notes from the original case-summary workbook's special-reasons field (column AP) were linked to the 423 analytic records using the verified patient-to-source-row alignment. Within 180 days, 28 patients died and 19 underwent ipsilateral below-knee amputation; four had both events, giving 28 + 19 − 4 = 43 patients with the severe composite. Exact dates were available for all 28 deaths and 19 amputations. The four patients with both events underwent amputation before death, with intervals of 1, 1, 13 and 16 days. Two additional deaths occurred at days 227 and 255, and one additional amputation occurred at day 217; these events were outside the observation window. The patient who died at day 227 had undergone amputation at day 59 and contributed an amputation event within 180 days. All three patients remained in the analytic cohort under the outcome definitions. Event intervals were measured from the first TTT date.

Clinical cause determination and review. Most deaths occurred in an ICU. The principal clinical reasons for death were determined by ICU physicians at the time of death and jointly reviewed by diabetic-foot specialists and ICU physicians using original records supplied by the hospital medical-records department. Separately, non-exclusive tags from the case-summary special-reasons field retained documented heart failure/cardiorenal failure, renal injury/failure, severe infection/necrotising fasciitis and other conditions (Supplementary Table 2). These tags capture overlapping coexisting conditions and terminal descriptions, separately from the determination and review of principal clinical causes. Three case-summary entries contained particularly limited detail: cardiopulmonary arrest with a heel wound, an unspecified reason, and necrotising fasciitis with no further cause detail in that field. Their wording was preserved. Malignancy was retained as accompanying history or clinical information; cardiopulmonary arrest was not recoded as heart failure. These descriptive tags were not added as predictors.

Predictors and score development

The six specified predictors were age, heart failure grade, palpated foot-temperature grade, estimated glomerular filtration rate (eGFR) grade, wound-area grade, and infection–inflammation burden grade (IBG). Clinical categories and source fields are described in the main Methods and the candidate-input dictionary in Supplementary Data. IBG construction and threshold provenance are reported in Supplementary Table 3. Random-forest ranking, one-versus-rest L1-penalised (OVR-L1) screening, predictive performance, and clinical availability informed the choice of predictors (Supplementary Table 17). One full-cohort imputation (M=1) yielded the locked 0–59-point card. Repeated internal evaluation re-estimated points and probability maps within training partitions (Supplementary Table 4).[5]

Clinical input definitions. Heart failure was coded from the preoperative cardiology diagnosis together with the preoperative echocardiogram: 0 none, 1 chronic, and 2 acute. Diagnosis combined clinical manifestations with objective cardiac abnormalities (main Reference 18). Categories reflected the documented preoperative clinical state. Palpated foot temperature was graded as cold, cool, warm or warmer (1–4). Renal function was classified using laboratory-reported eGFR standardised to 1.73 m²; grades 1–5 were ≥90, 60≤eGFR<90, 30≤eGFR<60, 15≤eGFR<30, and <15 mL/min/1.73 m² or documented uraemia. Wound area was estimated as maximum length × maximum width, summed across multiple wounds, and categorised as A<5, 5≤A≤20, 20<A≤100 or A>100 cm². Recorded categories were retained for the locked analyses. Supplementary Table 19 summarises extraction and coding.

Ordinal scoring. The model was $\text{logit}[P(Y\leq k|X)]=\alpha_k-\beta^T X$, k=0,1. Relative points were constrained to be non-negative, with the lowest-risk category as reference. Coefficients were scaled by the smallest positive coefficient, normalised to a 60-point target, rounded, and subjected to pool-adjacent-violators (PAV) monotonicity constraints. Monotonic directions for the final six predictors followed the predefined clinical coding: risk points were constrained not to decrease with higher age, heart-failure, eGFR, wound-area, or IBG categories, and not to increase with warmer palpated foot-temperature grades. The theoretical maximum was 59 points. The fixed point table was derived from one full-cohort imputation (M=1); cross-validated predictions were pooled across M=20 imputations.

M=20 fixed-score imputation sensitivity. This analysis held fixed the six predictors, all coding rules, monotonic directions, the 60-point target, and the 24/34 risk boundaries, without repeating predictor selection or cut-off optimisation. Twenty full-development-cohort imputations were generated, and the same proportional-odds model was fitted in each. Continuous coefficients were combined using Rubin’s rules before applying the same smallest-positive-coefficient scaling, 60-point normalisation, rounding, and PAV monotonicity procedure as the formal score. Patient scores were calculated within each imputation and then averaged. Both the M=1 and M=20 point tables were evaluated using the prespecified five-fold split, with score-to-outcome probability mappings fitted only in

training folds. Patient-level paired bootstrap resampling (B=500, seed=42) provided 95% CIs for performance differences. Because both point tables were derived from the complete development cohort, this analysis quantified imputation-related stability; the repeated internal evaluation is reported in Supplementary Tables 4, 5, and 6.

Repeated internal evaluation and clinical benchmarks

Models and procedures evaluated. The 0–59-point card and frozen 24/34 boundaries were derived from the whole development cohort. The earlier post-selection five-fold OOF evaluation re-estimated points after fixing predictor identities and upstream specifications. Matched repeated analyses evaluated scoring procedures whose point tables were re-estimated within each training partition. The specified automated workflow (AUTO) selected configurations and predictors; the two six-predictor procedures (SF and SL) retained age, heart failure, palpated foot temperature, eGFR, wound area, and IBG. The six-predictor procedure with existing IBG thresholds and training-only alignment (SF_align) provided the subsequent matched repeated estimates and summaries in Supplementary Table 10 and Supplementary Data (source tables 48 and 51). These estimates describe the fold-refitted procedure. External validation applied the whole-cohort card and locked complete-input maps in Supplementary Table 19 unchanged (Supplementary Tables 20–26). The primary historical post-selection OOF mAUC was 0.8047 (95% CI 0.7462–0.8497); the earlier repeated nested six-predictor-budget assessment yielded 0.7813 (SD 0.0151). The later SF_align estimate of 0.8052 is a separate procedure assessment.

Specified automated development. Ten repetitions of outer stratified five-fold evaluation use seed 20261002 + r. Within outer fold f, the five inner folds use seed 20262002 + 100r + f. Inner imputation seeds are 20271002 + 1000r + 100f + i; outer-refit seeds are 20281002 + 100r + f (r, f and i start at 0). M = 20 training imputations include the training outcome; evaluation transformations use a separate outcome-blind imputer. The fixed EfficientMI implementation uses posterior IterativeImputer with BayesianRidge, 10 iterations and skip_complete=True. Variables with more than 40% missingness in a training partition are excluded. Original IBG is excluded from imputation estimation and reconstructed from imputed constituents. For columns fully observed in training, missing evaluation values are initialised using the training mean. All scoring and ranking operations use one CPU thread. [6,7]

Training-local IBG. In each training partition, midpoints between observed NLR values define candidate thresholds. Low and high thresholds maximise the mean Youden indices for Y≥1 and Y=2 subject to low<high. Candidate ESR thresholds exceed 70 and are selected to maximise the Youden index for Y=2. At least 10 observed patients are required on each side of a candidate. The study-adopted WBC, CRP, PCT and ESR 70 rules and the A=max(WBC,NLR), B=max(ESR,CRP), C=PCT domain aggregation are retained. This deterministic rule was specified before the automated workflow was run. SF retains NLR 4.52/7.5 and ESR 94 as earlier cohort-derived specification choices; SL learns the internal thresholds within each training partition.

Configuration selection. Candidate K values are 5, 6, 7, 8 and 11; age is ≥65 or automatic training tertiles; score targets are 60 or 100; natriuretic peptide precursor is eligible or excluded. Random-forest ranking with 100 trees and OVR-L1 screening with C=0.05 are computed within each imputation and aggregated by Borda ranks. Inner OOF score mAUC identifies configurations within 0.01 of the highest value. The smallest eligible K is retained; the lowest leave-inner-fold-out multiclass Brier score determines the configuration. Ties within 1e-8 prefer 60 points, age 65, natriuretic peptide precursor excluded, then lexical ID. Clinical monotone directions and source ordinal coding are retained. Each point model uses the same proportional-odds likelihood, reference-category selection, coefficient scaling, integer rounding and PAV functions.

Matched six-predictor procedures. SF and SL fix the six predictor identities, the age boundary at 65 years and 60-point normalisation. Each inner and outer training partition re-estimates coefficients and integer points in 20 imputations. Each procedure compares the original inner-OOF sigmoid mapping with training-only score alignment. For imputation m and scoring model h, z=(S−mean(S_train,h,m))/SD(S_train,h,m), with an SD floor of 1e-7. Each inner held-out score uses its generating model's training mean and SD. Cumulative sigmoid maps are fitted to the assembled aligned inner OOF values; outer evaluation scores use the outer model's own training mean and SD. Centring and scaling are estimated entirely on the training side. The prespecified primary contrast of this matched experiment was the multiclass Brier difference for SL aligned minus SL original. SF retains the clinical integer card's IBG definition; SL evaluates training-local threshold development. Mapping diagnostics are shown in Supplementary Fig. 1.

Probabilities and groups. For each imputation, separate sigmoids estimate q1=P(Y≥1) and q2=P(Y=2), giving p=(1−q1,q1−q2,q2). When cumulative probabilities cross, negative intermediate probabilities are clipped, class probabilities are normalised, and predictions are averaged over M=20 imputations.

Training QWK boundaries use patient-pooled M=20 scores, with ties resolved using macroF1. Original and aligned mappings share scores and QWK groups. Current repeated risk strata use each outer training partition's boundaries (Table 10); external performance of the unchanged frozen 24/34 boundaries is reported separately in Supplementary Data.

Clinical benchmarks. The audited 423-patient alignment is retained, with complete retrospective SINBAD and UT coding. Discrimination based on raw SINBAD scores (0–6) is reported separately from discrimination based on training-fold ordinal-logit probabilities. The SINBAD probability map supplies Brier, calibration and maximum-probability QWK. UT uses stage B as reference with an indicator for D. Grade is uniformly 3 and is omitted; all patients are infection-positive, so this stage contrast reflects ischaemia within infected wounds extending to bone or joint. Wagner uses numerical grade in a training-fold proportional-odds map. Benchmarks use the same 50 outer partitions as the scoring procedures. AUCs are probability-based except explicitly labelled raw SINBAD rows. Generalised C uses expected severity p1+2p2 for all probability models. UT grade 3 denotes wound extension to bone or joint; stage separately records infection and ischaemia, with A indicating neither, B infection alone, C ischaemia alone, and D both. SINBAD neuropathy was coded using documented clinical diabetic peripheral neuropathy diagnoses as a retrospective proxy for loss of protective sensation. The depth, area and infection components were each 1 in all 423 patients; the varying components were site, diagnosis-based neuropathy and ischaemia.

Analysis timing and integrity. Training-only alignment was defined after the first same-cohort automated-workflow evaluation, and the matched follow-up protocol was locked before new predictions. Original and aligned AUTO predictions were reused from the recorded calibration study after verification of patient indices, outcomes and input hashes. Imputation, points, six-predictor predictions and benchmark maps were fitted anew. All 50 outer-fold prediction records were completed before performance was summarised. The reproducibility package contains patient-level predictions, partitions, point tables and source checksums. The descriptive summaries in Supplementary Table 10 and Supplementary Data (source tables 48, 51) use these saved predictions. For the source-record correction, SINBAD and UT maps were refitted using the same 50 outer training partitions and the earlier seed-42 five-fold partitions. All other model predictions were retained. The repeated comparisons reused the original 2,000 patient-cluster weights; the earlier comparisons reused their original 2,000 patient resamples. Updated full-cohort SINBAD and UT maps were separately locked for future independent evaluation. The correction protocol was documented before these models were refitted.

Uncertainty and stability. Each repetition supplied one complete set of 423 OOF predictions for cumulative mAUC, both cumulative AUCs, Brier scores, generalised concordance, and quadratic weighted kappa (QWK). Estimates were arithmetic means of the ten repetition-specific metrics; sample SDs and ranges described variation across training partitions. Paired patient-cluster bootstrap resampling used 2,000 draws (seed 20261013), retaining each patient's ten predictions together. Each resample recomputed and averaged the repetition-specific metrics. Intervals were conditional on saved OOF predictions; scoring models were not refitted. Calibration intercepts and slopes were jointly re-estimated with resampled patient weights for each cumulative endpoint and averaged across repetitions. Observed-to-expected event ratios (O/E) were summarised by repetition. Variability was described using patient-level severe-risk SD and median absolute deviation (MAD), changes in stratum assignment, and fold-level mean-risk error.

The seven-arm follow-up comparison of calibration transfer and training-selected K=5 is retained in Supplementary Data (source Tables 41–42) with its original protocol. It evaluates distinct prediction procedures and is reported separately from fixed-six matched temperature removal or the current clinical-reference comparison.

Four-predictor clinical reference and CRP comparisons

The prespecified four-predictor clinical reference combined age, heart failure grade, eGFR grade, and wound-area grade. It provided a reproducible comparator based on recorded clinical information. Clinician judgement was not measured. The comparison protocol was saved before new comparator predictions were generated. Each procedure used the same 423 patients and ten repetitions of five outer folds.

Matched validation. The comparison retained the same 423 patients and ten repetitions with five outer and five inner folds. For the four-predictor reference, each training partition fits M=20 BayesianRidge posterior imputations using only its four inputs; observed values are preserved and missing heart failure grades are constrained to training categories. Age is dichotomised at 65, and ordinal integer scoring, reference-category selection, 60-point target scaling, training standardisation and inner-OOF cumulative mapping match the previous procedure. No evaluation outcome enters imputation, scoring or mapping. All 50 new

outer prediction records were sealed before evaluation. Across the 300 training units, 9,575 ordinal fits converged, including conditional reference-category refits; their content digests and disjoint partitions were independently checked.

Comparisons and uncertainty. The primary incremental comparison is the severe-outcome binary Brier difference between required-input six-predictor IBG and the required-input four-predictor reference, calculated as the former minus the latter. Severe AUC, cumulative mAUC and multiclass Brier are secondary. The previously reported common-auxiliary six-predictor IBG and the required-input six-predictor CRP are additional comparisons; common-auxiliary versus required-input comparisons also change the imputation information scope. Metrics are averaged over ten complete OOF repeats. Two thousand paired patient-cluster resamples (seed 20261009) retain all ten predictions per patient. Intervals are conditional on saved predictions and do not repeat original predictor selection. Secondary and threshold intervals are exploratory and unadjusted.

Clinical question and prediction procedure. This same-cohort follow-up evaluates whether graded CRP can replace the IBG representation while retaining age, heart failure, palpated foot temperature, eGFR and wound area. It was prompted by the earlier continuous-CRP regression point estimates in Supplementary Table 13. The protocol was locked before generating new predictions. Predictor identities, age ≥65 years and existing IBG cut-offs were fixed after development. The analysis evaluates scoring procedures refitted within training data and retains the original whole-cohort 0–59 card and frozen 24/34 boundaries.

Comparisons and input scope. Common-auxiliary comparisons use the same original candidate matrix, posterior imputation draws and seeds for IBG and CRP. Existing aligned fixed-six IBG OOF predictions were reused after verification of patients, labels, folds, hashes, and deterministic imputation. Required-input branches restrict the imputation matrix to the five shared predictors plus the five IBG constituents (ten fields), or the five shared predictors plus raw CRP (six fields). The CRP required-input analysis excludes WBC, NLR, ESR, PCT, and other laboratory auxiliaries. IBG is reconstructed after imputing its constituents. Raw CRP is imputed before grading as 0 for <10, 1 for 10 to <100 and 2 for ≥100 mg/L; these thresholds are inherited from the existing IBG definition.

Matched fitting and mapping. The 423 patients, locked outcomes and outer partitions match the existing ten-repeat stratified five-fold procedure (outer seed 20261002 plus repeat index). Each outer training partition uses five inner folds and M=20 posterior IterativeImputer/BayesianRidge imputations (maximum ten iterations), training-range constraints and valid categorical rounding. Evaluation outcomes do not enter imputation or fitting. The categorical joint proportional-odds model, reference-category selection, coefficient scaling, 60-point normalisation target, conversion to integer points and monotonic constraints are shared across comparisons. Each scoring model is standardised using its own training mean and SD. Cumulative probability maps are fitted to aligned inner OOF scores; crossing probabilities are clipped and renormalised before M=20 pooling. Group boundaries maximise training QWK on patient-pooled scores. All fifty outer prediction records were completed before overall performance was evaluated. All 36,040 new ordinal fits converged without the ridge fallback.

Outcomes and uncertainty. The primary contrast is CRP minus IBG multiclass Brier under common-auxiliary imputation. The required-input contrast assesses sensitivity to implementation requirements; limited-minus-common contrasts assess changes in imputation input scope. Each metric is computed once per complete OOF repeat and averaged across ten repeats. The paired patient-cluster bootstrap used 2,000 resamples (seed 20261023), retaining each patient's ten predictions together. Intervals are conditional on saved OOF predictions, with scoring models fixed and calibration regressions re-estimated during resampling. The study did not specify an equivalence or non-inferiority margin. Repeat SDs describe variation across training partitions; patient-level probability SDs and changes in QWK groups describe prediction stability.

Observed input completeness was assessed before imputation. All required inputs were available in 301/423 patients for the IBG score, 396/423 for the CRP score, and 405/423 for the four-predictor reference. Common-auxiliary and required-input imputation results are reported separately (Supplementary Tables 7–9). Nine severe outcomes occurred among the 122 patients with incomplete IBG inputs. Fold-refitted internal performance used within-fold imputation in all 423 patients; the locked calculator returned no risk estimate when required inputs were missing.

Matched sensitivities and outcome summaries

The specifications in Supplementary Table 13 use age ≥65 years and numerical clinical grades in a stacked binomial cumulative-logit likelihood. The shared-slope reference, age-specific model, all cut-point-specific model, ejection fraction (EF) substitutions, and IBG removal or substitutions share the original five folds and M=20 imputation. For crossing cumulative probabilities, the two probabilities were averaged before probability constraints and normalisation. All 100

fits for each specification converged. Performance differences are alternative minus the shared-slope reference, with 500 valid patient-level paired bootstrap resamples. The calibration slope reported in Supplementary Table 13 is the mean of three class-specific OVR slopes. This framework has its own six-predictor reference; the categorical joint-ordinal likelihood used in the foot-temperature analysis is specified separately below.

The original 48-feature imputation pipeline was retained: 20 posterior imputations, BayesianRidge IterativeImputer, seeds 42–61, and at most ten iterations. Training-side imputation could use training outcomes; evaluation-side imputation received no evaluation outcomes. IBG was reconstructed from imputed components. Age was coded as ≥65 years and other grades as fixed categorical indicators. Both the five-predictor model and its matched six-predictor reference used the same joint ordinal proportional-odds likelihood. For each patient, the predicted probabilities were averaged across the 20 imputations. Foot temperature was removed from the five-predictor prediction equation and retained as an auxiliary imputation variable. All 100 fits in each model converged. Differences are five predictors minus the matched six predictors; comparisons with the earlier five-fold point-score reference are also tabulated.

The matched foot-temperature analysis and earlier single five-fold benchmark analyses used 2,000 non-stratified patient-level paired bootstrap resamples (seed 42), with identical sampled patients for every model and percentile 95% CIs. Saved OOF predictions were held fixed. These intervals quantify patient-sampling uncertainty conditional on fitted OOF predictions. The full original development process was not repeated. No multiplicity adjustment was applied. Joint intercept-and-slope logistic recalibration was fitted separately for Y≥1 and Y=2; ideal values are 0 and 1. These intercepts differ from calibration-in-the-large estimates obtained with the slope fixed to 1.

Repeated risk gradient and class assignment. SF_align provides ten sets of 423 OOF predictions. Within each repeat, low, intermediate and high strata use the corresponding outer training-fold QWK boundaries. Within each repetition and stratum, patient counts and the three observed outcome proportions are computed first; any adverse outcome combines persistent non-healing and severe outcomes. Supplementary Table 10 reports arithmetic means and sample SDs of the ten stratum summaries, and Supplementary Data (source table 48) reports all repeat-specific counts and event rates. Main Fig. 2A uses the mean of the per-repeat proportions. For Supplementary Table 10 and main Fig. 2B, each observed-outcome row is normalised by its 319, 61 or 43 patients before averaging the ten repeat-specific assignment proportions. Diagonal entries give class-specific sensitivity. Supplementary Table 10 and main Fig. 2C count the distinct strata assigned to each patient across the ten repeats. Repeated predictions remain clustered within the same 423 patients.

Calibration illustration. Supplementary Fig. 8B–D uses the first recorded repeat of SF_align, with one OOF prediction per patient. Predictions are grouped into up to eight quantiles separately for Y≥1 and Y=2, allowing tied predictions to reduce the number of groups. Supplementary Data (source table 51) reports the group size, event count, mean predicted probability, observed event proportion and Wilson 95% interval for the observed proportion. The continuation of the source table provides counts and percentages for probability histograms with a bin width of 0.05. These first-repeat plot coordinates illustrate calibration; Supplementary Table 5 supplies the calibration estimates averaged across all ten repeats.

In the same-centre temporal sensitivity analysis, imputers, points, probability maps, and risk boundaries were fitted using data from 302 patients treated in 2020–2024 and evaluated in 121 patients treated in 2025–2026. Predictor identities and upstream rules originated from the overall development process. Results are reported in Table 15 and Fig. 3. Earlier single five-fold component ranking, results from setting temperature points to zero, and full fixed-K reselection outputs are retained in Supplementary Data with their own estimands.

Exploratory ranking of component events

Exploratory component ranking. Existing saved OOF composite severe-outcome probabilities, P(Y=2), ranked patients with all-cause death within 180 days (28 versus 395 without death) and ipsilateral below-knee amputation within 180 days (19 versus 404 without amputation). The four patients with both events were positive in both analyses. For the death contrast, the 15 patients with amputation alone were included among controls; for the amputation contrast, the 24 patients with death alone were included among controls. Composite probabilities served as ranking scores. No separate component model was fitted, and no component-specific Brier score or decision curve was calculated.

Procedures and uncertainty. The matched comparison used required-input six-predictor IBG versus required-input four-predictor reference. Required-input six-predictor CRP and common-auxiliary six-predictor IBG (SF_align) were additional procedures; the latter also changes the information available for imputation.

Metrics were calculated for each of ten complete OOF repeats and averaged. Two thousand paired patient-cluster bootstrap resamples (seed 202610091) retained all ten predictions per patient. The percentile intervals are conditional on fixed predictions, exploratory and unadjusted for multiplicity. Differences between the death and amputation incremental AUCs were not formally tested. The component analyses supplement the prespecified severe-composite binary Brier comparison in Supplementary Table 7. Supplementary Table 11 reports the current component analyses using composite severe probabilities from ten repeated evaluations; earlier component analyses are identified in the historical Supplementary Data archive.

Detection scenario. At the previously specified 10% composite severe-risk threshold, identified counts were calculated within each repeat and then averaged. Detection divides the average identified count by the observed component-event count. Supplementary Table 11 reports mean counts and patient-cluster intervals for detection rates. These measures quantify scenario-based event identification.

Additional preoperative cardiorenal review

The intended use is to inform additional cardiorenal review and perioperative planning for patients selected for first TTT, after routine preoperative assessment. A proposed review would integrate cardiac and renal diagnoses, current functional status, and infection and wound burden to define management priorities. Consultations indicated by clinical findings would continue independently of the score. Consultation plans before score display and actual additional consultations were not recorded; threshold-positive counts therefore represent potential review volume.[8]

Severe-risk thresholds of 5%, 10%, 15% and 20% were specified before the new predictions; the complete 5–40% grid uses 2.5% increments. Threshold-positive and identified-event counts are calculated within each repetition and then averaged. Sensitivity divides average identified events by 43, specificity uses 380 non-events, and positive predictive value (PPV) is the ratio of average identified events to average threshold-positive patients. Net benefit is $TP/n - FP/n \times pt/(1-pt)$. At a 10% threshold, the net-benefit calculation weights a false positive by 1/9 of a true positive; this expresses the assumed trade-off for the additional-review scenario. Model and comparator scenarios retain usual care and differ in the threshold-based additional-review strategy. Intervals are exploratory and unadjusted, and no threshold was selected by maximising observed net benefit. Table 18 reports the four prespecified scenarios; Supplementary Data contains the complete grid. Decision scenarios use fold-refitted predictions with within-fold imputation in all 423 patients. Use of the locked complete-input score also depends on the availability of observed inputs (Tables 14D and 19).

A prospective implementation assessment should record routine consultation and perioperative plans before showing the score, and then record new or revised plans, consultation completion and patient burden. Together with 180-day outcomes, function, and quality of life, these measures would test whether model-guided review adds to existing care and changes management or outcomes.

Locked specifications for independent evaluation

Locked complete-input specifications. The six-predictor 0–59-point card was retained unchanged. Two cumulative logistic maps were fitted once to observed scores in 301 complete-input development records (222/45/34 outcomes) and locked. The four-predictor reference card used full-cohort M=1 required-input derivation (seed 42); its maps were fitted in 405 complete records (308/54/43 outcomes). These specifications were applied unchanged in external validation (Supplementary Tables 20–26). Populations used to fit the maps and external input availability are reported by centre. Calculators reject missing or invalid grades and provide no default for unmeasured PCT. The six-predictor interface accepts observed age and five clinical grades; the IBG helper requires all five measured markers. Score and probability checks passed for all 1,920 age-boundary and grade combinations and the hand-calculated example (32 points).

Independent external evaluation applied the locked complete-input point cards and probability maps without development refitting. Calibration, discrimination, input availability, paired four-predictor reference comparisons and the prespecified review scenarios are reported in the main manuscript, main Fig. 3 and Supplementary Tables 20–26. Table 23 reports supplementary baseline-risk updating.

External validation and supplementary baseline risk updating

Centres and ascertainment periods. The external centres were Shanxi Provincial People's Hospital, Zhongshan Hospital, Xiamen University, and The First Affiliated Hospital of Hunan University of Chinese Medicine. Their respective ascertainment periods were January 1, 2020–April 1, 2026; January 14, 2024–

April 1, 2026; and December 8, 2018–April 1, 2026. Each centre verified first-TTT dates and 180-day outcomes. The six external amputations were ipsilateral below-knee amputations within 180 days, without subsequent death in that window. The 38 included Hunan procedures occurred from December 8, 2018, to September 3, 2025. Outcomes were assessed at 180 days.

Locked prediction. The IBG card and complete-input cumulative maps were applied without predictor selection, point refitting, or threshold changes. External data were not statistically imputed. The locked four-predictor reference was compared in identical patients; severe-outcome binary Brier score was the primary paired contrast. The supplementary CRP version was fitted exclusively in development data and locked after the earlier two-centre external results had been examined. The availability pathway used IBG when its inputs were complete and CRP otherwise when its inputs were complete, without reference to outcome or predicted risk. Verified records provided 31 ESR values and the remaining Hunan eGFR value; all 38 Hunan patients had calculable scores. Unmeasured biomarkers were not assigned normal values. Detection limits were used only when they uniquely identified a score category. Documented uraemia was assigned renal grade 5; a matched sensitivity analysis required numerical eGFR.

Inference. Locked predictions were evaluated by centre and pooled, using both cumulative AUCs, their mean (mAUC), binary and multiclass Brier scores, calibration-in-the-large, freely estimated calibration slopes, observed/expected ratios, and prespecified review scenarios. Patient bootstrap resampling (2,000 draws, seed 20261016) preserved centre sample sizes and the pairing of predictions. AUC and slope were not estimated for single-outcome centres. Event-rate, threshold-sensitivity and PPV intervals used exact binomial methods. Calibration intercept and slope intervals used likelihood methods; O/E intervals used a Poisson count interval conditional on the predicted event count. Pooled intervals are conditional on these three centres and their case mix. Centre-restricted concordance used only within-centre event/non-event pairs. Component-event and Wagner analyses were supplementary explorations; external source tables did not contain the fields needed to reconstruct SINBAD or UT.

Updating procedure. Two cumulative intercept offsets were estimated, retaining points and slopes and imposing coherent probabilities over the full score range. IBG, CRP, and the four-predictor reference used identical centre-by-outcome-stratified partitions for 20 repetitions of five-fold patient cross-validation. Each patient supplied one held-out prediction per repetition; metrics were calculated within each repetition and then averaged. Leave-one-centre-out updating fitted offsets in the other centres and applied them to the held-out centre. Paired loss intervals were conditional on saved held-out predictions; 500 bootstrap refits separately assessed optimism in the updating step. Offsets estimated in the full sample define candidate parameters for subsequent independent evaluation. Supplementary Table 23 reports held-out evaluation of the updating procedure. The earlier two-centre offset candidate was also applied unchanged, with overlap between evaluation cohorts identified explicitly.

Literature context and methodological sources

For the submission update, we searched PubMed from database inception through October 17, 2026, without language restrictions, for TTT clinical cohorts, diabetic-foot prognostic studies, systematic reviews and guidelines. Search terms included tibial transverse transport, tibial cortex transverse, transverse tibial transport, diabetic foot, prediction, prognosis, amputation, healing and mortality. Clinical studies and reviews were considered for prognostic context; animal studies served only as mechanistic background. Additional studies of blood-cell ratios, nutritional–immune indicators, and amputation prediction are included here. [9–11]

Risk-of-bias considerations followed PROBAST+AI. The development and external cohorts used the available consecutive cases; no prospective sample-size calculation was performed. Sample-size guidance informs interpretation of event counts and uncertainty rather than determining the historical cohorts. [12–14]

Supplementary Results

Supplementary Table 1 Candidate information and development cohort characteristics by outcome

Variable	Type	Missing (%)	Overall	Healed (n=319)	Persistently unhealed (n=61)	Below-knee amputation or death (n=43)
Age, years	Continuous	0.0%	64.26 ± 10.45	63.79 ± 10.59	63.11 ± 10.27	69.35 ± 8.28
Sex	Binary	0.0%	0=105(24.8%); 1=318(75.2%)	0=75(23.5%); 1=244(76.5%)	0=20(32.8%); 1=41(67.2%)	0=10(23.3%); 1=33(76.7%)
Diabetes duration category	Categorical/ ordinal	1.7%	0=34(8.2%); 1=166(39.9%); 2=192(46.2%); 3=24(5.8%)	0=29(9.3%); 1=132(42.2%); 2=135(43.1%); 3=17(5.4%)	0=3(4.9%); 1=24(39.3%); 2=30(49.2%); 3=4(6.6%)	0=2(4.8%); 1=10(23.8%); 2=27(64.3%); 3=3(7.1%)
Smoking history	Binary	0.0%	0=233(55.1%); 1=190(44.9%)	0=175(54.9%); 1=144(45.1%)	0=35(57.4%); 1=26(42.6%)	0=23(53.5%); 1=20(46.5%)
Wagner grade	Categorical/ ordinal	0.0%	3=93(22.0%); 4=296(70.0%); 5=34(8.0%)	3=81(25.4%); 4=217(68.0%); 5=21(6.6%)	3=6(9.8%); 4=45(73.8%); 5=10(16.4%)	3=6(14.0%); 4=34(79.1%); 5=3(7.0%)
Popliteal artery occlusion	Binary	0.0%	0=403(95.3%); 1=20(4.7%)	0=311(97.5%); 1=8(2.5%)	0=55(90.2%); 1=6(9.8%)	0=37(86.0%); 1=6(14.0%)
Previous revascularisation	Binary	0.0%	0=353(83.5%); 1=70(16.5%)	0=272(85.3%); 1=47(14.7%)	0=46(75.4%); 1=15(24.6%)	0=35(81.4%); 1=8(18.6%)
Hypertension	Binary	0.0%	0=173(40.9%); 1=250(59.1%)	0=136(42.6%); 1=183(57.4%)	0=29(47.5%); 1=32(52.5%)	0=8(18.6%); 1=35(81.4%)
Coronary heart disease	Binary	0.0%	0=308(72.8%); 1=115(27.2%)	0=243(76.2%); 1=76(23.8%)	0=42(68.9%); 1=19(31.1%)	0=23(53.5%); 1=20(46.5%)
Cerebral infarction	Binary	0.0%	0=323(76.4%); 1=100(23.6%)	0=246(77.1%); 1=73(22.9%)	0=52(85.2%); 1=9(14.8%)	0=25(58.1%); 1=18(41.9%)
Lower-limb arteriosclerosis obliterans	Binary	0.0%	0=196(46.3%); 1=227(53.7%)	0=160(50.2%); 1=159(49.8%)	0=24(39.3%); 1=37(60.7%)	0=12(27.9%); 1=31(72.1%)
Diabetic vascular disease	Binary	0.0%	0=23(5.4%); 1=400(94.6%)	0=20(6.3%); 1=299(93.7%)	0=1(1.6%); 1=60(98.4%)	0=2(4.7%); 1=41(95.3%)
Diabetic neuropathy	Binary	0.0%	0=134(31.7%); 1=289(68.3%)	0=108(33.9%); 1=211(66.1%)	0=15(24.6%); 1=46(75.4%)	0=11(25.6%); 1=32(74.4%)
Diabetic nephropathy	Binary	0.0%	0=212(50.1%); 1=211(49.9%)	0=180(56.4%); 1=139(43.6%)	0=22(36.1%); 1=39(63.9%)	0=10(23.3%); 1=33(76.7%)
Diabetic retinopathy	Binary	0.0%	0=369(87.2%); 1=54(12.8%)	0=286(89.7%); 1=33(10.3%)	0=48(78.7%); 1=13(21.3%)	0=35(81.4%); 1=8(18.6%)
Heart failure status	Categorical/ ordinal	4.3%	0=310(76.5%); 1=67(16.5%); 2=28(6.9%)	0=257(83.4%); 1=44(14.3%); 2=7(2.3%)	0=34(63.0%); 1=14(25.9%); 2=6(11.1%)	0=19(44.2%); 1=9(20.9%); 2=15(34.9%)
EF category	Categorical/ ordinal	14.2%	0=325(89.5%); 1=30(8.3%); 2=8(2.2%)	0=257(93.5%); 1=18(6.5%); 2=0(0.0%)	0=41(85.4%); 1=3(6.2%); 2=4(8.3%)	0=27(67.5%); 1=9(22.5%); 2=4(10.0%)
Glycated haemoglobin, %	Continuous	18.7%	8.84 ± 2.57	8.94 ± 2.61	8.70 ± 2.74	8.32 ± 2.00
White blood cell count, 10^9/L	Continuous	0.0%	10.14 [7.76, 13.68]	9.78 [7.67, 13.46]	11.12 [8.56, 13.22]	11.56 [9.10, 16.37]

Variable	Type	Missing (%)	Overall	Healed (n=319)	Persistently unhealed (n=61)	Below-knee amputation or death (n=43)
Haemoglobin	Continuous	0.0%	97.30 ± 21.66	98.75 ± 22.51	93.11 ± 16.91	92.46 ± 20.06
Platelet count, 10^9/L	Continuous	0.0%	343.00 [265.20, 469.55]	341.10 [264.45, 459.80]	348.30 [268.00, 480.00]	357.60 [291.05, 486.50]
Procalcitonin, ng/mL	Continuous	26.7%	0.17 [0.07, 0.45]	0.15 [0.07, 0.37]	0.23 [0.09, 0.84]	0.39 [0.16, 1.02]
High-sensitivity troponin T, ng/L	Continuous	48.2%	27.70 [17.15, 60.35]	25.54 [15.95, 46.80]	44.35 [15.70, 76.50]	76.00 [38.00, 157.00]
Serum troponin I, μg/L	Continuous	32.9%	0.01 [0.00, 0.02]	0.01 [0.00, 0.01]	0.01 [0.01, 0.02]	0.03 [0.01, 0.09]
Brain natriuretic peptide precursor	Continuous	21.7%	706.00 [247.15, 3261.00]	594.50 [214.75, 1826.75]	1299.00 [301.00, 5820.00]	6257.00 [853.75, 22139.00]
CRP, mg/L	Continuous	2.4%	50.10 [16.60, 106.14]	41.65 [12.38, 99.90]	64.60 [26.20, 110.20]	99.90 [48.00, 147.96]
Albumin category	Categorical/ ordinal	0.0%	0=2(0.5%); 1=46(10.9%); 2=98(23.2%); 3=245(57.9%); 4=32(7.6%)	0=2(0.6%); 1=32(10.0%); 2=69(21.6%); 3=189(59.2%); 4=27(8.5%)	0=0(0.0%); 1=6(9.8%); 2=17(27.9%); 3=35(57.4%); 4=3(4.9%)	0=0(0.0%); 1=8(18.6%); 2=12(27.9%); 3=21(48.8%); 4=2(4.7%)
Prealbumin category	Categorical/ ordinal	0.0%	0=107(25.3%); 1=150(35.5%); 2=166(39.2%)	0=82(25.7%); 1=106(33.2%); 2=131(41.1%)	0=11(18.0%); 1=28(45.9%); 2=22(36.1%)	0=14(32.6%); 1=16(37.2%); 2=13(30.2%)
Lymphocyte category	Categorical/ ordinal	0.0%	0=35(8.3%); 1=95(22.5%); 2=194(45.9%); 3=99(23.4%)	0=25(7.8%); 1=61(19.1%); 2=150(47.0%); 3=83(26.0%)	0=5(8.2%); 1=16(26.2%); 2=27(44.3%); 3=13(21.3%)	0=5(11.6%); 1=18(41.9%); 2=17(39.5%); 3=3(7.0%)
Triglyceride category	Categorical/ ordinal	6.6%	0=261(66.1%); 1=53(13.4%); 2=72(18.2%); 3=9(2.3%)	0=191(64.3%); 1=43(14.5%); 2=54(18.2%); 3=9(3.0%)	0=43(74.1%); 1=4(6.9%); 2=11(19.0%); 3=0(0.0%)	0=27(67.5%); 1=6(15.0%); 2=7(17.5%); 3=0(0.0%)
Total cholesterol category	Categorical/ ordinal	6.9%	0=99(25.1%); 1=249(63.2%); 2=31(7.9%); 3=15(3.8%)	0=67(22.6%); 1=188(63.5%); 2=29(9.8%); 3=12(4.1%)	0=20(34.5%); 1=33(56.9%); 2=2(3.4%); 3=3(5.2%)	0=12(30.0%); 1=28(70.0%); 2=0(0.0%); 3=0(0.0%)
ESR category	Categorical/ ordinal	0.9%	0=45(10.7%); 1=360(85.9%); 2=14(3.3%)	0=40(12.7%); 1=265(84.1%); 2=10(3.2%)	0=4(6.6%); 1=55(90.2%); 2=2(3.3%)	0=1(2.3%); 1=40(93.0%); 2=2(4.7%)
ESR, mm/h	Continuous	0.9%	70.92 ± 34.35	68.12 ± 34.91	77.87 ± 30.59	81.61 ± 32.44
eGFR category	Categorical/ ordinal	0.0%	1=61(14.4%); 2=133(31.4%); 3=151(35.7%); 4=31(7.3%); 5=47(11.1%)	1=52(16.3%); 2=108(33.9%); 3=114(35.7%); 4=22(6.9%); 5=23(7.2%)	1=7(11.5%); 2=18(29.5%); 3=22(36.1%); 4=4(6.6%); 5=10(16.4%)	1=2(4.7%); 2=7(16.3%); 3=15(34.9%); 4=5(11.6%); 5=14(32.6%)
Bacterial culture category	Categorical/ ordinal	59.3%	0=27(15.7%); 1=105(61.0%); 2=40(23.3%)	0=15(11.9%); 1=77(61.1%); 2=34(27.0%)	0=9(29.0%); 1=16(51.6%); 2=6(19.4%)	0=3(20.0%); 1=12(80.0%); 2=0(0.0%)
NLR, ratio	Continuous	0.0%	5.55 [3.48, 9.47]	5.26 [3.25, 8.74]	5.65 [3.78, 10.22]	8.72 [5.31, 14.59]
Previous debridement	Binary	0.0%	0=244(57.7%); 1=179(42.3%)	0=182(57.1%); 1=137(42.9%)	0=37(60.7%); 1=24(39.3%)	0=25(58.1%); 1=18(41.9%)
Previous anti-infective treatment	Binary	0.0%	0=113(26.7%); 1=310(73.3%)	0=81(25.4%); 1=238(74.6%)	0=21(34.4%); 1=40(65.6%)	0=11(25.6%); 1=32(74.4%)

Variable	Type	Missing (%)	Overall	Healed (n=319)	Persistently unhealed (n=61)	Below-knee amputation or death (n=43)
Previous negative-pressure therapy	Binary	0.0%	0=396(93.6%); 1=27(6.4%)	0=300(94.0%); 1=19(6.0%)	0=57(93.4%); 1=4(6.6%)	0=39(90.7%); 1=4(9.3%)
Wound area category	Categorical/ ordinal	0.0%	1=65(15.4%); 2=128(30.3%); 3=135(31.9%); 4=95(22.5%)	1=56(17.6%); 2=104(32.6%); 3=90(28.2%); 4=69(21.6%)	1=2(3.3%); 2=15(24.6%); 3=27(44.3%); 4=17(27.9%)	1=7(16.3%); 2=9(20.9%); 3=18(41.9%); 4=9(20.9%)
Palpated foot temperature grade	Categorical/ ordinal	0.0%	1=15(3.5%); 2=117(27.7%); 3=166(39.2%); 4=125(29.6%)	1=10(3.1%); 2=72(22.6%); 3=123(38.6%); 4=114(35.7%)	1=3(4.9%); 2=22(36.1%); 3=27(44.3%); 4=9(14.8%)	1=2(4.7%); 2=23(53.5%); 3=16(37.2%); 4=2(4.7%)
Posterior tibial artery occlusion	Binary	0.0%	0=305(72.1%); 1=118(27.9%)	0=239(74.9%); 1=80(25.1%)	0=41(67.2%); 1=20(32.8%)	0=25(58.1%); 1=18(41.9%)
Anterior tibial artery occlusion	Binary	0.0%	0=321(75.9%); 1=102(24.1%)	0=253(79.3%); 1=66(20.7%)	0=41(67.2%); 1=20(32.8%)	0=27(62.8%); 1=16(37.2%)
Peroneal artery occlusion	Binary	0.0%	0=371(87.7%); 1=52(12.3%)	0=285(89.3%); 1=34(10.7%)	0=50(82.0%); 1=11(18.0%)	0=36(83.7%); 1=7(16.3%)
Dorsalis pedis artery occlusion	Binary	0.0%	0=334(79.0%); 1=89(21.0%)	0=259(81.2%); 1=60(18.8%)	0=43(70.5%); 1=18(29.5%)	0=32(74.4%); 1=11(25.6%)
Preoperative random blood glucose, mmol/L	Continuous	0.2%	9.63 [7.21, 13.06]	9.42 [7.12, 13.14]	9.41 [6.86, 12.47]	11.15 [7.86, 13.44]
Ulcer duration category	Categorical/ ordinal	0.0%	0=126(29.8%); 1=149(35.2%); 2=72(17.0%); 3=76(18.0%)	0=98(30.7%); 1=111(34.8%); 2=55(17.2%); 3=55(17.2%)	0=19(31.1%); 1=21(34.4%); 2=9(14.8%); 3=12(19.7%)	0=9(20.9%); 1=17(39.5%); 2=8(18.6%); 3=9(20.9%)
Composite IBG	Categorical/ ordinal	27.4%	0=88(28.7%); 1=107(34.9%); 2=66(21.5%); 3=46(15.0%)	0=75(33.6%); 1=76(34.1%); 2=44(19.7%); 3=28(12.6%)	0=9(18.0%); 1=21(42.0%); 2=13(26.0%); 3=7(14.0%)	0=4(11.8%); 1=10(29.4%); 2=9(26.5%); 3=11(32.4%)

Data are observed values before imputation. Continuous variables are shown as mean ± SD or median [IQR], as indicated; categorical percentages use non-missing observations within each variable and outcome group as the denominator. For binary variables, 0 denotes absence and 1 presence, except sex, for which 0 denotes female and 1 male. Heart failure categories 0/1/2 denote none/chronic/acute; palpated foot temperature grades 1/2/3/4 denote cold/cool/warm/warmer. Renal grades use standardised eGFR (mL/min/1.73 m²), with documented uraemia assigned grade 5; numerical boundaries are defined in Supplementary Table 19. Wound-area categories 1–4 correspond to <5, 5–20, 20<A≤100 and >100 cm². Variables with >40% missingness in a given training fold were excluded from imputation and modelling in that fold but are retained here to document the complete preoperative candidate set. IBG and its component markers were not entered simultaneously into the AutoScore candidate space. Laboratory units, archived category definitions and source-field names are documented in Supplementary Data (candidate_input_dictionary.csv). The original field label “First random blood glucose” is displayed here as preoperative random blood glucose, following the latest-available preoperative measurement rule.

Supplementary Table 2 Death-related clinical conditions and event timing

Part A Recorded conditions and terminal descriptions

Recorded condition/description	n	%
Explicit heart failure/cardiorenal failure	10	35.7
Explicit renal injury/failure	1	3.6

Recorded condition/description	n	%
Heart-failure or renal label (union)	10	35.7
Severe infection/necrotising fasciitis	10	35.7
Heart-failure/renal label and infection	2	7.1
Cerebrovascular event	4	14.3
Gastrointestinal bleeding	2	7.1
Vascular thrombosis	1	3.6
Multiorgan failure	1	3.6
Malignancy history/bone tumour	2	7.1
Terminal cardiopulmonary arrest	4	14.3
Limited detail in case-summary field	3	10.7

Labels are non-exclusive and derived from source records, with a denominator of 28 throughout. Heart-failure and renal labels may overlap. Three entries with limited detail in the case-summary field also overlap with other tags. The table preserves this field-level information; the contemporaneous ICU clinical determination and joint review of the principal reasons for death are described in Supplementary Methods.

Part B Time from first TTT to the component event

Component event	n	Days, median (IQR)	Range	By day 30, n (%)	By day 90, n (%)
Death	28	64 (32.75–88)	7–170	7 (25.0)	21 (75.0)
Below-knee amputation	19	49 (37.5–82.5)	14–154	4 (21.1)	15 (78.9)

Timing describes all 28 deaths and 19 amputations with exact dates. Twenty-one deaths (75.0%) occurred on days 31–180. Among the four patients with both events within 180 days, the median interval from amputation to death was 7 days (range, 1–16). Event times are descriptive intervals from first TTT.

Supplementary Table 3 IBG definition and threshold provenance

Part A Component thresholds

Measure	0 points	1 point	2 points	Unit
WBC	4–10	>10–12	>12 or <4	$\times 10^9$/L
NLR	≤4.52	>4.52–7.5	>7.5	—
ESR	<70	70–94	>94	mm/h
CRP	<10	10–<100	≥100	mg/L
PCT	<0.10	0.10–0.50	>0.50	ng/mL

A, B and C are domain scores derived from component points: A=max(WBC,NLR), B=max(ESR,CRP) and C=PCT. A and B each use the higher within-domain score and contribute 0–2 points, avoiding addition of two markers from the same inflammatory dimension. The three domains represent leucocyte/neutrophil

response, acute-phase response, and PCT information, respectively. Sums of 0–2, 3–4, 5 and 6 correspond to IBG grades 0, 1, 2 and 3. WBC extremes (>12 or <4 $\times 10^9$/L) correspond to thresholds used in systemic inflammatory-response criteria. The WBC intermediate boundary, ESR ≥70, CRP and PCT boundaries were adopted in the study dictionary. NLR thresholds of >4.52 and >7.5 and ESR >94 were developed within the cohort. IBG was treated as a study-specific structured composite predictor. The updated complete-marker PCT-omission analysis is reported in Supplementary Table 14B and Supplementary Data (PCT_omission_complete_markers_updated.csv). The earlier available-component analysis is retained as labelled historical data. An isolated CRP elevation, even at ≥100 mg/L, gives B=2. With A=C=0, the total is 2 and IBG remains grade 0. Grades 2 and 3 require contributions from all three domains. Grade 0 denotes the lowest composite category and can include localised infection. Clinical infection is assessed from local and systemic findings alongside the composite grade.

Part B Sources of fixed thresholds

Variable/threshold	Source	Role of current cohort	Outcome information in derivation	Implication for internal assessment
WBC >12 or <4; >10	WBC >12 or <4: systemic inflammatory-response thresholds; >10: study-adopted intermediate boundary	No (adopted directly)	No	Fixed; not tuned in this cohort
ESR ≥70	Study-adopted threshold; ESR ≥70 mm/h has been evaluated in diabetic foot osteomyelitis diagnosis	No (adopted directly)	No	Fixed
ESR >94	Cohort-derived threshold; no external reference identified in available records	Yes	Available archived code did not document outcome-guided threshold searching; original derivation records were incomplete	Entered the current internal assessments as a fixed development specification
CRP ≥10 / ≥100	Study-adopted CRP boundaries recorded in the variable dictionary	No (adopted directly)	No	Fixed
PCT ≥0.1 / >0.5	Study-adopted PCT boundaries recorded in the variable dictionary	No (adopted directly)	No	Fixed
NLR >4.52 / >7.5	Cohort-derived threshold; no external reference identified in available records	Yes	Available archived code did not document outcome-guided threshold searching; original derivation records were incomplete	Entered the current internal assessments as a fixed development specification
IBG domain maxima and mapping to grades 0–3	Study-specific composite rule	Yes	Fixed composite rule	Fixed development specification; reproducibility should be assessed externally
Age≥65 years	Clinical convention and cohort development comparisons	Yes	Yes	Selected during development; cut-off fixed in repeated CV
Wound area category	Clinical categories defined before modelling	Applied only	No outcome-guided optimisation documented	Fixed
eGFR categories 1–5	Study-adopted renal-function categories recorded in the clinical dictionary	Applied only	No	Fixed
Foot temperature grades 1–4; heart failure categories 0–2	Original clinical records/definitions	Applied only	No	Fixed clinical definitions; cross-centre measurement consistency remains to be assessed

Variable/threshold	Source	Role of current cohort	Outcome information in derivation	Implication for internal assessment
Risk cut-offs 24/34	Fixed after training-data QWK optimisation	Yes	Yes	Development-stage risk cut-offs
60-point target (maximum 59)	Cohort development selection	Yes	Yes	Selected during development; specification fixed in repeated CV

Threshold sources were traced from available grading records, development code and fixed configurations. For the upper NLR and ESR thresholds, original derivation records were incomplete; available archived code did not document outcome-guided threshold searching. These cut-offs were therefore treated as fixed cohort-derived specifications in the current analyses.

Supplementary Table 4 Models, scoring procedures, and development choices

Assessment	Identity / complexity	IBG and encoding	Weights and mapping
Whole-cohort integer card	Six; whole-cohort derived	Existing IBG; age 65; 60 target	Fixed 0–59 points; 24/34 boundaries
Primary post-selection five-fold OOF	Six identities fixed	Existing IBG; age 65; 60 target	Points/maps refitted within folds; mAUC 0.8047
Earlier fixed-K = 6 repeated CV	K = 6; identities reselected	Existing IBG; age 65; 60 target	Points/maps refitted; mAUC 0.7813±0.0151
AUTO original / align	Inner-selected K and identities	Training-learned IBG; selected age/scale/precursor	Refitted points; original or aligned inner OOF mapping
SF original / align	Exactly six identities fixed	Existing IBG; age 65; 60 target	Refitted points; original or aligned mapping
SL original / align	Exactly six identities fixed	Training-learned IBG; age 65; 60 target	Refitted points; original or aligned mapping
Same-centre temporal assessment	Six identities fixed	Whole-development upstream specifications	2020–24 training; 2025–26 later assessment

The whole-cohort card, fold-refitted six-predictor procedure, and specified automated workflow are evaluated separately. Subsequent matched repeated estimates use SF_align; locked complete-input probability maps are specified in Supplementary Table 19.

Supplementary Table 5 Repeated predictive performance, calibration, and stability

Part A Predictive performance

Procedure	mAUC (95% CI)	AUC Y≥1	AUC Y=2	Multiclass Brier (95% CI)	mAUC repeat SD
Six existing IBG original	0.8044 (0.7499–0.8514)	0.7817	0.8270	0.3338 (0.2877–0.3820)	0.0089
SF_align follow-up: six existing IBG aligned	0.8052 (0.7507–0.8519)	0.7824	0.8279	0.3335 (0.2869–0.3818)	0.0086
Six learned IBG original	0.8052 (0.7518–0.8510)	0.7797	0.8307	0.3342 (0.2880–0.3821)	0.0085
Six learned IBG aligned	0.8070 (0.7534–0.8527)	0.7815	0.8325	0.3337 (0.2870–0.3818)	0.0086
Automatic original	0.7687 (0.7143–0.8163)	0.7446	0.7927	0.3500 (0.3045–0.3974)	0.0240
Automatic aligned	0.7825 (0.7262–0.8318)	0.7557	0.8093	0.3447 (0.2985–0.3933)	0.0131
SINBAD raw score	0.6363 (0.5782–0.6973)	0.6445	0.6280	0.3854 (0.3378–0.4369)	0.0000

Procedure	mAUC (95% CI)	AUC Y≥1	AUC Y=2	Multiclass Brier (95% CI)	mAUC repeat SD
SINBAD mapped probability	0.6182 (0.5571–0.6816)	0.6289	0.6075	0.3854 (0.3378–0.4369)	0.0087
UT stage B/D	0.5816 (0.5183–0.6456)	0.5555	0.6077	0.3946 (0.3457–0.4463)	0.0088
Wagner mapped probability	0.5518 (0.4968–0.6029)	0.5763	0.5273	0.3943 (0.3451–0.4465)	0.0086

423 patients; ten repeated outer five-fold evaluations; M = 20 for scoring procedures. Confidence intervals use 2,000 patient-cluster resamples conditional on fixed OOF predictions. Raw SINBAD AUCs use the original score; its Brier score uses the same map as probability-based SINBAD. Supplementary Data supplies all six metric intervals, ranges and repeat values.

Part B Cumulative calibration

Procedure	Intercept Y≥1	Slope Y≥1	O/E Y≥1	Intercept Y=2	Slope Y=2	O/E Y=2
SF_original	0.030 (-0.306–0.401)	1.061 (0.822–1.360)	0.984 (0.838–1.129)	0.010 (-0.586–0.658)	1.027 (0.745–1.386)	0.975 (0.743–1.224)
SF_align	0.036 (-0.301–0.413)	1.035 (0.803–1.324)	1.001 (0.853–1.149)	0.023 (-0.576–0.675)	1.014 (0.737–1.365)	1.000 (0.764–1.254)
SL_original	0.041 (-0.295–0.422)	1.062 (0.825–1.373)	0.989 (0.841–1.136)	0.028 (-0.575–0.676)	1.031 (0.753–1.388)	0.983 (0.750–1.232)
SL_align	0.037 (-0.300–0.420)	1.035 (0.803–1.343)	1.002 (0.852–1.151)	0.030 (-0.571–0.682)	1.017 (0.738–1.364)	1.001 (0.765–1.256)
AUTO_original	0.102 (-0.247–0.471)	1.143 (0.880–1.456)	0.979 (0.834–1.127)	-0.015 (-0.678–0.689)	1.014 (0.711–1.397)	0.981 (0.746–1.240)
AUTO_align	0.130 (-0.225–0.525)	1.124 (0.867–1.440)	1.004 (0.855–1.152)	0.074 (-0.599–0.807)	1.042 (0.737–1.431)	1.002 (0.765–1.265)
SINBAD_raw	-0.022 (-0.518–0.514)	0.975 (0.567–1.458)	1.003 (0.843–1.168)	-0.348 (-1.612–0.945)	0.831 (0.270–1.463)	1.004 (0.745–1.295)
SINBAD_probability	-0.022 (-0.518–0.514)	0.975 (0.567–1.458)	1.003 (0.843–1.168)	-0.348 (-1.612–0.945)	0.831 (0.270–1.463)	1.004 (0.745–1.295)
UT_stage	-0.202 (-0.884–0.421)	0.823 (0.274–1.372)	0.992 (0.835–1.156)	0.753 (-0.920–2.298)	1.368 (0.607–2.145)	0.990 (0.734–1.278)
Wagner_probability	0.017 (-0.776–0.762)	1.008 (0.345–1.679)	1.006 (0.848–1.169)	-1.435 (-3.304–0.378)	0.335 (-0.473–1.173)	1.006 (0.740–1.299)

Values are means across repetitions with 2,000 patient-cluster 95% intervals conditional on fixed OOF predictions. Joint weighted logit estimates are recomputed for each cumulative endpoint in each resample, then averaged across ten repeats, using the same patient weights as the performance bootstrap. Supplementary Data supplies repeat SDs and ranges. Ideal intercept, slope and O/E are 0, 1 and 1, respectively. Negative probabilities caused by crossings are clipped and class probabilities renormalised before M = 20 pooling.

Part C Patient and fold variability

Procedure	Patient-level severe-risk SD, median [IQR]	Assigned to ≥2 strata, n/423	Fold-level mean-risk error, median	Fold-level mean-risk error, maximum
SF_original	0.0114 [0.0065, 0.0229]	141/423	0.0133	0.0398
SF_align	0.0103 [0.0060, 0.0207]	141/423	0.0085	0.0319
SL_original	0.0124 [0.0067, 0.0232]	119/423	0.0150	0.0371
SL_align	0.0109 [0.0061, 0.0211]	119/423	0.0096	0.0328
AUTO_original	0.0287 [0.0182, 0.0504]	196/423	0.0203	0.0929
AUTO_align	0.0211 [0.0114, 0.0415]	196/423	0.0122	0.0332
SINBAD_raw	0.0036 [0.0030, 0.0042]	0/423	0.0081	0.0260
SINBAD_probability	0.0036 [0.0030, 0.0042]	0/423	0.0081	0.0260
UT_stage	0.0036 [0.0030, 0.0061]	0/423	0.0072	0.0185
Wagner_probability	0.0019 [0.0015, 0.0044]	0/423	0.0060	0.0187

SD of patient-level severe-risk predictions summarises ten OOF probabilities. Group instability denotes assignment to at least two distinct strata across repetitions; SF, SL and AUTO use training-QWK groups, whereas benchmark groups use the highest predicted class probability. Fold error is the absolute difference between mean predicted severe risk and observed severe event proportion. Original and aligned scoring groups are identical. Supplementary Table 10 details the number of distinct strata for SF_align. Across-repeat stratum changes quantify sensitivity to training partitions.

Part D Class-specific discrimination and probability calibration

Outcome	n	OVR AUC (95% CI)	Binary Brier (95% CI)	Calibration intercept (95% CI)	Calibration slope (95% CI)	O/E (95% CI)
Healing	319	0.782 (0.732 to 0.830)	0.1458 (0.1262 to 0.1664)	-0.036 (-0.413 to 0.301)	1.035 (0.803 to 1.324)	1.000 (0.951 to 1.049)
Persistent non-healing	61	0.683 (0.618 to 0.746)	0.1176 (0.0956 to 0.1399)	0.046 (-0.709 to 0.889)	1.027 (0.637 to 1.529)	1.002 (0.779 to 1.223)
Below-knee amputation or death	43	0.828 (0.755 to 0.889)	0.0702 (0.0527 to 0.0893)	0.023 (-0.576 to 0.675)	1.014 (0.737 to 1.365)	1.000 (0.764 to 1.254)

SF_align predictions were retained unchanged. For each class k, one-versus-rest (OVR) AUC and binary Brier score compare I(Y=k) with P(Y=k); intercept and slope are jointly estimated in logit P(Y=k). Values are the means of ten complete OOF repeats; intervals use the same 2,000 patient-cluster bootstrap resamples as the matched repeated assessment, with calibration regressions refitted in each resample. Ideal intercept, slope and observed/expected event ratio (O/E) are 0, 1 and 1. The three binary Brier means sum to the SF_align multiclass Brier of 0.3335. Supplementary Fig. 5 presents first-repeat calibration; repeat estimates and coordinates are in Supplementary Data.

Supplementary Table 6 Clinical grading benchmarks

Part A SINBAD and UT distributions

Classification	Score or level	n	Percent
SINBAD total	2	0	0.0

Classification	Score or level	n	Percent
SINBAD total	3	52	12.3
SINBAD total	4	187	44.2
SINBAD total	5	155	36.6
SINBAD total	6	29	6.9
UT grade	3	423	100.0
UT stage	A	0	0.0
UT stage	B	315	74.5
UT stage	C	0	0.0
UT stage	D	108	25.5
SINBAD site component	1	187	44.2
SINBAD neuropathy diagnosis proxy	1	289	68.3
SINBAD ischaemia component	1	108	25.5
SINBAD depth component	1	423	100.0
SINBAD area component	1	423	100.0
SINBAD infection component	1	423	100.0

All 423 patients had complete source-reviewed retrospective SINBAD and UT coding. SINBAD totals ranged from 3 to 6. Its depth, area and infection components were each 1 throughout; site was 1 in 187 patients, diagnosis-based neuropathy in 289 and ischaemia in 108. Neuropathy was coded from documented clinical diabetic peripheral neuropathy diagnoses as a proxy for loss of protective sensation. UT grade was 3 throughout; infection was present in all patients, with stage B in 315 and D in 108. UT benchmarking therefore assesses variation in ischaemia within this infected grade-3 cohort.

Part B Paired comparisons in the repeated assessment

Difference direction	ΔmAUC (95% CI)	Δ multiclass Brier (95% CI)
SF_align minus SF_original	+0.0008 (-0.0003–+0.0021)	-0.0003 (-0.0010–+0.0005)
SL_align minus SL_original	+0.0018 (+0.0005–+0.0032)	-0.0005 (-0.0013–+0.0003)
AUTO_align minus AUTO_original	+0.0138 (+0.0065–+0.0215)	-0.0053 (-0.0082–-0.0024)
SL_align minus SF_align	+0.0019 (-0.0029–+0.0073)	+0.0002 (-0.0018–+0.0022)
SL_align minus AUTO_align	+0.0245 (+0.0093–+0.0407)	-0.0110 (-0.0190–-0.0033)
SF_align minus SINBAD_raw	+0.1689 (+0.1014–+0.2310)	-0.0519 (-0.0770–-0.0271)
SF_align minus SINBAD_probability	+0.1870 (+0.1186–+0.2514)	-0.0519 (-0.0770–-0.0271)
SF_align minus UT_stage	+0.2236 (+0.1536–+0.2944)	-0.0611 (-0.0865–-0.0372)
SF_align minus Wagner_probability	+0.2534 (+0.1817–+0.3175)	-0.0608 (-0.0871–-0.0353)
SL_align minus SINBAD_raw	+0.1708 (+0.1043–+0.2335)	-0.0517 (-0.0773–-0.0271)
SL_align minus SINBAD_probability	+0.1889 (+0.1212–+0.2525)	-0.0517 (-0.0773–-0.0271)

Difference direction	ΔmAUC (95% CI)	Δ multiclass Brier (95% CI)
SL_align minus UT_stage	+0.2254 (+0.1561–+0.2966)	-0.0609 (-0.0866–-0.0373)
SL_align minus Wagner_probability	+0.2553 (+0.1853–+0.3184)	-0.0606 (-0.0867–-0.0354)
AUTO_align minus SINBAD_raw	+0.1462 (+0.0773–+0.2092)	-0.0407 (-0.0631–-0.0184)
AUTO_align minus SINBAD_probability	+0.1644 (+0.0947–+0.2286)	-0.0407 (-0.0631–-0.0184)
AUTO_align minus UT_stage	+0.2009 (+0.1317–+0.2695)	-0.0499 (-0.0719–-0.0292)
AUTO_align minus Wagner_probability	+0.2308 (+0.1599–+0.2952)	-0.0496 (-0.0726–-0.0277)

Positive mAUC and negative Brier differences favour the first procedure. Comparisons use identical patient resamples. Raw and probability-based SINBAD use the same probability map for error assessment but assess discrimination differently. These are matched same-cohort follow-up comparisons.

Supplementary Table 7 Incremental prediction beyond the four-predictor clinical reference

Procedure	Metric	Estimate (95% interval)	Difference versus four-predictor reference (95% interval)
Four-predictor reference / required	Severe binary Brier	0.0734 (0.0549 to 0.0925)	Reference
Four-predictor reference / required	Severe AUC	0.7927 (0.7071 to 0.8691)	Reference
Four-predictor reference / required	Cumulative mAUC	0.7626 (0.7002 to 0.8190)	Reference
Four-predictor reference / required	Multiclass Brier	0.3499 (0.3022 to 0.3963)	Reference
six-predictor IBG / required	Severe binary Brier	0.0701 (0.0524 to 0.0887)	-0.0032 (-0.0075 to 0.0010)
six-predictor IBG / required	Severe AUC	0.8265 (0.7526 to 0.8891)	0.0338 (-0.0139 to 0.0770)
six-predictor IBG / required	Cumulative mAUC	0.8049 (0.7490 to 0.8535)	0.0424 (0.0061 to 0.0761)
six-predictor IBG / required	Multiclass Brier	0.3331 (0.2871 to 0.3769)	-0.0168 (-0.0283 to -0.0057)
six-predictor CRP / required	Severe binary Brier	0.0687 (0.0513 to 0.0871)	-0.0046 (-0.0092 to -0.0003)
six-predictor CRP / required	Severe AUC	0.8322 (0.7594 to 0.8952)	0.0395 (-0.0091 to 0.0850)
six-predictor CRP / required	Cumulative mAUC	0.8093 (0.7562 to 0.8571)	0.0467 (0.0089 to 0.0824)

Procedure	Metric	Estimate (95% interval)	Difference versus four-predictor reference (95% interval)
six-predictor CRP / required	Multiclass Brier	0.3313 (0.2853 to 0.3763)	-0.0185 (-0.0312 to -0.0061)
six-predictor IBG / common	Severe binary Brier	0.0702 (0.0524 to 0.0888)	-0.0032 (-0.0074 to 0.0010)
six-predictor IBG / common	Severe AUC	0.8279 (0.7550 to 0.8896)	0.0352 (-0.0113 to 0.0788)
six-predictor IBG / common	Cumulative mAUC	0.8052 (0.7493 to 0.8533)	0.0426 (0.0070 to 0.0767)
six-predictor IBG / common	Multiclass Brier	0.3335 (0.2874 to 0.3778)	-0.0164 (-0.0276 to -0.0053)

Required-input six-predictor IBG minus required-input four-predictor reference for severe-outcome binary Brier score is the prespecified primary additional contrast. Multiclass Brier and cumulative mAUC are secondary. Common-auxiliary IBG also changes the imputation information scope. All values retain the dedicated comparison bootstrap intervals; therefore identical point estimates can have different intervals from Table 5. CRP comparisons with the four-predictor clinical reference are secondary and unadjusted for multiplicity; direct IBG–CRP contrasts are reported in Table 8.

Supplementary Table 8 IBG and graded CRP prediction comparisons

Part A Predictive performance

Procedure	Cumulative mAUC	AUC, any adverse	AUC, severe	Multiclass Brier	QWK	Log loss
IBG, common auxiliaries	0.8052 (0.7524 to 0.8532)	0.7824 (0.7312 to 0.8330)	0.8279 (0.7557 to 0.8920)	0.3335 (0.2882 to 0.3801)	0.4322 (0.3275 to 0.5235)	0.6066 (0.5318 to 0.6826)
CRP, common auxiliaries	0.8087 (0.7575 to 0.8554)	0.7848 (0.7367 to 0.8327)	0.8327 (0.7642 to 0.8949)	0.3318 (0.2865 to 0.3771)	0.4472 (0.3473 to 0.5348)	0.6002 (0.5276 to 0.6753)
IBG, required inputs	0.8049 (0.7516 to 0.8531)	0.7834 (0.7325 to 0.8343)	0.8265 (0.7531 to 0.8913)	0.3331 (0.2881 to 0.3797)	0.4432 (0.3422 to 0.5313)	0.6062 (0.5313 to 0.6827)
CRP, required inputs	0.8093 (0.7581 to 0.8556)	0.7864 (0.7379 to 0.8346)	0.8322 (0.7630 to 0.8947)	0.3313 (0.2862 to 0.3765)	0.4389 (0.3383 to 0.5292)	0.5996 (0.5270 to 0.6746)

Mean of ten complete OOF repeats; parentheses show 95% patient-cluster intervals conditional on fixed OOF predictions. Each repeat includes 423 patients with 319 healed, 61 persistently unhealed and 43 severe outcomes. Higher mAUC, AUC, and QWK indicate better performance, as do lower Brier score and log loss. IBG/common point estimates reproduce the existing SF_align procedure; its interval uses the dedicated comparison bootstrap seed. Generalised concordance, repeat SDs and ranges are archived in summary.csv and repeat_metrics.csv.

Part B Paired contrasts for marker representation and input scope

Contrast	Metric	Mean difference	Lower 95%	Upper 95%
CRP, common auxiliaries minus IBG, common auxiliaries	Cumulative mAUC	0.0036	-0.0092	0.0165
CRP, common auxiliaries minus IBG, common auxiliaries	AUC, any adverse	0.0024	-0.0135	0.0193
CRP, common auxiliaries minus IBG, common auxiliaries	AUC, severe	0.0048	-0.0105	0.0198
CRP, common auxiliaries minus IBG, common auxiliaries	Multiclass Brier	-0.0017	-0.0075	0.0039
CRP, common auxiliaries minus IBG, common auxiliaries	Generalised C	0.0031	-0.0113	0.0183
CRP, common auxiliaries minus IBG, common auxiliaries	QWK	0.0151	-0.0200	0.0487
CRP, common auxiliaries minus IBG, common auxiliaries	Log loss	-0.0064	-0.0171	0.0042
CRP, required inputs minus IBG, required inputs	Cumulative mAUC	0.0044	-0.0091	0.0179

Contrast	Metric	Mean difference	Lower 95%	Upper 95%
CRP, required inputs minus IBG, required inputs	AUC, any adverse	0.0030	-0.0133	0.0195
CRP, required inputs minus IBG, required inputs	AUC, severe	0.0057	-0.0106	0.0215
CRP, required inputs minus IBG, required inputs	Multiclass Brier	-0.0017	-0.0076	0.0041
CRP, required inputs minus IBG, required inputs	Generalised C	0.0035	-0.0114	0.0188
CRP, required inputs minus IBG, required inputs	QWK	-0.0042	-0.0337	0.0234
CRP, required inputs minus IBG, required inputs	Log loss	-0.0066	-0.0172	0.0042
CRP, required inputs minus CRP, common auxiliaries	Cumulative mAUC	0.0006	-0.0008	0.0022
CRP, required inputs minus CRP, common auxiliaries	AUC, any adverse	0.0016	-0.0006	0.0049
CRP, required inputs minus CRP, common auxiliaries	AUC, severe	-0.0005	-0.0020	0.0009
CRP, required inputs minus CRP, common auxiliaries	Multiclass Brier	-0.0004	-0.0013	0.0003
CRP, required inputs minus CRP, common auxiliaries	Generalised C	0.0013	-0.0008	0.0040
CRP, required inputs minus CRP, common auxiliaries	QWK	-0.0083	-0.0195	0.0024
CRP, required inputs minus CRP, common auxiliaries	Log loss	-0.0006	-0.0022	0.0007
IBG, required inputs minus IBG, common auxiliaries	Cumulative mAUC	-0.0002	-0.0018	0.0014
IBG, required inputs minus IBG, common auxiliaries	AUC, any adverse	0.0010	-0.0007	0.0030
IBG, required inputs minus IBG, common auxiliaries	AUC, severe	-0.0014	-0.0037	0.0006
IBG, required inputs minus IBG, common auxiliaries	Multiclass Brier	-0.0005	-0.0012	0.0002
IBG, required inputs minus IBG, common auxiliaries	Generalised C	0.0008	-0.0009	0.0027
IBG, required inputs minus IBG, common auxiliaries	QWK	0.0110	-0.0019	0.0250
IBG, required inputs minus IBG, common auxiliaries	Log loss	-0.0004	-0.0015	0.0008

Contrasts were paired using identical patient resamples. The primary contrast was CRP/common minus IBG/common multiclass Brier score. Required-input comparisons assessed reduced collection requirements. For both imputation strategies, intervals for Brier and mAUC differences included zero and were compatible with differences in either direction. Equivalence and non-inferiority were not tested.

Supplementary Table 9 IBG and CRP calibration, availability, and stability

Part A Cumulative calibration

Procedure	Endpoint	Joint intercept	Joint slope	Observed/expected	Fixed-slope intercept
IBG, common auxiliaries	Any adverse outcome	0.036 (-0.297 to 0.416)	1.035 (0.805 to 1.326)	1.001 (0.856 to 1.149)	0.002 (-0.253 to 0.238)
IBG, common auxiliaries	Severe	0.023 (-0.575 to 0.690)	1.014 (0.737 to 1.375)	1.000 (0.764 to 1.259)	-0.001 (-0.363 to 0.335)
CRP, common auxiliaries	Any adverse outcome	0.016 (-0.308 to 0.385)	1.020 (0.804 to 1.300)	0.998 (0.852 to 1.146)	-0.003 (-0.261 to 0.232)
CRP, common auxiliaries	Severe	-0.003 (-0.550 to 0.588)	1.005 (0.754 to 1.347)	0.993 (0.762 to 1.249)	-0.011 (-0.375 to 0.326)
IBG, required inputs	Any adverse outcome	0.037 (-0.298 to 0.416)	1.038 (0.811 to 1.328)	1.000 (0.855 to 1.149)	0.000 (-0.256 to 0.237)
IBG, required inputs	Severe	0.017 (-0.578 to 0.677)	1.012 (0.737 to 1.375)	0.997 (0.763 to 1.255)	-0.004 (-0.367 to 0.331)
CRP, required inputs	Any adverse outcome	0.021 (-0.299 to 0.391)	1.026 (0.808 to 1.306)	0.998 (0.853 to 1.147)	-0.003 (-0.258 to 0.234)
CRP, required inputs	Severe	0.002 (-0.548 to 0.598)	1.007 (0.756 to 1.350)	0.993 (0.763 to 1.252)	-0.010 (-0.376 to 0.328)

Mean of ten repeats with 95% intervals. Calibration intercept and slope are jointly estimated for each cumulative endpoint; the fixed-slope intercept (CITL) estimates calibration in the large with the slope fixed at 1. Ideal intercepts are 0 and ideal slope and observed/expected ratio are 1. Calibration regressions are refitted within each patient-cluster bootstrap sample. The severe-outcome intervals reflect the 43 observed events.

Part B Observed input completeness

Score input set	Raw fields	All fields observed, n	Complete, %	Missing ≥1 field, n
IBG	10	301	71.2%	122
CRP	6	396	93.6%	27
Four-predictor reference	4	405	95.7%	18

The five common fields are age, heart failure, palpated foot temperature, eGFR and wound area. IBG additionally requires WBC, NLR, ESR, CRP and PCT; the CRP score additionally requires CRP alone. Before imputation, heart failure was missing in 18/423 (4.3%), ESR in 4/423 (0.9%), CRP in 10/423 (2.4%) and PCT in 113/423 (26.7%); the remaining six fields were complete. The raw completeness difference is 22.5 percentage points. Completeness describes score inputs; common-auxiliary imputation additionally uses the candidate matrix.

Part C Variability in predictions and stratum assignment

Procedure	Median severe-risk SD	SD interquartile range	Changed group, n	Changed group, %
IBG, common auxiliaries	0.0103	0.0060 to 0.0207	141	33.3%
CRP, common auxiliaries	0.0106	0.0055 to 0.0222	130	30.7%
IBG, required inputs	0.0100	0.0059 to 0.0212	141	33.3%
CRP, required inputs	0.0108	0.0057 to 0.0220	138	32.6%

One record per patient. Severe-risk SD is calculated from ten OOF probabilities. A group change denotes assignment to at least two QWK-defined strata across ten repetitions and summarises variability in boundaries and predictions. All branches have median severe-risk SD near 0.01. These are descriptive stability summaries.

Supplementary Table 10 Outcome gradient and category assignment

Part A Repeated risk gradient

Stratum	Patients, mean ± SD	Range of patient counts	Healed % mean ± SD	Persistent non-healing % mean ± SD	Severe % mean ± SD	Any adverse % mean ± SD
Low	302.7 ± 12.3	283–317	84.6 ± 0.8	11.3 ± 0.7	4.1 ± 0.5	15.4 ± 0.8
Intermediate	87.2 ± 10.6	73–110	61.2 ± 4.3	22.3 ± 4.5	16.5 ± 2.2	38.8 ± 4.3
High	33.1 ± 6.7	25–47	28.0 ± 3.8	21.5 ± 6.2	50.5 ± 7.6	72.0 ± 3.8

SF_align; 423 patients per repeat. Means and sample SDs summarise stratum-specific results across ten repetitions. Stratum size can change across repeats. Percentages are averaged after division by the corresponding repeat-specific stratum size. Any adverse outcome is Y≥1; severe outcome is Y=2. Main Fig. 2A shows the three mean outcome proportions.

Part B Class-specific performance

Procedure	Outcome	Sensitivity mean ± SD	Specificity mean ± SD	Positive predictive value mean ± SD
SF_align follow-up: six existing IBG aligned	Healed	80.3% ± 2.9%	55.2% ± 3.6%	84.6% ± 0.8%

Procedure	Outcome	Sensitivity mean ± SD	Specificity mean ± SD	Positive predictive value mean ± SD
SF_align follow-up: six existing IBG aligned	Persistent non-healing	31.8% ± 6.9%	81.3% ± 2.6%	22.3% ± 4.5%
SF_align follow-up: six existing IBG aligned	Below-knee amputation or death	37.9% ± 2.9%	95.6% ± 1.6%	50.5% ± 7.6%
Six learned IBG aligned	Healed	82.6% ± 2.5%	54.0% ± 3.0%	84.7% ± 0.7%
Six learned IBG aligned	Persistent non-healing	34.3% ± 5.0%	83.0% ± 2.3%	25.4% ± 2.7%
Six learned IBG aligned	Below-knee amputation or death	36.5% ± 4.1%	96.5% ± 1.4%	55.5% ± 8.1%
Automatic workflow aligned	Healed	80.6% ± 2.5%	52.8% ± 3.8%	84.0% ± 0.9%
Automatic workflow aligned	Persistent non-healing	33.8% ± 5.9%	81.7% ± 2.4%	23.8% ± 3.4%
Automatic workflow aligned	Below-knee amputation or death	40.5% ± 4.7%	96.6% ± 0.8%	58.0% ± 6.7%

Values are arithmetic means and sample SDs of ten complete 423-patient OOF evaluations. SD describes variation across training partitions. QWK boundaries are estimated from patient-pooled M = 20 training scores, and scoring points are re-estimated within training partitions. Original and aligned maps share scores and class assignments. The full assignment matrix and individual distinct-stratum counts are reported in Supplementary Table 10.

Part C Observed outcome and assigned stratum

Observed outcome	Patients, n	Assigned low % mean ± SD	Assigned intermediate % mean ± SD	Assigned high % mean ± SD
Healed	319	80.3 ± 2.9	16.8 ± 2.6	2.9 ± 0.9
Persistent non-healing	61	56.1 ± 4.2	31.8 ± 6.9	12.1 ± 5.6
Below-knee amputation or death	43	28.8 ± 4.4	33.3 ± 4.8	37.9 ± 2.9

Rows contain the three observed 180-day outcomes; columns contain training-QWK stratum assignments. Each row percentage is computed within a repeat, then averaged across ten repeats. Sample SD describes variation across training partitions. Row means sum to 100% before rounding. The diagonal corresponds to class sensitivity in Supplementary Table 10 and main Fig. 2B.

Part D Number of distinct assigned strata

Distinct strata assigned over ten repeats	Patients, n	Patients, %
1	282	66.7
2	138	32.6
3	3	0.7

SF_align; one record per patient. A total of 141/423 patients (33.3%) were assigned to at least two distinct strata. The three rows are mutually exclusive and sum to 423. Original and aligned SF mappings share these assignments. This table supplies main Fig. 2C and expands the group-instability summary in Supplementary Table 5.

Supplementary Table 11 Exploratory component ranking and detection

Part A Ranking using the predicted probability of severe outcomes

Component event	Procedure	Ranking AUC (95% interval)	Paired AUC difference (95% interval)
Death (n=28)	Four-predictor reference / required	0.7769 (0.6674 to 0.8676)	Reference
Death (n=28)	six-predictor IBG / required	0.7804 (0.6813 to 0.8711)	0.0036 (-0.0491 to 0.0559)
Death (n=28)	six-predictor CRP / required	0.7926 (0.7014 to 0.8725)	0.0157 (-0.0416 to 0.0757)
Death (n=28)	six-predictor IBG / common	0.7826 (0.6846 to 0.8718)	0.0057 (-0.0468 to 0.0586)
Amputation (n=19)	Four-predictor reference / required	0.8188 (0.6996 to 0.9088)	Reference
Amputation (n=19)	six-predictor IBG / required	0.8811 (0.8097 to 0.9370)	0.0623 (0.0133 to 0.1256)
Amputation (n=19)	six-predictor CRP / required	0.8763 (0.7950 to 0.9407)	0.0575 (0.0111 to 0.1137)
Amputation (n=19)	six-predictor IBG / common	0.8807 (0.8095 to 0.9369)	0.0619 (0.0130 to 0.1243)

AUCs and paired differences use existing composite severe-outcome probabilities to rank death events (28 versus 395 without death) and amputation events (19 versus 404 without amputation). Four patients contribute events to both targets. All comparisons are exploratory with patient-cluster intervals conditional on fixed OOF predictions. Required-input and common-auxiliary imputation scopes remain distinct.

Part B Detection at the 10% composite-risk threshold

Component event	Procedure	Mean identified, n	Detection %, 95% interval
Death (n=28)	Four-predictor reference / required	18.5	66.1 (48.9 to 81.4)
Death (n=28)	six-predictor IBG / required	19.7	70.4 (54.0 to 84.8)
Death (n=28)	six-predictor CRP / required	19.9	71.1 (55.3 to 85.5)
Death (n=28)	six-predictor IBG / common	19.0	67.9 (51.3 to 82.6)
Amputation (n=19)	Four-predictor reference / required	14.0	73.7 (55.4 to 90.0)
Amputation (n=19)	six-predictor IBG / required	15.7	82.6 (64.5 to 96.9)
Amputation (n=19)	six-predictor CRP / required	15.9	83.7 (65.5 to 98.5)

Component event	Procedure	Mean identified, n	Detection %, 95% interval
Amputation (n=19)	six-predictor IBG / common	15.5	81.6 (63.6 to 95.6)

Counts are means over ten OOF evaluations. Required-input IBG identifies a mean 19.7 patients with death and 15.7 with amputation, including 3.9 with both events. The union is 19.7 + 15.7 − 3.9 = 31.5/43 (73.3%). For the four-predictor reference, the corresponding union is 18.5 + 14.0 − 4.0 = 28.5/43 (66.3%). Detection uses the 10% composite-risk threshold, not a component-specific probability threshold.

Supplementary Table 12 Matched contribution of palpated foot temperature

Part A Matched refitting performance

Model	Measure	Estimate (95% CI)
Earlier five-fold point-score reference	Cumulative mAUC	0.8047 (0.7501 to 0.8545)
Earlier five-fold point-score reference	Generalised C	0.7766 (0.7291 to 0.8221)
Earlier five-fold point-score reference	AUC Y≥1	0.7800 (0.7310 to 0.8282)
Earlier five-fold point-score reference	AUC Y=2	0.8293 (0.7554 to 0.8910)
Earlier five-fold point-score reference	Multiclass Brier	0.3323 (0.2878 to 0.3784)
Earlier five-fold point-score reference	Mean outcome-pair ranking AUC	0.7660 (0.7142 to 0.8146)
Matched six-predictor CLM	Cumulative mAUC	0.7932 (0.7352 to 0.8467)
Matched six-predictor CLM	Generalised C	0.7629 (0.7133 to 0.8101)
Matched six-predictor CLM	AUC Y≥1	0.7678 (0.7167 to 0.8172)
Matched six-predictor CLM	AUC Y=2	0.8185 (0.7424 to 0.8849)
Matched six-predictor CLM	Multiclass Brier	0.3402 (0.2927 to 0.3884)
Matched six-predictor CLM	Mean outcome-pair ranking AUC	0.7539 (0.6968 to 0.8074)
Five-predictor CLM	Cumulative mAUC	0.7544 (0.6935 to 0.8110)
Five-predictor CLM	Generalised C	0.7169 (0.6650 to 0.7695)
Five-predictor CLM	AUC Y≥1	0.7178 (0.6638 to 0.7743)
Five-predictor CLM	AUC Y=2	0.7909 (0.7055 to 0.8662)
Five-predictor CLM	Multiclass Brier	0.3537 (0.3054 to 0.4040)
Five-predictor CLM	Mean outcome-pair ranking AUC	0.7213 (0.6623 to 0.7779)

The matched six- and five-predictor CLMs share categorical coding, joint ordinal likelihood, folds and imputations. The earlier five-fold point-score reference uses its original integer-score and probability-map pipeline; it differs from both CLMs and from the current ten-repeat SF_align assessment.

Part B Paired removal contrasts

Difference direction	Measure	Difference (95% CI)
Five minus matched six	Cumulative mAUC	−0.0388 (−0.0729 to −0.0039)
Five minus matched six	Generalised C	−0.0460 (−0.0808 to −0.0130)
Five minus matched six	AUC Y≥1	−0.0500 (−0.0874 to −0.0143)
Five minus matched six	AUC Y=2	−0.0275 (−0.0691 to 0.0166)
Five minus matched six	Multiclass Brier	0.0135 (−0.0000 to 0.0275)
Five minus matched six	Mean outcome-pair ranking AUC	−0.0326 (−0.0627 to −0.0019)
Five minus earlier point-score	Cumulative mAUC	−0.0503 (−0.0860 to −0.0150)
Five minus earlier point-score	Generalised C	−0.0598 (−0.0954 to −0.0247)
Five minus earlier point-score	AUC Y≥1	−0.0622 (−0.0999 to −0.0261)
Five minus earlier point-score	AUC Y=2	−0.0384 (−0.0830 to 0.0063)
Five minus earlier point-score	Multiclass Brier	0.0214 (0.0082 to 0.0358)
Five minus earlier point-score	Mean outcome-pair ranking AUC	−0.0447 (−0.0773 to −0.0130)

For the five-predictor model minus its matched six-predictor reference, the ΔBrier interval is −0.0000123 to 0.0275476 and includes zero. The severe-outcome AUC difference interval also includes zero.

Supplementary Table 13 Model structure and alternative predictors

Part A Cumulative logit structure

Specification	Cumulative mAUC	Multiclass Brier	ΔmAUC (95% CI)	Δ multiclass Brier (95% CI)	Mean OVR slope
Six-predictor shared slope	0.8082	0.3294	Reference	Reference	0.8971
Age-specific slopes	0.8109	0.3260	0.0027 (−0.0039 to 0.0104)	−0.0033 (−0.0094 to 0.0026)	0.9125
All cut-point-specific slopes	0.8113	0.3311	0.0031 (−0.0046 to 0.0115)	0.0018 (−0.0076 to 0.0111)	0.7537

All specifications use numerical-grade coding and the stacked cumulative-binomial likelihood. Differences are calculated as the alternative minus the shared-slope reference, with intervals from 500 paired bootstrap resamples. The final column averages three class-specific OVR calibration slopes.

Part B EF substitutions

Specification	Cumulative mAUC	Multiclass Brier	ΔmAUC (95% CI)	Δ multiclass Brier (95% CI)
Six-predictor shared slope	0.8082	0.3294	Reference	Reference
EF replacing age	0.8053	0.3305	−0.0029 (−0.0172 to 0.0121)	0.0011 (−0.0072 to 0.0087)

Specification	Cumulative mAUC	Multiclass Brier	ΔmAUC (95% CI)	Δ multiclass Brier (95% CI)
EF replacing eGFR	0.7954	0.3306	−0.0127 (−0.0313 to 0.0062)	0.0012 (−0.0082 to 0.0104)

EF was missing in 60/423 patients (14.2%). Numerical coding, likelihood, folds and imputation were matched within this comparison.

Part C IBG removal and single-marker substitutions

Specification	Cumulative mAUC	Multiclass Brier	ΔmAUC (95% CI)	Δ multiclass Brier (95% CI)
Six-predictor shared slope	0.8082	0.3294	Reference	Reference
IBG removed	0.7967	0.3384	−0.0115 (−0.0265 to 0.0029)	0.0090 (0.0018 to 0.0168)
CRP replacing IBG	0.8150	0.3245	0.0068 (−0.0069 to 0.0210)	−0.0049 (−0.0126 to 0.0028)
NLR replacing IBG	0.8075	0.3341	−0.0007 (−0.0143 to 0.0134)	0.0047 (−0.0039 to 0.0137)
ESR replacing IBG	0.8044	0.3371	−0.0037 (−0.0190 to 0.0127)	0.0077 (−0.0021 to 0.0171)
WBC replacing IBG	0.8049	0.3309	−0.0033 (−0.0178 to 0.0115)	0.0015 (−0.0063 to 0.0084)

Numerical coding and the stacked cumulative-binomial likelihood were matched. Removal of IBG increased Brier error with a difference interval above zero; mAUC and Brier difference intervals for all single-marker substitutions included zero. Continuous CRP substitution in this regression analysis is distinct from the graded CRP integer-scoring comparison in Supplementary Tables 8, 9 and Supplementary Data (source table 57), which uses categorical coding, training-refitted integer points and aligned probability maps.

Supplementary Table 14 Sensitivity to heart failure coding, PCT omission, and outcome confirmation

Part A Heart failure coding

Measure	Primary scenario B	Scenario A: 18 uncertain records assigned HF=0	Δ(A−B)
mAUC	0.8047	0.8064	+0.0017
Generalised C-index	0.7766	0.7762	−0.0005
QWK	0.4598	0.4306	−0.0292
Multiclass Brier	0.3323	0.3321	−0.0002

This sensitivity analysis uses the earlier single five-fold OOF assessment. Scenario B was the primary analysis: heart failure status for the 18 records coded as "0, echocardiography not performed" was treated as missing and imputed within each training fold using the prespecified M=20 procedure. Scenario A assigned these 18 records to HF=0. The comparison quantifies sensitivity to this coding decision; differences are reported as A−B.

Part B Simulated PCT omission in previously tested patients

Rule	IBG agreement (n=307)	Risk-stratum migration (n=301)	High → intermediate (n=301)
C=1 (exploratory fallback)	215/307 (70.0%)	8/301 (2.7%)	1
C=0 (candidate comparator)	160/307 (52.1%)	27/301 (9.0%)	4

Rule	IBG agreement (n=307)	Risk-stratum migration (n=301)	High → intermediate (n=301)
C≈observed mean (=1)	215/307 (70.0%)	8/301 (2.7%)	1
C=round(mean(A,B))	153/307 (49.8%)	16/301 (5.3%)	0

The primary analysis reconstructed IBG after within-fold imputation. In the updated omission sensitivity, observed PCT was masked in 307 patients with all five inflammatory markers measured, and reconstructed IBG grades were compared with their complete-marker grades. Risk-stratum migration used the 301 patients who also had all five other point-card items available, using the development-derived card and 24/34 boundaries. The earlier 310/304-patient analysis allowed partially observed A or B domains and is retained only as labelled historical data. Separately, 113 patients had genuinely unmeasured PCT; 112 had at least one available marker in each of A and B, and one did not. The table describes simulated omission among fully measured reference patients. The current locked complete-input policy returns no risk output when a required input is missing. PCT-testing and six-predictor input availability by outcome are reported in Part D.

Part C Healing confirmed by telephone alone

Analysis	n	Cumulative mAUC	Generalised C	Multiclass Brier	AUC Y≥1	AUC Y=2
Earlier five-fold OOF, all 423	423	0.8047	0.7766	0.3323	0.7800	0.8293
Earlier five-fold OOF, excluding 3	420	0.8070	0.7795	0.3323	0.7829	0.8310
Shared-slope reference, all 423	423	0.8082	0.7795	0.3294	0.7855	0.8308
Shared-slope reference, excluding 3	420	0.8104	0.7820	0.3287	0.7882	0.8326

This sensitivity analysis uses the earlier single five-fold OOF assessment. Saved OOF predictions were retained; no model was refitted. This sensitivity concerns the three telephone-only confirmations after the completed outcome-source review.

Part D PCT testing and required-input availability by outcome

Outcome	n	PCT measured	PCT unmeasured	All six-predictor required inputs available	Required inputs incomplete
Healing	319	226 (70.8%)	93 (29.2%)	222 (69.6%)	97 (30.4%)
Persistent non-healing	61	50 (82.0%)	11 (18.0%)	45 (73.8%)	16 (26.2%)
Below-knee amputation or death	43	34 (79.1%)	9 (20.9%)	34 (79.1%)	9 (20.9%)
Total	423	310 (73.3%)	113 (26.7%)	301 (71.2%)	122 (28.8%)

Percentages use the outcome-row denominator. Six-predictor availability requires the five non-IBG predictors and all five component markers used to reconstruct IBG (WBC, NLR, ESR, CRP and PCT) to be observed. These counts summarise observed input availability. Locked complete-input mapping uses the 301 complete records, whereas fold-refitted internal assessment uses imputation in all 423 patients.

Supplementary Table 15 Same-centre temporal evaluation

Part A Development and evaluation periods

Dataset	Calendar period	n	Outcome Y=0/Y=1/Y=2
Development	2020–2024	302	228 / 44 / 30
Later evaluation	2025–2026	121	91 / 17 / 13

Predictor identities and some upstream specifications originated from the overall development process; imputers, point values, probability mappings and cut-offs were fitted to 2020–2024 data and then applied to the 2025–2026 evaluation period.

Part B Performance in the later cohort

Measure	Estimate	95% CI
mAUC	0.7810	0.6726–0.8665
Generalised C-index	0.7265	—
QWK	0.4457	—
Multiclass Brier	0.3471	—
Overall accuracy	0.6446	—

The 95% CI for mAUC was estimated from 500 bootstrap resamples. This was a same-centre temporal sensitivity analysis; intervals were not estimated for the remaining measures.

Part C Risk strata in the later cohort

Risk group	n	Any adverse outcome (95% CI)	Severe outcome (95% CI)
Low	85	17.6% (11.0–27.1)	4.7% (1.8–11.5)
Intermediate	16	25.0% (10.2–49.5)	12.5% (3.5–36.0)
High	20	55.0% (34.2–74.2)	35.0% (18.1–56.7)

Development-period cut-offs were approximately 26.9 and 33.0. Estimates for the intermediate- and high-risk strata were imprecise because these groups were small.

Supplementary Table 16 Sensitivity of whole-cohort point cards to imputation pooling

Part A M1 and pooled M20 points

Predictor	Level	Current M=1	M=20 pooled	Difference	M20 median (IQR)
Age	≥65 years	3	3	0	3 (3–3)
Heart failure	1	3	3	0	3 (3–3)
Heart failure	2	16	16	0	16 (16–16)
Palpated foot temperature	1 (cold)	16	16	0	16 (15–16)
Palpated foot temperature	2 (cool)	13	13	0	13 (13–13)
Palpated foot temperature	3 (warm)	8	7	−1	8 (8–8)
eGFR	2	0	1	+1	1 (0–1)
eGFR	3	1	2	+1	2 (2–2)
eGFR	4	1	2	+1	2 (2–2)
eGFR	5	8	8	0	8 (8–8.25)
Wound area	5–20 cm²	3	3	0	3 (3–3)

Predictor	Level	Current M=1	M=20 pooled	Difference	M20 median (IQR)
Wound area	20<A≤100 cm²	9	9	0	9 (9–9)
Wound area	>100 cm²	10	10	0	10 (10–11)
IBG	1	3	3	0	3 (3–4)
IBG	2	6	6	0	6 (6–6)
IBG	3	6	6	0	6 (6–6)

Reference levels remained at 0 points. Original eGFR 1–5 and IBG 0–3 levels are shown separately to display imputation-related variation; the formal 0–59-point table combines equal-point categories. Continuous coefficients were pooled using Rubin's rules before conversion to integer points; the 20 integer point tables were not averaged. All 20 models converged, no risk-direction reversals occurred, and PAV did not further alter the pooled points.

Part B Conditional predictions using whole-cohort point cards

Metric	M=1 fixed table	M=20 pooled table	Δ (M20−M1)	95% CI for Δ
mAUC	0.8320	0.8316	−0.0004	−0.0048–0.0040
Generalised C-index	0.8091	0.8091	−0.0001	−0.0042–0.0044
Multiclass Brier	0.3161	0.3164	+0.0003	−0.0020–0.0027
AUC (Y≥1)	0.8143	0.8147	+0.0004	−0.0039–0.0048
PR-AUC (Y≥1)	0.6391	0.6393	+0.0002	−0.0094–0.0098
Calibration slope (Y≥1)	0.9786	0.9799	+0.0013	−0.0252–0.0229
AUC (Y=2)	0.8496	0.8484	−0.0012	−0.0076–0.0047
PR-AUC (Y=2)	0.5413	0.5359	−0.0054	−0.0367–0.0216
Calibration slope (Y=2)	0.9584	0.9627	+0.0043	−0.0230–0.0290

Both point tables were derived from the whole development cohort; patient scores were averaged across the 20 imputations, and probability mappings were fitted only in training folds of the prespecified five-fold split. All 95% CIs for the listed differences included zero. These estimates quantify imputation sensitivity conditional on whole-cohort point derivation. The earlier five-fold OOF estimate with fold-refitted point tables was 0.8047; the current ten-repeat framework assessment is reported in Supplementary Table 5.

Supplementary Table 17 Model complexity and automated variable selection

Part A Development comparisons

Development stage	Specification	AutoScore-Ordinal mAUC	AutoScore-Custom mAUC
Model complexity	11-predictor reference	0.7723	0.7726
Model complexity	7-predictor candidate	0.7932	0.7758
Model complexity	6-predictor candidate (K=6)	0.7980	0.7824

Development stage	Specification	AutoScore-Ordinal mAUC	AutoScore-Custom mAUC
Fixed six-predictor score	Baseline target: 100 points	0.7982	0.7837
Fixed six-predictor score	Target changed to 60 points	0.7971	0.7855
Fixed six-predictor score	Age≥65 years + 100-point target	—	0.8030
Fixed six-predictor score	Age≥65 years + 60-point target (selected)	—	0.8047

The formal complexity comparison used constrained M=20 analyses and reported the six-, seven-, and 11-predictor specifications from the final analysis configuration; score-specification comparisons held the six predictors fixed. The values 0.7980 and 0.7982 arose from separate development runs, and the 65-year age cut-off and 60-point target were selected during development. An earlier independent V6 exploratory run compared 5–8-predictor candidates under a different analysis configuration: AutoScore-Ordinal mAUCs were 0.7380, 0.7890, 0.7970, and 0.7926, respectively, and AutoScore-Custom mAUCs were 0.7182, 0.7673, 0.7809, and 0.7621; the corresponding 11-predictor baselines were 0.7594 and 0.7671. This earlier run served to identify a parsimonious candidate range and was not combined with the formal constrained 6/7/11 comparison because the analysis configurations differed.

Part B Selection in the specified automated workflow

Predictor	Selected outer refits	Proportion
Palpated foot-temperature grade	50/50	100%
Heart failure grade	50/50	100%
IBG	45/50	90%
Wound-area grade	32/50	64%
Ejection-fraction grade	31/50	62%
eGFR grade	24/50	48%
Natriuretic peptide precursor	20/50	40%
Popliteal-artery occlusion	11/50	22%
Age	9/50	18%
Diabetic nephropathy	8/50	16%
Random glucose	6/50	12%
Platelet count	4/50	8%
Wagner grade	4/50	8%
Haemoglobin	3/50	6%
Serum troponin I	2/50	4%
Total-cholesterol grade	1/50	2%
Glycated haemoglobin	1/50	2%

All non-zero selection frequencies across 50 outer-refit models are shown, with inner-selected configurations and training-local IBG. Supplementary Data (source table 6) reports the earlier fixed-K = 6 procedure. Frequencies describe algorithmic selection across overlapping partitions.

Supplementary Table 18 Scenarios for additional preoperative cardiorenal review

Part A Threshold-positive patients and severe-outcome identification

Procedure	Threshold %	Threshold-positive / 100	Sensitivity %	PPV %	Net benefit / 100
Four-predictor reference / required	5	53.8 (49.8 to 58.1)	86.5 (76.1 to 95.2)	16.3 (11.7 to 21.0)	6.42 (3.79 to 9.13)
Four-predictor reference / required	10	26.5 (22.8 to 30.3)	66.3 (54.1 to 79.1)	25.4 (18.2 to 33.3)	4.54 (2.34 to 7.01)
Four-predictor reference / required	15	15.0 (11.8 to 18.4)	48.8 (34.9 to 63.0)	33.1 (22.5 to 44.5)	3.20 (1.21 to 5.45)
Four-predictor reference / required	20	11.4 (8.5 to 14.4)	41.2 (27.4 to 55.4)	36.8 (23.9 to 50.1)	2.39 (0.51 to 4.52)
six-predictor IBG / required	5	51.0 (46.6 to 55.2)	90.0 (80.9 to 97.3)	17.9 (13.0 to 22.9)	6.95 (4.24 to 9.79)
six-predictor IBG / required	10	29.1 (25.3 to 33.2)	73.3 (61.3 to 84.7)	25.6 (18.6 to 33.1)	5.04 (2.71 to 7.66)
six-predictor IBG / required	15	18.3 (15.0 to 21.9)	59.1 (44.2 to 73.0)	32.8 (22.5 to 43.0)	3.83 (1.61 to 6.25)
six-predictor IBG / required	20	12.7 (9.9 to 15.7)	50.7 (36.8 to 64.1)	40.4 (28.3 to 52.7)	3.26 (1.22 to 5.46)
six-predictor CRP / required	5	49.8 (45.6 to 54.2)	89.8 (81.1 to 97.1)	18.3 (13.2 to 23.4)	6.98 (4.27 to 9.85)
six-predictor CRP / required	10	29.7 (25.8 to 33.9)	74.2 (62.2 to 85.7)	25.4 (18.2 to 32.9)	5.08 (2.65 to 7.68)
six-predictor CRP / required	15	19.1 (15.7 to 22.6)	60.5 (46.2 to 74.5)	32.2 (22.5 to 42.0)	3.86 (1.62 to 6.27)
six-predictor CRP / required	20	13.2 (10.2 to 16.2)	52.3 (37.8 to 66.4)	40.3 (27.8 to 52.4)	3.35 (1.24 to 5.57)
six-predictor IBG / common	5	51.1 (46.7 to 55.3)	90.7 (81.8 to 97.8)	18.0 (13.1 to 23.0)	7.02 (4.29 to 9.87)
six-predictor IBG / common	10	29.0 (25.2 to 33.1)	71.6 (59.6 to 83.1)	25.1 (18.1 to 32.6)	4.87 (2.52 to 7.46)
six-predictor IBG / common	15	18.3 (15.0 to 21.7)	58.8 (44.0 to 72.2)	32.8 (22.5 to 43.0)	3.82 (1.60 to 6.25)
six-predictor IBG / common	20	12.6 (9.7 to 15.6)	50.0 (35.9 to 63.6)	40.3 (28.0 to 52.5)	3.20 (1.19 to 5.37)

Counts are means over ten evaluations of the same 423 patients. The number of threshold-positive patients per 100 represents potential review volume. Net benefit per 100 is the threshold-weighted balance of true- and false-positive classifications under the strategy defined in Supplementary Methods.

Part B Paired net benefit at the prespecified scenario thresholds

Six-predictor procedure	Threshold %	Change per 100 (95% interval)
six-predictor IBG / required	5	0.52 (-0.27 to 1.32)
six-predictor IBG / required	10	0.50 (-0.37 to 1.41)
six-predictor IBG / required	15	0.64 (-0.36 to 1.75)
six-predictor IBG / required	20	0.87 (0.08 to 1.77)
six-predictor IBG / common	5	0.59 (-0.23 to 1.41)
six-predictor IBG / common	10	0.33 (-0.56 to 1.22)
six-predictor IBG / common	15	0.62 (-0.34 to 1.68)
six-predictor IBG / common	20	0.82 (0.04 to 1.65)
six-predictor CRP / required	5	0.56 (-0.22 to 1.28)

Six-predictor procedure	Threshold %	Change per 100 (95% interval)
six-predictor CRP / required	10	0.54 (-0.53 to 1.63)
six-predictor CRP / required	15	0.67 (-0.45 to 1.89)
six-predictor CRP / required	20	0.96 (-0.09 to 2.10)

Differences are six-predictor IBG minus four-predictor reference, per 100 patients, at the four prespecified thresholds. Intervals are exploratory and unadjusted for multiplicity. Supplementary Data (source table 61) provides the complete 5–40% grid, including intermediate thresholds and all comparisons.

Supplementary Table 19 Locked scoring specifications for complete inputs

Part A Cumulative probability mappings

Scoring specification	Complete n	a0	a1	b0	b1
Fixed six-predictor IBG	301	-5.08709090	0.17190274	-6.94376759	0.18545075
Fixed four-predictor reference	405	-3.23041220	0.10992344	-4.70552650	0.12008210

For fixed total score s, P(Y≥1)=expit(a0+a1s) and P(Y=2)=expit(b0+b1s), where expit(z)=1/(1+exp(−z)). P(Y=0)=1−P(Y≥1) and P(Y=1)=P(Y≥1)−P(Y=2). The two maps do not cross within their score domains. Six-predictor IBG score range is 0–59; four-predictor reference range is 0–60. Coefficients are shown to eight decimal places; machine-readable specifications retain full precision. These maps were applied unchanged in the three-centre external validation (Tables 20–26). Maps were fitted in 301 six-predictor complete-input patients and 405 four-predictor complete-input patients. They require their defined inputs to be complete; no risk estimate is returned if required inputs are missing. Internal predictions for all 423 patients used within-fold imputation and refitted points/maps, as specified in Table 4.

Part B Point card for the four-predictor reference

Variable	Category	Points
Age	<65 years	0
Age	≥65 years	8
Heart failure grade	0	0
Heart failure grade	1	5
Heart failure grade	2	25
eGFR grade	1	0
eGFR grade	2	1
eGFR grade	3	2
eGFR grade	4	2
eGFR grade	5	15
Wound-area grade	1	0
Wound-area grade	2	4
Wound-area grade	3	12
Wound-area grade	4	12

The comparator was derived with the same scoring algorithm and a 60-point normalisation target. The age boundary is ≥65 years; heart-failure, eGFR and wound-area definitions match the main Methods and the candidate-input dictionary in Supplementary Data. This is a study-specific clinical reference. The six-predictor points remain those in the main score table.

Part C Preoperative input definitions and extraction

Input	Preoperative extraction and coding	Locked-score requirement
Age	Recorded age at first TTT; <65 or ≥65 years.	Observed age required.
Heart failure	Latest preoperative cardiology diagnosis together with preoperative echocardiography: 0 none, 1 chronic, 2 acute.	Use the documented clinical state.
Palpated foot temperature	Latest preoperative palpation: 1 cold, 2 cool, 3 warm, 4 warmer. Development records retained the two ward physicians' consensus.	Use the recorded grade; independent observer ratings were not retained.
Renal function	Standardised eGFR, mL/min/1.73 m²: 1 ≥90; 2 60≤eGFR<90; 3 30≤eGFR<60; 4 15≤eGFR<30; 5 <15 or documented uraemia.	Use the laboratory report. Documented uraemia retains grade 5; numerical-only sensitivity is separate.
Wound area	Maximum length × maximum width; sum multiple wounds. Grades 1–4: A<5, 5≤A≤20, 20<A≤100 and A>100 cm².	Area is not rounded before categorization. Analyses retain the recorded grades.
IBG	Latest preoperative WBC, NLR, ESR, CRP and PCT; domain maxima and cut-offs in Table 3.	All components needed to calculate the observed grade.
CRP supplementary version	CRP <10, 10 to <100, or ≥100 mg/L replaces IBG; the other five inputs are retained.	Use only when all six inputs for that version are observed.
Time and missing inputs	All information available before first TTT; the latest preoperative result is selected.	No risk estimate is returned if a required input is unavailable.

Supplementary Table 20 External cohorts, outcomes, and input availability

Centre	Eligible n	Healed	Persistently unhealed	Severe	Amputation / death	Severe-outcome rate (95% CI)
Shanxi	94	62	26	6	5 / 1	6.4% (2.4 to 13.4)
Xiamen	81	63	12	6	1 / 5	7.4% (2.8 to 15.4)
Hunan	38	38	0	0	0 / 0	0.0% (0.0 to 9.3)
Pooled	213	163	38	12	6 / 6	5.6% (2.9 to 9.6)

Centre	Input version	Scorable n / eligible n	Coverage	Severe events represented
Shanxi	IBG	94/94	100.0%	6/6
Shanxi	CRP	94/94	100.0%	6/6
Shanxi	Availability pathway	94/94	100.0%	6/6
Xiamen	IBG	73/81	90.1%	6/6
Xiamen	CRP	79/81	97.5%	6/6
Xiamen	Availability pathway	79/81	97.5%	6/6
Hunan	IBG	38/38	100.0%	0/0
Hunan	CRP	38/38	100.0%	0/0
Hunan	Availability pathway	38/38	100.0%	0/0
Pooled	IBG	205/213	96.2%	12/12
Pooled	CRP	211/213	99.1%	12/12
Pooled	Availability pathway	211/213	99.1%	12/12

Six Shanxi Wagner grade 2 patients were excluded before performance evaluation. Hunan contributed no severe events. All 12 external severe events had complete IBG and CRP inputs. The eight patients without complete IBG inputs were in Xiamen; CRP covered six of these. Source-file checksums and the final data version are archived.

Observed case mix is summarised below; the expanded four-cohort clinical comparison is in main Table 1. Categorical denominators use the relevant non-missing observations. Full distributions are provided in the case-mix and missingness source tables.

Characteristic	Development n=423	Shanxi n=94	Xiamen n=81	Hunan n=38
Age years median IQR	64 (57 to 71.5)	68 (64 to 72)	62 (55 to 70)	67 (59.25 to 72)
Heart failure grade 1 or 2	95/405 (23.5%)	10/94 (10.6%)	10/81 (12.3%)	1/38 (2.6%)
Acute heart failure grade 2	28/405 (6.9%)	0/94 (0.0%)	5/81 (6.2%)	0/38 (0.0%)
Renal grade 4 or 5	78/423 (18.4%)	10/94 (10.6%)	18/81 (22.2%)	7/38 (18.4%)
Renal grade 5	47/423 (11.1%)	10/94 (10.6%)	13/81 (16.0%)	7/38 (18.4%)
Wound area grade 3 or 4	230/423 (54.4%)	57/94 (60.6%)	45/81 (55.6%)	16/38 (42.1%)
Clinical temperature grade 1 or 2	132/423 (31.2%)	78/94 (83.0%)	31/81 (38.3%)	18/38 (47.4%)
IBG grade 2 or 3	112/307 (36.5%)	21/94 (22.3%)	33/74 (44.6%)	6/38 (15.8%)
Wagner grade 3	93/423 (22.0%)	65/94 (69.1%)	27/81 (33.3%)	2/38 (5.3%)
Wagner grade 4	296/423 (70.0%)	26/94 (27.7%)	46/81 (56.8%)	36/38 (94.7%)
Wagner grade 5	34/423 (8.0%)	3/94 (3.2%)	8/81 (9.9%)	0/38 (0.0%)

Supplementary Table 21 Locked model performance and paired comparisons

Centre	Model	n	Any-adverse-outcome AUC (95% CI)	Severe-outcome AUC (95% CI)	Severe-outcome Brier (95% CI)	Multiclass Brier (95% CI)
Shanxi	IBG	94	0.984 (0.962 to 0.998)	0.937 (0.845 to 1.000)	0.0727 (0.0475–0.0987)	0.3123 (0.2429–0.3815)

Centre	Model	n	Any-adverse-outcome AUC (95% CI)	Severe-outcome AUC (95% CI)	Severe-outcome Brier (95% CI)	Multiclass Brier (95% CI)
Shanxi	CRP	94	0.962 (0.921 to 0.993)	0.903 (0.729 to 1.000)	0.0570 (0.0370–0.0808)	0.3329 (0.2651–0.3986)
Shanxi	Four-predictor reference	94	0.970 (0.937 to 0.994)	0.899 (0.726 to 1.000)	0.0425 (0.0208–0.0699)	0.3633 (0.2708–0.4591)
Shanxi	Availability pathway	94	0.984 (0.962 to 0.998)	0.937 (0.845 to 1.000)	0.0727 (0.0475–0.0987)	0.3123 (0.2429–0.3815)
Xiamen	IBG	73	0.867 (0.767 to 0.949)	0.867 (0.732 to 0.967)	0.0795 (0.0370–0.1257)	0.3062 (0.2005–0.4176)
Xiamen	CRP	79	0.808 (0.692 to 0.909)	0.773 (0.530 to 0.960)	0.0806 (0.0392–0.1264)	0.3362 (0.2259–0.4510)
Xiamen	Four-predictor reference	80	0.711 (0.574 to 0.839)	0.845 (0.705 to 0.972)	0.0665 (0.0286–0.1077)	0.3466 (0.2338–0.4658)
Xiamen	Availability pathway	79	0.851 (0.746 to 0.937)	0.869 (0.748 to 0.964)	0.0766 (0.0380–0.1210)	0.3159 (0.2135–0.4254)
Hunan	IBG	38	NE	NE	0.0081 (0.0036–0.0142)	0.1022 (0.0611–0.1501)
Hunan	CRP	38	NE	NE	0.0087 (0.0035–0.0155)	0.0882 (0.0487–0.1334)
Hunan	Four-predictor reference	38	NE	NE	0.0108 (0.0048–0.0199)	0.0929 (0.0576–0.1358)
Hunan	Availability pathway	38	NE	NE	0.0081 (0.0036–0.0142)	0.1022 (0.0611–0.1501)
Pooled	IBG	205	0.936 (0.897 to 0.971)	0.895 (0.803 to 0.963)	0.0632 (0.0432–0.0833)	0.2712 (0.2214–0.3207)
Pooled	CRP	211	0.893 (0.838 to 0.941)	0.841 (0.685 to 0.966)	0.0572 (0.0390–0.0784)	0.2900 (0.2404–0.3456)
Pooled	Four-predictor reference	212	0.840 (0.777 to 0.898)	0.865 (0.755 to 0.956)	0.0459 (0.0287–0.0655)	0.3085 (0.2491–0.3695)
Pooled	Availability pathway	211	0.925 (0.878 to 0.964)	0.896 (0.806 to 0.962)	0.0625 (0.0437–0.0832)	0.2758 (0.2281–0.3312)

Populations with available inputs differ; paired comparisons use the same 205 patients and 12 severe events. Positive Brier differences indicate greater error for IBG. The locked IBG model had lower multiclass Brier but higher severe-outcome Brier than the four-predictor clinical reference. NE means not estimable. The availability pathway uses IBG when its inputs are complete and otherwise uses CRP when its inputs are complete. Updating results are reported separately in Table 23.

Paired contrast	Metric	IBG	Four-predictor reference	Difference (95% CI)
IBG minus reference	AUC: any adverse outcome	0.9361	0.8450	0.0911 (0.0495 to 0.1367)
IBG minus reference	AUC: severe outcome	0.8949	0.8620	0.0328 (-0.0196 to 0.0963)
IBG minus reference	mAUC	0.9155	0.8535	0.0620 (0.0238 to 0.1070)
IBG minus reference	Severe-outcome Brier	0.0632	0.0472	0.0160 (0.0019 to 0.0292)
IBG minus reference	Multiclass Brier	0.2712	0.3065	-0.0353 (-0.0677 to -0.0057)
IBG minus reference	Log loss	0.4930	0.5468	-0.0538 (-0.0998 to -0.0101)

Class-specific external performance

180-day outcome	Events / n	AUC (95% CI)	Binary Brier (95% CI)	Observed / predicted	Calibration slope (95% CI)
Healed	157/205	0.936 (0.897 to 0.971)	0.0918 (0.0722 to 0.1129)	76.6% / 68.8%	1.612 (1.177 to 2.159)
Persistent non-healing	36/205	0.855 (0.794 to 0.910)	0.1162 (0.0916 to 0.1428)	17.6% / 16.3%	2.291 (1.475 to 3.325)
Severe	12/205	0.895 (0.803 to 0.963)	0.0632 (0.0432 to 0.0833)	5.9% / 15.0%	0.976 (0.574 to 1.478)

All rows use the 205 complete-input external patients and unchanged IBG predictions. Each outcome is evaluated against the other two. The three binary Brier scores sum to the multiclass Brier. Class-specific calibration slopes are estimated jointly with an intercept; a finite slope of 1 is ideal. Cumulative calibration is reported in Table 22.

Within-centre discrimination

Model	Outcome	Within-centre AUC (95% CI)
six-predictor IBG	Any adverse outcome	0.947 (0.914–0.977)
six-predictor IBG	Severe	0.906 (0.828–0.969)
six-predictor CRP; supplementary	Any adverse outcome	0.907 (0.858–0.949)
six-predictor CRP; supplementary	Severe	0.844 (0.700–0.963)
Four-predictor reference	Any adverse outcome	0.877 (0.821–0.929)
Four-predictor reference	Severe	0.874 (0.765–0.960)
Availability route	Any adverse outcome	0.937 (0.895–0.971)
Availability route	Severe	0.906 (0.827–0.968)

Concordance retains only event/non-event pairs from the same centre and weights each centre by its number of such pairs. Centre-stratified patient-bootstrap intervals use 2,000 resamples. Analysis populations correspond to the available-input rows above; Hunan contributes no event/non-event pairs. This analysis avoids cross-centre pairs while retaining each model's observed input availability.

Supplementary Table 22 Cumulative calibration of the locked IBG model

Centre / target	n / events	Observed / predicted	O/E 95% CI	Intercept (95% CI)	Slope (95% CI)
Shanxi / any	94/32	34.0% / 38.9%	0.875 (0.599 to 1.236)	-0.326 (-0.865 to 0.198)	4.277 (2.461 to 7.227)
Shanxi / severe	94/6	6.4% / 19.9%	0.321 (0.118 to 0.698)	-1.739 (-2.750 to -0.899)	1.693 (0.775 to 3.154)
Xiamen / any	73/16	21.9% / 27.4%	0.801 (0.458 to 1.301)	-0.473 (-1.205 to 0.205)	0.823 (0.433 to 1.320)
Xiamen / severe	73/6	8.2% / 13.2%	0.623 (0.228 to 1.355)	-0.844 (-1.970 to 0.121)	0.663 (0.226 to 1.210)
Hunan / any	38/0	0.0% / 19.7%	0.000 (0.000 to 0.492)	NE	NE
Hunan / severe	38/0	0.0% / 6.1%	0.000 (0.000 to 1.602)	NE	NE
Pooled / any	205/48	23.4% / 31.2%	0.750 (0.553 to 0.994)	-0.611 (-1.018 to -0.219)	1.612 (1.177 to 2.159)
Pooled / severe	205/12	5.9% / 15.0%	0.391 (0.202 to 0.684)	-1.469 (-2.181 to -0.840)	0.976 (0.574 to 1.478)

Calibration-in-the-large intercepts were estimated with the locked logit prediction as an offset and slope fixed at 1. Free calibration slopes were estimated jointly with an intercept. In Hunan, AUC and finite calibration intercepts and slopes were not estimated because no adverse outcome occurred.

Supplementary Table 23 Supplementary baseline-risk updating

Stage	Model	Severe AUC	Severe-outcome Brier (95% CI)	Multiclass Brier (95% CI)	Mean severe risk
Locked model	IBG	0.895	0.0632 (0.0432–0.0833)	0.2712 (0.2214–0.3207)	15.0%
Locked model	CRP	0.840	0.0577 (0.0386–0.0783)	0.2858 (0.2339–0.3387)	13.6%
Locked model	Four-predictor reference	0.862	0.0472 (0.0295–0.0681)	0.3065 (0.2456–0.3682)	10.1%
Updating-step CV	IBG	0.892	0.0439 (0.0251–0.0672)	0.2278 (0.1781–0.2787)	5.9%
Updating-step CV	CRP	0.837	0.0468 (0.0265–0.0715)	0.2631 (0.2107–0.3172)	5.9%
Updating-step CV	Four-predictor reference	0.859	0.0460 (0.0257–0.0707)	0.3010 (0.2432–0.3631)	5.9%
Leave-one-centre-out updating	IBG	0.876	0.0443 (0.0253–0.0675)	0.2533 (0.1992–0.3101)	6.5%
Leave-one-centre-out updating	CRP	0.833	0.0461 (0.0260–0.0706)	0.2856 (0.2251–0.3455)	5.9%
Leave-one-centre-out updating	Four-predictor reference	0.846	0.0461 (0.0255–0.0716)	0.3449 (0.2758–0.4157)	5.5%

Stage / contrast	Metric	Difference (95% CI)
Locked: IBG minus reference	Severe-outcome Brier	0.0160 (0.0019 to 0.0292)
Locked: IBG minus reference	Multiclass Brier	-0.0353 (-0.0677 to -0.0057)
Updating-step CV: IBG minus reference	Severe-outcome Brier	-0.0022 (-0.0085 to 0.0039)
Updating-step CV: IBG minus reference	Multiclass Brier	-0.0732 (-0.0987 to -0.0498)
Leave-one-centre-out updating: IBG minus reference	Severe-outcome Brier	-0.0019 (-0.0103 to 0.0054)
Leave-one-centre-out updating: IBG minus reference	Multiclass Brier	-0.0916 (-0.1196 to -0.0651)
Updating-step CV minus locked IBG	Severe-outcome Brier	-0.0193 (-0.0348 to -0.0031)
Updating-step CV minus locked IBG	Multiclass Brier	-0.0434 (-0.0700 to -0.0144)

Held-out centre	Train / test n	Test events	Predicted / observed severe risk	Severe Brier
Shanxi	111/94	6	10.2% / 6.4%	0.0404
Xiamen	132/73	6	4.3% / 8.2%	0.0719
Hunan	167/38	0	1.7% / 0.0%	0.0007

All rows above use 205 identical patients. CV evaluates the updating procedure using 20 repeated five-fold partitions; LOCO trains offsets in other centres. Intervals for paired differences use patient bootstrap resampling conditional on saved predictions. Full-cohort apparent fit is archived separately as candidate-model development. Bootstrap optimism checks and the availability-route update are provided in the external analysis workbook and CSV tables.

Supplementary Table 24 Additional cardiorenal review scenarios in external patients

Model	Threshold	Review n/205	Detected /12	Missed	Sensitivity % (95% CI)	PPV % (95% CI)	Net benefit /100 (95% CI)
six-predictor IBG	5%	110	12	0	100.0 (73.5–100.0)	10.9 (5.8–18.3)	3.34 (0.28–7.14)
six-predictor IBG	10%	77	11	1	91.7 (61.5–99.8)	14.3 (7.4–24.1)	1.79 (-1.14–5.42)

Model	Threshold	Review n/205	Detected /12	Missed	Sensitivity % (95% CI)	PPV % (95% CI)	Net benefit /100 (95% CI)
six-predictor IBG	15%	60	11	1	91.7 (61.5–99.8)	18.3 (9.5–30.4)	1.15 (-1.92–4.85)
six-predictor IBG	20%	51	11	1	91.7 (61.5–99.8)	21.6 (11.3–35.3)	0.49 (-2.80–4.39)
Four-predictor reference	5%	95	12	0	100.0 (73.5–100.0)	12.6 (6.7–21.0)	3.72 (0.67–7.50)
Four-predictor reference	10%	82	9	3	75.0 (42.8–94.5)	11.0 (5.1–19.8)	0.43 (-2.22–3.85)
Four-predictor reference	15%	32	8	4	66.7 (34.9–90.1)	25.0 (11.5–43.4)	1.84 (-0.69–4.85)
Four-predictor reference	20%	18	7	5	58.3 (27.7–84.8)	38.9 (17.3–64.3)	2.07 (-0.24–5.00)

Prediction stage	Threshold %	IBG-minus-reference net benefit/100 (95% CI)
Locked model	5	-0.39 (-0.64–-0.13)
Locked model	10	1.36 (0.22–2.93)
Locked model	15	-0.69 (-2.41–1.35)
Locked model	20	-1.59 (-3.66–0.85)
Updating-step CV	5	1.50 (0.53–2.88)
Updating-step CV	10	0.02 (-0.60–0.73)
Updating-step CV	15	-0.76 (-1.95–0.06)
Updating-step CV	20	-0.18 (-1.57–1.20)

Counts describe potential review volume. Sensitivity and PPV use Clopper–Pearson exact binomial 95% intervals. Absolute and paired net-benefit intervals use centre-stratified patient bootstrap resampling conditional on saved predictions. The four thresholds were retained from the prior scenario specification. Paired updating-step CV differences average per-patient threshold contributions across 20 repeats before resampling; these supplementary intervals use 2,000 resamples (seed 20261017). Intervals are exploratory and unadjusted. Supplementary Data additionally reports outcomes within the unchanged 0–23, 24–33, and 34–59 point strata, by centre and pooled. These score strata are distinct from the probability thresholds defining additional-review scenarios.

Supplementary Table 25 External sensitivity analyses

Sensitivity	Version	n	Severe AUC	Severe Brier	Multiclass Brier
Matched numerical renal definition	IBG	204	0.929	0.0591	0.2650
Matched numerical renal definition	IBG, numerical eGFR	204	0.931	0.0536	0.2592
Omit adjudicated death with unsynchronized source note	IBG	204	0.929	0.0591	0.2650
Omit patient with isolated area value 5	IBG	204	0.896	0.0630	0.2695

The numerical-eGFR comparison retains only patients for whom scores could be calculated under both renal definitions. One Xiamen patient had a previously clinically adjudicated 180-day death while the source note retained uncertain/lost-follow-up wording; omission assesses its influence. One Hunan record contained a single area value of 5 in a range-style field; the primary analysis used the 5–20 cm² category and omission is reported. Recomputing NLR from the same recorded differential percentages did not change any complete-input IBG grade. Source-data queries and transformation records are archived.

Supplementary Table 26 Locked CRP specification and updating candidate parameters

Predictor	Category	Points
Age	<65 years	0
Age	≥65 years	3
Heart failure grade	0	0
Heart failure grade	1	2
Heart failure grade	2	15
Foot-temperature grade	1	15
Foot-temperature grade	2	13
Foot-temperature grade	3	8
Foot-temperature grade	4	0
eGFR grade	1	0
eGFR grade	2	1
eGFR grade	3	2
eGFR grade	4	3
eGFR grade	5	9
Wound-area grade	1	0
Wound-area grade	2	1
Wound-area grade	3	7
Wound-area grade	4	8
CRP grade	0	0
CRP grade	1	7
CRP grade	2	11

CRP categories are <10, 10 to <100 and ≥100 mg/L (grades 0, 1 and 2). Age is coded at 65 years; other clinical categories follow the main Methods. This supplementary score was fitted only in development data and locked after the earlier two-centre results had been inspected. Its coefficients and point values use development data exclusively.

Locked model	Score range	Any intercept	Any slope	Severe intercept	Severe slope
IBG	[0, 59]	-5.087090901836143	0.17190274442248235	-6.943767593629686	0.18545074627493308
CRP	[0, 61]	-5.668529114922584	0.17123629031445656	-7.502784718085149	0.18675925767064208
Four-predictor reference	[0, 60]	-3.2304121963823134	0.10992343920264437	-4.705526502287174	0.1200821001294649

Candidate model	Any intercept offset	Severe intercept offset
IBG	-0.61074777	-1.46944093
CRP	-0.41989254	-1.21729667
Four-predictor reference	-0.01735310	-0.69784604

For score S, q1=logistic(a0+a1×S), q2=logistic(b0+b1×S), and the class probabilities are (1−q1, q1−q2, q2). Candidate offsets are added only to the corresponding locked intercept. Fitting to all 205 patients defines candidate parameters for subsequent independent evaluation. Original maps, full-precision parameters, score-level probabilities, fold assignments and held-out predictions are included in the external reproducibility package.

Supplementary Table 27 Internal predictive performance and reference comparisons

Procedure	Any-adverse-outcome AUC (95% CI)	Severe-outcome AUC (95% CI)	mAUC (95% CI)	Multiclass Brier (95% CI)	Severe-outcome calibration slope (95% CI)
Fixed six, original IBG	0.7824 (0.7318–0.8301)	0.8279 (0.7552–0.8887)	0.8052 (0.7507–0.8519)	0.3335 (0.2869–0.3818)	1.014 (0.737–1.365)
Fixed six, re-estimated IBG cut-offs	0.7815 (0.7310–0.8293)	0.8325 (0.7627–0.8914)	0.8070 (0.7534–0.8527)	0.3337 (0.2870–0.3818)	1.017 (0.738–1.364)
Specified automated development	0.7557 (0.7037–0.8033)	0.8093 (0.7332–0.8735)	0.7825 (0.7262–0.8318)	0.3447 (0.2985–0.3933)	1.042 (0.737–1.431)

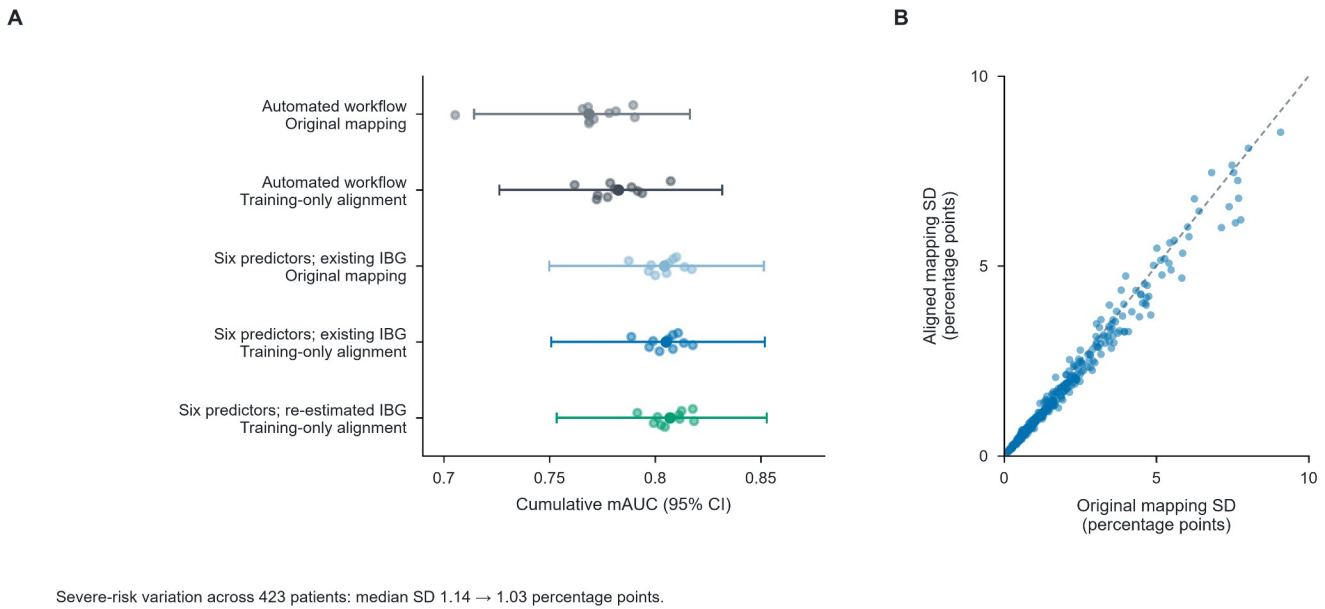
Values are means across ten subsequent matched assessments with patient-cluster 95% CIs conditional on saved OOF predictions. The fixed-six SF_align procedure retained predictor identities and clinical coding while refitting the score in training partitions. The re-estimated-IBG procedure also re-estimated component cut-offs; the automated workflow selected variables and configurations within training. Training-only probability alignment was used. These follow-up estimates do not replace the primary historical five-fold OOF mAUC or its original bootstrap interval.

Supplementary Table 28 Locked-model performance in the same 205 external patients

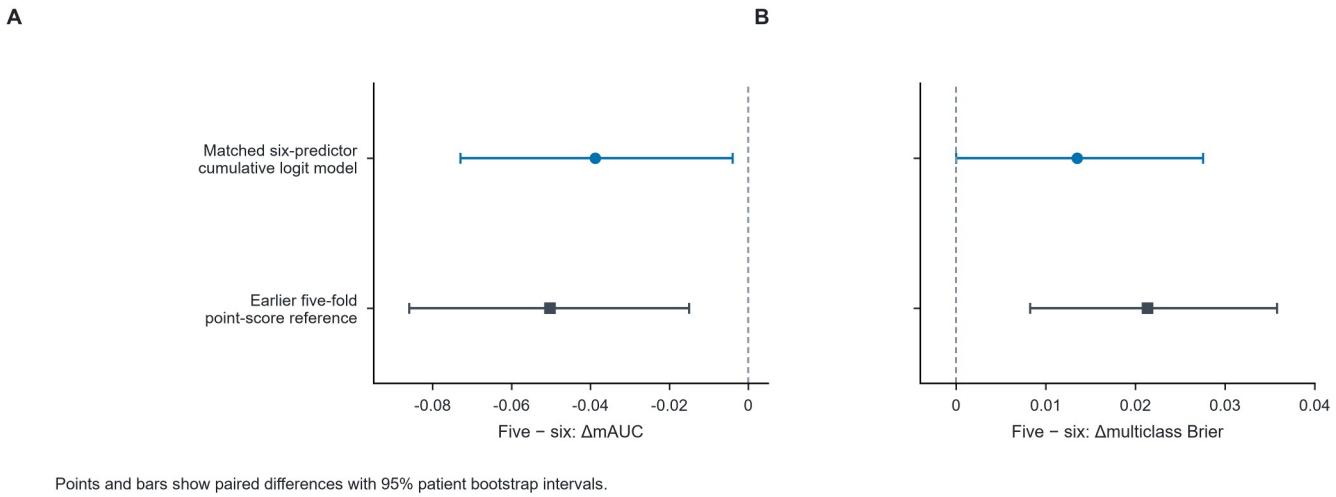
Locked version	Severe AUC (95% CI)	Severe-outcome Brier (95% CI)	Multiclass Brier (95% CI)	Mean predicted / observed severe risk
six-predictor IBG	0.895 (0.803–0.963)	0.0632 (0.0432–0.0833)	0.2712 (0.2214–0.3207)	15.0%／5.9%
six-predictor CRP supplementary candidate	0.840 (0.686–0.967)	0.0577 (0.0386–0.0783)	0.2858 (0.2339–0.3387)	13.6%／5.9%
Four-predictor reference	0.862 (0.759–0.953)	0.0472 (0.0295–0.0681)	0.3065 (0.2456–0.3682)	10.1%／5.9%

Severe events were ipsilateral below-knee amputation or all-cause death within 180 days (12 events). All models used the same 205 patients; intervals used 2,000 within-centre patient bootstrap samples. Severe-outcome Brier score evaluates the binary severe endpoint; multiclass Brier score is the patient-level mean of summed squared errors across three outcome probabilities. Lower values indicate smaller error for both metrics. CRP was a supplementary candidate version. Detailed paired estimates are in Supplementary Table 21.

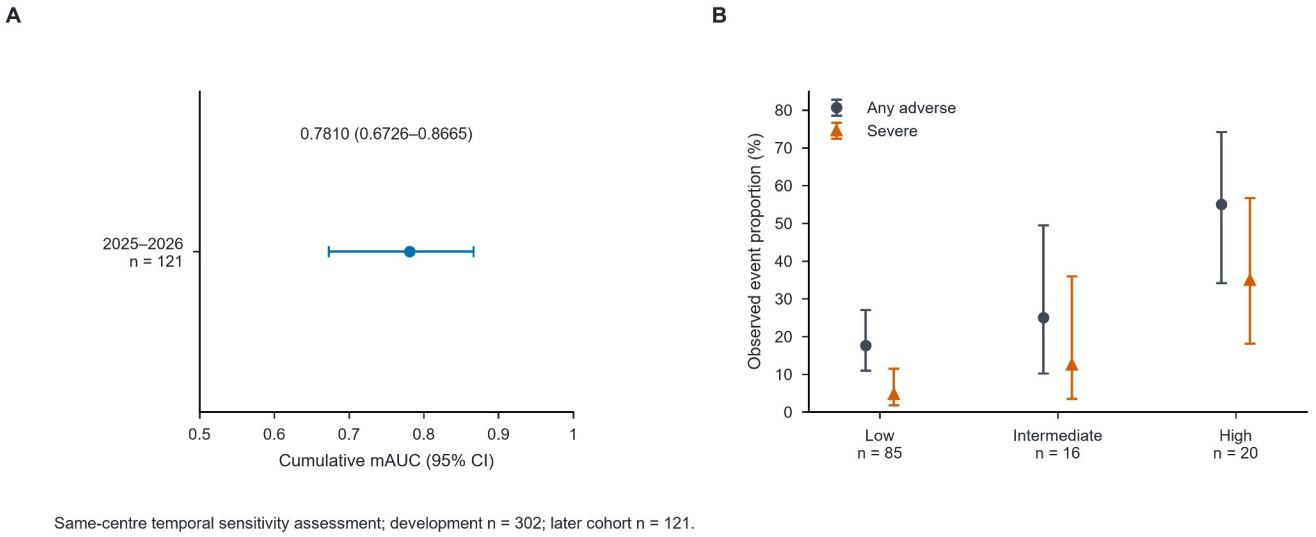
Supplementary Figures



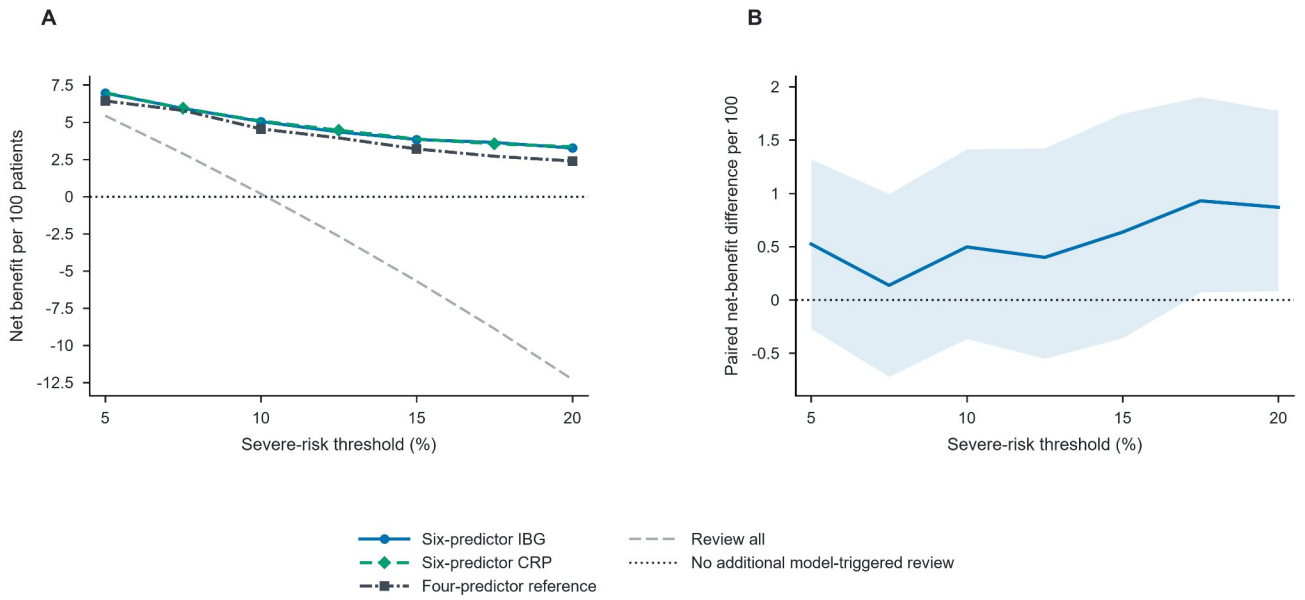
Supplementary Fig. 1 Training-only probability alignment and variability in repeated predictions. (A) Cumulative mAUC for automated and six-predictor scoring procedures, with ten repetition-specific estimates and patient-cluster intervals based on 2,000 resamples conditional on fixed OOF predictions. (B) Sample SD of each patient's ten predicted severe-outcome probabilities under original and aligned SF mappings. Both axes show percentage points; the diagonal indicates equal SDs. Source estimates are in Supplementary Table 5 and Supplementary Data.



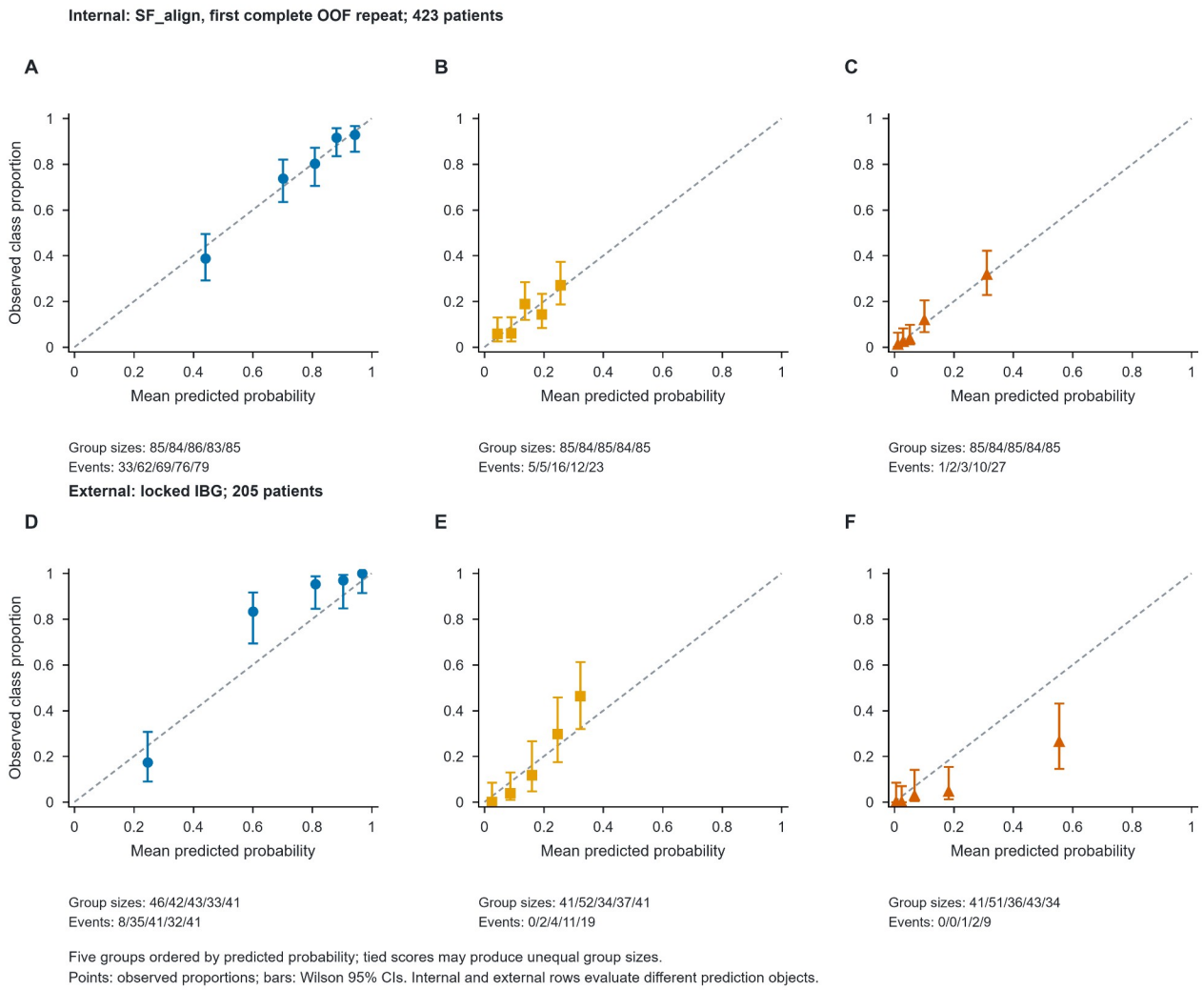
Supplementary Fig. 2 Paired changes after removal of palpated foot temperature. The left panel shows cumulative mAUC differences and the right panel shows multiclass Brier differences. Contrasts are five predictors minus the matched six-predictor model or earlier five-fold point-score reference, with intervals from 2,000 paired patient bootstrap resamples (Supplementary Table 12).



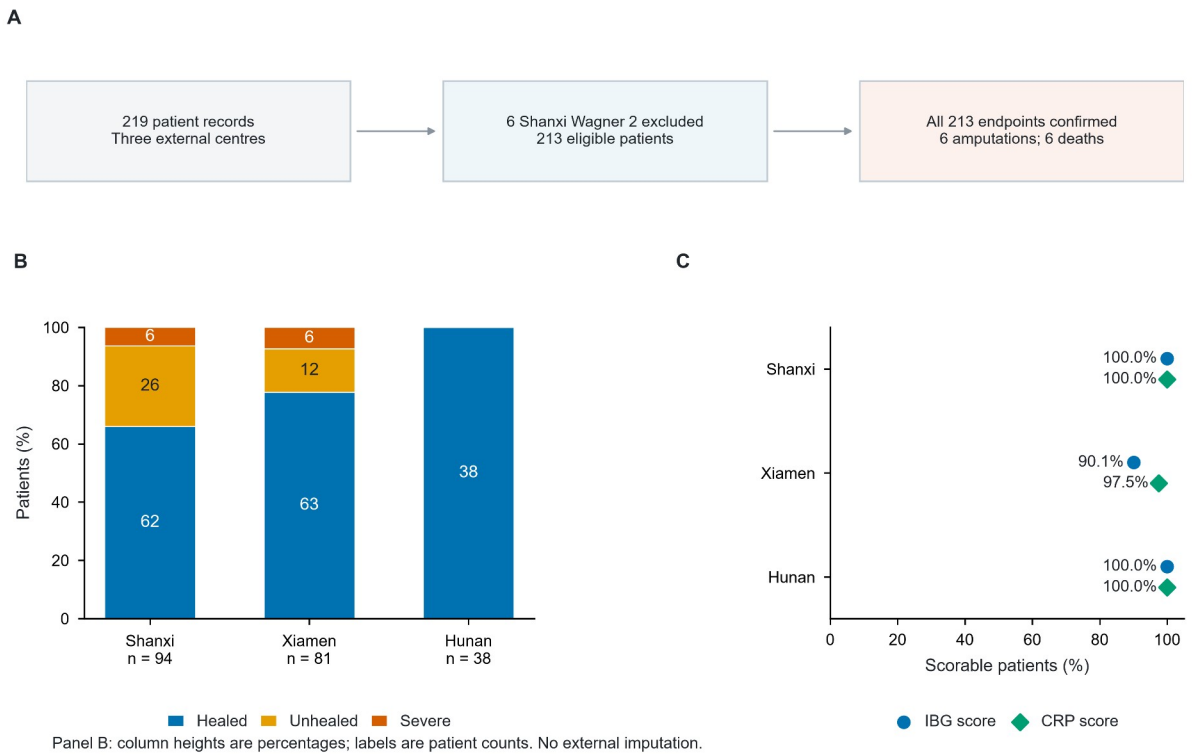
Supplementary Fig. 3 Same-centre temporal evaluation. (A) Cumulative mAUC and its interval from 500 bootstrap resamples in the 121-patient later cohort. (B) Any-adverse-outcome and severe-outcome proportions with Wilson 95% CIs, using development-period boundaries of approximately 26.9 and 33.0. Training included 302 patients; predictor identities and upstream specifications originated from overall development. These estimates describe a temporal sensitivity analysis (Supplementary Table 15).



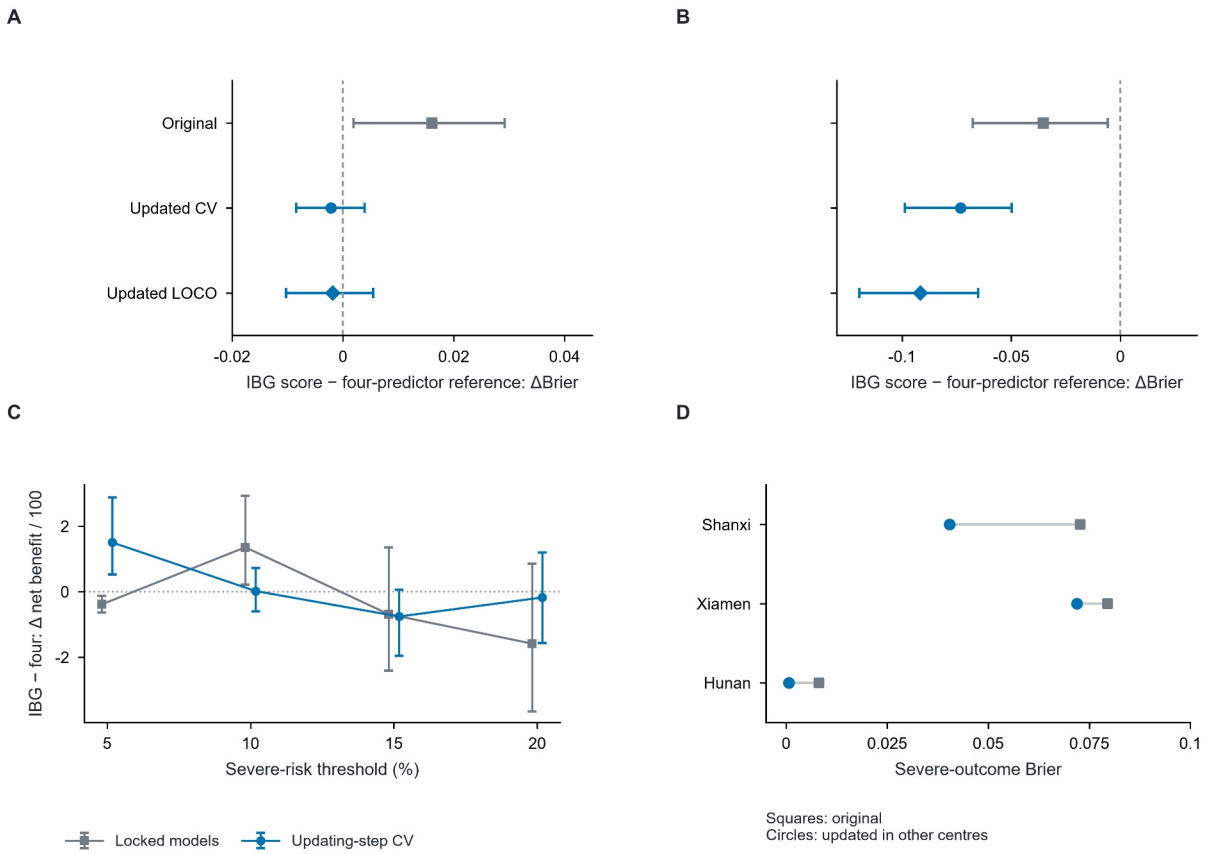
Supplementary Fig. 4 Severe-risk scenarios for additional preoperative cardiorenal review. (A) Net benefit for required-input IBG, CRP, and four-predictor procedures, alongside review-all and no additional model-triggered review strategies. (B) Paired IBG-minus-reference net-benefit difference with a patient-cluster 95% CI conditional on saved OOF predictions. The panels show thresholds of 5–20%; Supplementary Table 18 presents prespecified scenarios and Supplementary Data provides the full grid. Routine care continues under each strategy; threshold-positive counts describe potential review volume.



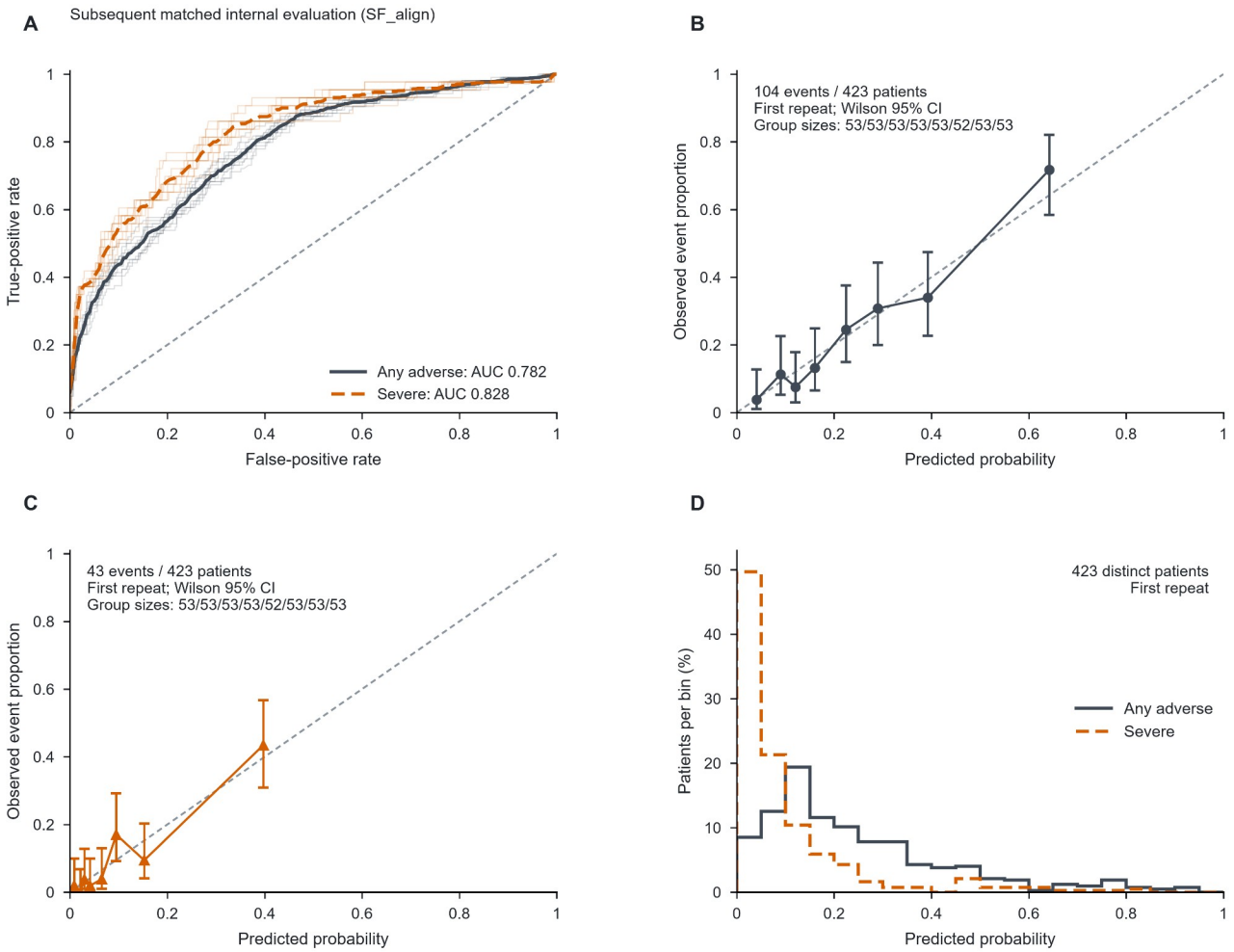
Supplementary Fig. 5 Internal and external class-specific calibration. (A–C) First complete OOF repetition of the subsequent matched SF_align procedure (423 patients). (D–F) Locked IBG predictions in 205 external patients. Columns represent healing, persistent non-healing, and ipsilateral below-knee amputation or death. Five groups are ordered by predicted class probability; tied scores can yield unequal group sizes. Points show observed proportions, bars show Wilson 95% CIs, and dashed lines indicate equality. Group sizes and event counts appear below each panel. All panels use common 0–1 probability scales, with small display margins to retain boundary markers. Numerical results are in Supplementary Tables 5D and 21.



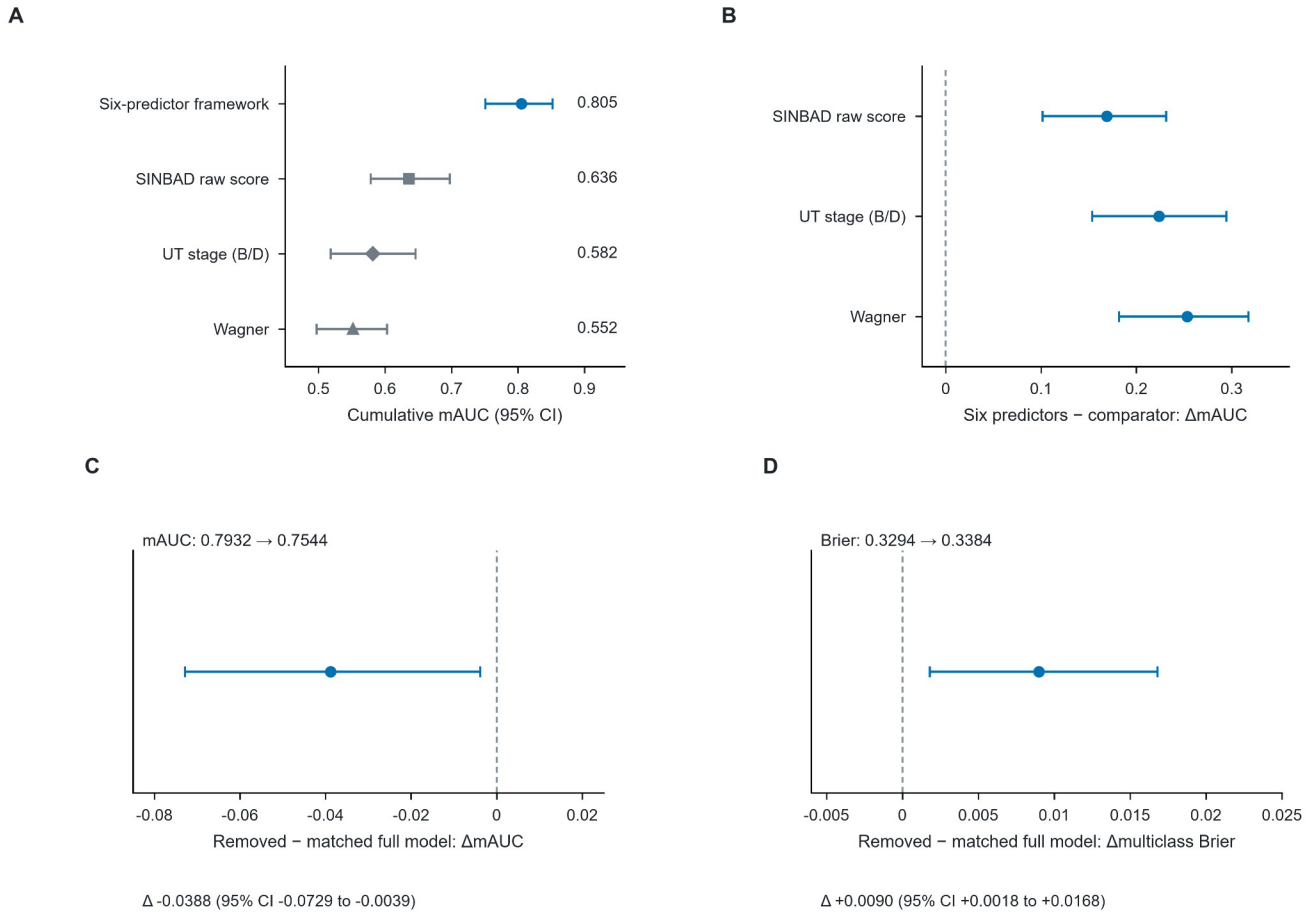
Supplementary Fig. 6 External cohorts and input availability. (A) Source records, exclusion of six Shanxi patients with Wagner grade 2 ulcers, and confirmed 180-day outcomes. (B) Mutually exclusive outcomes by centre; column heights show percentages and labels show patient counts. (C) Observed-input coverage of locked IBG and CRP models. All 38 Hunan patients had complete IBG inputs and clinically confirmed 180-day outcomes.



Supplementary Fig. 7 Probability error and paired review scenarios in matched external patients. (A) IBG-minus-reference severe-outcome Brier difference. (B) Multiclass Brier difference. Negative values favour IBG. (C) Paired net-benefit differences at the four prespecified thresholds for locked and updating-step cross-validation predictions; positive values favour IBG. Bars in panels A–C are centre-stratified patient-bootstrap 95% CIs conditional on saved predictions. (D) Severe-outcome Brier before updating and after fitting offsets in other centres. All panels use the same 205 patients; updating comparisons are supplementary.



Supplementary Fig. 8 Discrimination and calibration in subsequent matched internal evaluation. (A) Receiver operating characteristic (ROC) curves and mean AUC across ten SF_align OOF repetitions; thin lines show individual repetitions and the thick line shows their mean. (B–C) Calibration for any adverse outcome and severe outcomes in the first repetition, with prediction-group means, observed proportions, and Wilson 95% CIs. (D) Predicted probability distributions in the same repetition. Panels B–D each use one prediction per patient (n=423). Group coordinates and counts are in Supplementary Data.



Supplementary Fig. 9 Clinical grading comparisons and matched predictor contributions. (A) Cumulative mAUC and 95% CIs for the six-predictor procedure and clinical grading systems in subsequent matched repeated evaluation. (B) Paired mAUC differences using identical patients and repeated outer folds. (C) mAUC difference after removing palpated foot temperature from the matched categorical ordinal model. (D) Multiclass Brier difference after removing IBG from the matched graded model. Panels C–D use their respective matched references and assess discrimination and probability error, respectively.

Supplementary References

1. van Geloven, N. et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur. J. Epidemiol.* **35,** 619–630 (2020). https://doi.org/10.1007/s10654-020-00636-1

2. Jeffcoate, W. et al. Causes, prevention, and management of diabetes-related foot ulcers. *Lancet Diabetes Endocrinol.* **12,** 472–482 (2024). https://doi.org/10.1016/S2213-8587(24)00110-4

3. Fitridge, R. et al. The intersocietal IWGDF, ESVS, SVS guidelines on peripheral artery disease in people with diabetes and a foot ulcer. *Diabetes Metab. Res. Rev.* **40,** e3686 (2024). https://doi.org/10.1002/dmrr.3686

4. Hua, Q. et al. Treatment experiences of 516 cases of diabetic foot treated with tibial transverse transport. *Zhongguo Xiu Fu Chong Jian Wai Ke Za Zhi* **34,** 959–963 (2020). https://doi.org/10.7507/1002-1892.202003099

5. Saffari, S. E. et al. AutoScore-Ordinal: an interpretable machine learning framework for generating scoring models for ordinal outcomes. *BMC Med. Res. Methodol.* **22,** 286 (2022). https://doi.org/10.1186/s12874-022-01770-y

6. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: issues and guidance for practice. *Stat. Med.* **30,** 377–399 (2011). https://doi.org/10.1002/sim.4067

7. van Buuren, S. *Flexible Imputation of Missing Data* 2nd edn (CRC Press, Boca Raton, 2018).

8. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26,** 565–574 (2006). https://doi.org/10.1177/0272989X06295361

9. Yang, S. et al. Correlation analysis of patients with diabetic foot ulcers treated with tibial cortex transverse transport surgery and platelet-to-lymphocyte ratio and monocyte-to-neutrophil ratio. *Adv. Clin. Exp. Med.* **34,** 549–559 (2025). https://doi.org/10.17219/acem/187765

10. Zhu, Y. et al. Association of nutritional–immune indicators and their dynamic changes with disease severity and wound healing in diabetic foot ulcers treated by tibial transverse transport: a retrospective study. *Front. Nutr.* **13,** 1863436 (2026). https://doi.org/10.3389/fnut.2026.1863436

11. Tao, H. et al. An interpreting machine learning models to predict amputation risk in patients with diabetic foot ulcers: a multi-center study. *Front. Endocrinol.* **16,** 1526098 (2025). https://doi.org/10.3389/fendo.2025.1526098

12. 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). https://doi.org/10.1136/bmj-2024-082505

13. Riley, R. D. et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* **368,** m441 (2020). https://doi.org/10.1136/bmj.m441

14. Riley, R. D. et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ* **384,** e074821 (2024). https://doi.org/10.1136/bmj-2023-074821