Journal of the American Medical Informatics Association, 2026, 1–16
https://doi.org/10.1093/jamia/ocag184
Research and Applications

Subgroup deployment-readiness audit of calibration and clinical utility in perioperative infection prediction

Hugo Guillen-Ramirez, PhD[1,2,3,*], Katharina Lucia Lucas, MD[2], Yves Max Wintsch, MD[2], Tobias Ueli Blatter, PhD[2,4], Karen Triep, MD[5], Olga Endrich, MD[5,6], Guido Beldi, MD[*,1,2]

[1]Department of Surgery, Fribourg Hospital, University of Fribourg, 1752 Fribourg, Switzerland
[2]Department of Visceral Surgery and Medicine, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland
[3]Department of Digital Medicine, University of Bern, 3008 Bern, Switzerland
[4]Multidisciplinary Center for Infectious Diseases, University of Bern, 3012 Bern, Switzerland
[5]Medical Directorate, Medizincontrolling, Inselspital, University Hospital Bern, Insel Gruppe, 3010 Bern, Switzerland
[6]Department of Clinical Chemistry, Inselspital, Bern University Hospital, University of Bern, 3010 Bern, Switzerland
*Corresponding authors: Hugo Guillen-Ramirez, PhD, Department of Visceral Surgery and Medicine, Inselspital, Bern University Hospital, University of Bern, Freiburgstrasse 20, 3010 Bern, Switzerland (hugo.guillenramirez@unibe.ch) and Guido Beldi, MD, Department of Surgery, Fribourg Hospital, University of Fribourg, Chemin des Pensionnats 2-6, 1752 Villars-sur-Glâne, Fribourg, Switzerland (guido.beldi@unifr.ch)

Abstract

Objective: **We audited deployment readiness of an end-of-surgery postoperative infection model across clinically relevant subgroups and assessed mitigation of subgroup miscalibration.**

Materials and Methods: **We analyzed out-of-fold predictions from 10 719 procedures (2014-2022) with 504 present-on-admission (POA)–coded postoperative infection events across recorded sex, age, and a clinical physiological-reserve proxy. We assessed discrimination, calibration, threshold-specific errors, and decision-curve net benefit, treating $t = 0.10$ as a reference anchor. We compared no recalibration with common and group-specific cross-fitted isotonic recalibration, evaluated a refit without the Synthetic Minority Oversampling Technique (SMOTE), and performed a temporal-holdout analysis.**

Results: **Overall area under the receiver operating characteristic curve (AUROC) was 0.876, with similar sex-marginal discrimination in women and men (0.878 vs 0.875). Calibration-in-the-large was −1.02 overall, −0.86 in frail, and −1.34 in prefrail patients; the 6-event nonfrail stratum was not estimable. At $t = 0.10$, net benefit was positive in frail and negative in prefrail patients in the source cohort, but negative prefrail utility did not persist in the ascertainable-period (2018-2022) and temporal analyses. Common recalibration moved overall calibration-in-the-large from −1.020 to −0.007. Removing SMOTE shifted overall calibration but did not reduce the reserve gradient. Temporal-holdout AUROC was 0.851 with persistent overprediction. Present-on-admission coding was effectively absent before 2018; prevalence-dependent results were therefore additionally reported for the ascertainable 2018-2022 period, and ascertainment differed by physiological reserve.**

Discussion: **Discrimination-only or sex-marginal evaluation would have missed deployment-relevant subgroup heterogeneity, while endpoint ascertainment affected prevalence-dependent interpretation.**

Conclusion: **Joint subgroup evaluation can identify deployment-relevant limitations missed by aggregate metrics. Prospective validation with comparable outcome ascertainment remains necessary before clinical deployment.**

Key words: machine learning; postoperative infection; algorithmic bias; algorithmic fairness; health equity; clinical prediction models.

Introduction

Postoperative bacterial infection drives preventable morbidity, length of stay, cost, and mortality.[1–3] Although infections are typically apparent by postoperative day 5 and C-reactive protein can shift identification to about day 3,[4] our prior work aimed to identify risk as early as the end of surgery. We first used 91 794 perioperative electronic health records (EHR) to predict infection within 2 days (area under the receiver operating characteristic curve [AUROC] of 0.83), with strategic missing-laboratory imputation improving recall.[5] We then used interpretable summary, trend, and distributional descriptors of intraoperative vital signs in 10 719 procedures, achieving AUROC 0.88 (95% CI, 0.85-0.91), cluster-level calibration, and SHAP-based interpretability.[6] However, strong

Received: June 19, 2026. **Revised:** September 29, 2026. **Accepted:** October 19, 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

aggregate discrimination is not sufficient evidence of subgroup deployment readiness.

Strong average performance can mask subgroup degradation. Examples in medical AI are severity underestimation in Black patients,[7] underdiagnosis by chest-radiograph classifiers in women and younger patients,[8] and temporal fairness drift in postsurgical risk scores despite maintained aggregate calibration.[9] Accuracy and calibration can degrade, and fairness gaps increase, even while aggregate discrimination remains stable.[9] Potential causes for such subgroup performance differences include unrepresentative training cohorts, unmodeled sex-associated differences,[10] incomplete inclusion of physiological reserve,[11] varying signal quality,[12,13] and differences in surveillance intensity or data completeness.[14–20]

Accordingly, TRIPOD+AI recommends reporting model performance, including uncertainty, within clinically relevant sociodemographic and clinical subgroups.[21] Subgroup deployment readiness in clinical AI requires joint assessment of discrimination, calibration, and clinical utility alongside relevant fairness criteria, rather than reliance on error-rate parity alone.[22,23] Existing software supports many components of such an assessment, including general fairness metrics and mitigation methods, decision-curve analysis, and fairness evaluation in medical imaging.[24–28] However, these capabilities are distributed across tools with different primary aims, making integrated subgroup assessment of calibration and clinical utility less straightforward for tabular and time-series clinical prediction models.

We retrospectively audited a previously developed end-of-surgery infection prediction model across recorded sex, age stratum, and a 3-level EHR-derived physiological-reserve classification. For each axis, we evaluated 3 clinically distinct dimensions: (1) discrimination, whether infected patients received higher predicted risks than noninfected patients; (2) calibration, whether predicted probabilities matched observed risks; and (3) clinical utility, whether acting at the intended threshold provided net benefit. We implemented the audit in isitfair, an open-source Python package used to conduct these analyses and generate their structured reports.

Methods

Cohort, model, and outcome

The primary audit evaluates the published model within its original 2014-2022 source cohort: 10 719 surgical procedures performed at Inselspital, Bern University Hospital, including 504 present-on-admission (POA)–coded postoperative bacterial infection events (Table 1; sex-stratified effect sizes for the characterization variables are reported in Table S1). Present-on-admission flags are used to exclude infections coded as already present at admission; where those flags are absent, infection timing cannot be established, and the infection case was excluded from cohort construction rather than retained as a negative outcome. Because POA flags entered routine use only in 2018, the 2014-2017 portion of the cohort contributes very few coded events; the full 2014-2022 cohort is nevertheless retained for the primary audit because it corresponds to the published model and its original out-of-fold predictions. Outcome ascertainment across the cohort window is characterized at the beginning of the Results, and key prevalence-dependent analyses are additionally reported for the ascertainable 2018-2022 period. The source cohort was not restricted to operating room or invasive cases. The cohort, outcome definition, model architecture, and training pipeline were those of the previously reported end-of-surgery model.[6] Briefly, the model used summary, trend, and distributional descriptors from intraoperative vital-sign time series to estimate the probability of postoperative bacterial infection at the end of surgery. We audited out-of-fold predictions from 5-fold stratified cross-validation, with the Synthetic Minority Oversampling Technique (SMOTE) used solely to rebalance the outcome classes within each training fold. Synthetic Minority Oversampling Technique was applied only after fold assignment and never to held-out observations. Its effect on calibration was examined in a no-SMOTE refit reported in Table S2. The outcome was bacterial postoperative infection during the index admission, defined as the continuous inpatient stay containing the surgery, from admission to discharge, using the prespecified ICD-10-GM coding pipeline from the source cohort. Present-on-admission flags, where available, were used to exclude infections coded as already present at admission.

Sensitive attributes and subgroup axes

The audit prespecified 3 subgroup axes available at surgery: recorded sex, age group, and physiological reserve. Recorded sex and age were sociodemographic audit axes, whereas physiological reserve was clinical. Physiological reserve was included because unnecessary monitoring escalation was expected to have different consequences by baseline reserve. Recorded sex was extracted from the administrative hospital information system and was binary in the source data. Gender identity was not captured; analyses therefore refer to recorded sex, not gender. Age was grouped a priori as 0-17, 18-44, 45-64, 65-79, and $\geq$80 years to represent clinically recognizable perioperative age strata while maintaining usable subgroup sizes. Because the institutional pediatric pathway covers patients aged $\leq$15 years, whereas the prespecified first stratum was 0-17 years, sensitivity analyses re-binning age at the institutional $\leq$15-year cutoff, as well as analyses using equal-frequency quintiles, are reported in Table S3.

The physiological-reserve proxy was a 3-level variable (frail, prefrail, and nonfrail) combining the American Society of Anesthesiologists (ASA) score, Hospital Frailty Risk Score (HFRS),[11] and a 4-item modified Frailty Index (mFI-4; mFI-5[29] without functional status, unavailable in our EHR). Components were not analyzed separately. Frail required ASA-PS $\geq$ 4, HFRS $>$ 15, or mFI-4 $\geq$ 2; prefrail required no frail criterion but ASA-PS 3, HFRS 5-15, or mFI-4 = 1; all remaining patients were nonfrail. The mFI-4 comprised diabetes mellitus, chronic obstructive pulmonary disease, congestive heart failure, and hypertension; item definitions and HFRS weights are provided in Tables S4 and S5. All reserve components were derived from routinely collected clinical and administrative EHR data available at the time of surgery; no additional prospective assessment was performed. Because hypertension was approximated from ICD-10 codes without medication linkage, we also evaluated an ASA/HFRS-only reserve definition (Table S6). We use “frail,” “prefrail,” and “nonfrail” only as labels for the reserve proxy, not as diagnoses

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

Table 1. Cohort characterization by sex.

Variable	Level	Overall	Female	Male	*P*-value	FDR
n		10 719	3689	7030		
Age, median [Q1, Q3]		66.0 [55.0, 74.0]	66.0 [53.0, 75.0]	66.0 [56.0, 73.0]	.467	0.467
Age group, *n* (%)	0-17	367 (3.4)	175 (4.7)	192 (2.7)	<.001	0.0016
	18-44	941 (8.8)	408 (11.1)	533 (7.6)		
	45-64	3715 (34.7)	1179 (32.0)	2536 (36.1)		
	65-79	4684 (43.7)	1470 (39.8)	3214 (45.7)		
	80+	1012 (9.4)	457 (12.4)	555 (7.9)		
ASA score, *n* (%)	1	1227 (11.4)	630 (17.1)	597 (8.5)	<.001	0.0016
	2	213 (2.0)	78 (2.1)	135 (1.9)		
	3	3803 (35.5)	1549 (42.0)	2254 (32.1)		
	4	5476 (51.1)	1432 (38.8)	4044 (57.5)		
Physiological reserve, *n* (%)	Frail	6139 (57.3)	1666 (45.2)	4473 (63.6)	<.001	0.0016
	Prefrail	3558 (33.2)	1501 (40.7)	2057 (29.3)		
	Nonfrail	1022 (9.5)	522 (14.2)	500 (7.1)		
Emergency surgery, *n* (%)	No	9250 (86.3)	3221 (87.3)	6029 (85.8)	.028	0.0373
	Yes	1469 (13.7)	468 (12.7)	1001 (14.2)		
Surgery duration, median [Q1, Q3]		213.0 [147.0, 281.0]	193.0 [135.0, 267.0]	221.0 [156.0, 285.0]	<.001	0.0016
Procedure cluster, *n* (%)	Nervous system	903 (8.4)	365 (9.9)	538 (7.7)	<.001	0.0016
	Thoracic	807 (7.5)	317 (8.6)	490 (7.0)		
	Cardiac	3764 (35.1)	836 (22.7)	2928 (41.7)		
	Vascular	769 (7.2)	189 (5.1)	580 (8.3)		
	Abdominal wall	530 (4.9)	249 (6.7)	281 (4.0)		
	Gastrointestinal	298 (2.8)	105 (2.8)	193 (2.7)		
	Hepatobiliary	306 (2.9)	95 (2.6)	211 (3.0)		
	Uroandrological	346 (3.2)	98 (2.7)	248 (3.5)		
	Gynecological	303 (2.8)	303 (8.2)	0 (0.0)		
	Musculoskeletal	1623 (15.1)	711 (19.3)	912 (13.0)		
	Cutaneous	213 (2.0)	95 (2.6)	118 (1.7)		
	Unclassified	857 (8.0)	326 (8.8)	531 (7.6)		
Postoperative infection, *n* (%)	No	10 215 (95.3)	3499 (94.8)	6716 (95.5)	.123	0.141
	Yes	504 (4.7)	190 (5.2)	314 (4.5)		

Counts (column percentages) for categorical variables and median [Q1, Q3] for continuous variables, stratified by administratively recorded sex. *P*-values are from the χ^2 test for categorical variables and the Mann-Whitney *U* test for continuous variables. FDR is the Benjamini-Hochberg–adjusted *q*-value across the 8 tested variables, controlling the false discovery rate at $q < 0.05$. Physiological reserve is the 3-level composite of ASA Physical Status, the HFRS, and a 4-item modified Frailty Index (mFI-4; the mFI-5 without functional status) described in Methods.
Abbreviations: ASA, American Society of Anesthesiologists; FDR, false discovery rate.

of a validated frailty syndrome. Surgical specialty and procedure clusters were included as case-mix reference axes rather than as data-driven patient subgroups; procedure clusters were inherited from the previously developed model and its MAP-CARE-based procedure representation.[30] Pairwise intersections were prespecified for sex by age, sex by physiological reserve, and age by physiological reserve. No data-driven subgroup discovery was performed.

Subgroup performance metrics

We evaluated subgroup performance across discrimination, calibration, and clinical utility: whether infected patients received higher scores, whether predicted risks matched observed risks, and whether using the score to trigger intensified surveillance, here termed monitoring, produced clinical benefit. Because the score is used as an individualized probability for threshold-based surveillance, the primary deployment criteria were subgroup calibration and subgroup clinical utility. Threshold-specific parity gaps were treated as descriptive characteristics of the decision rule rather than as evidence of unfairness by themselves. Raw differences across age, physiological reserve, and procedure-cluster strata were interpreted in the context of case mix and outcome prevalence.

Discrimination

We assessed whether patients who developed infection received higher predicted risks using AUROC and area under the precision-recall curve (AUPRC). We prespecified $t = 0.10$ as a reference operating anchor for intensified surveillance rather than as an evidence-derived optimum. At $t = 0.10$, the decision-curve false-positive weight is $t/(1-t) = 0.10/0.90 = 1/9$ relative to the true-positive term; this weighting follows mathematically from the threshold and was not independently elicited as a clinical preference. Threshold-specific parity gaps are reported

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

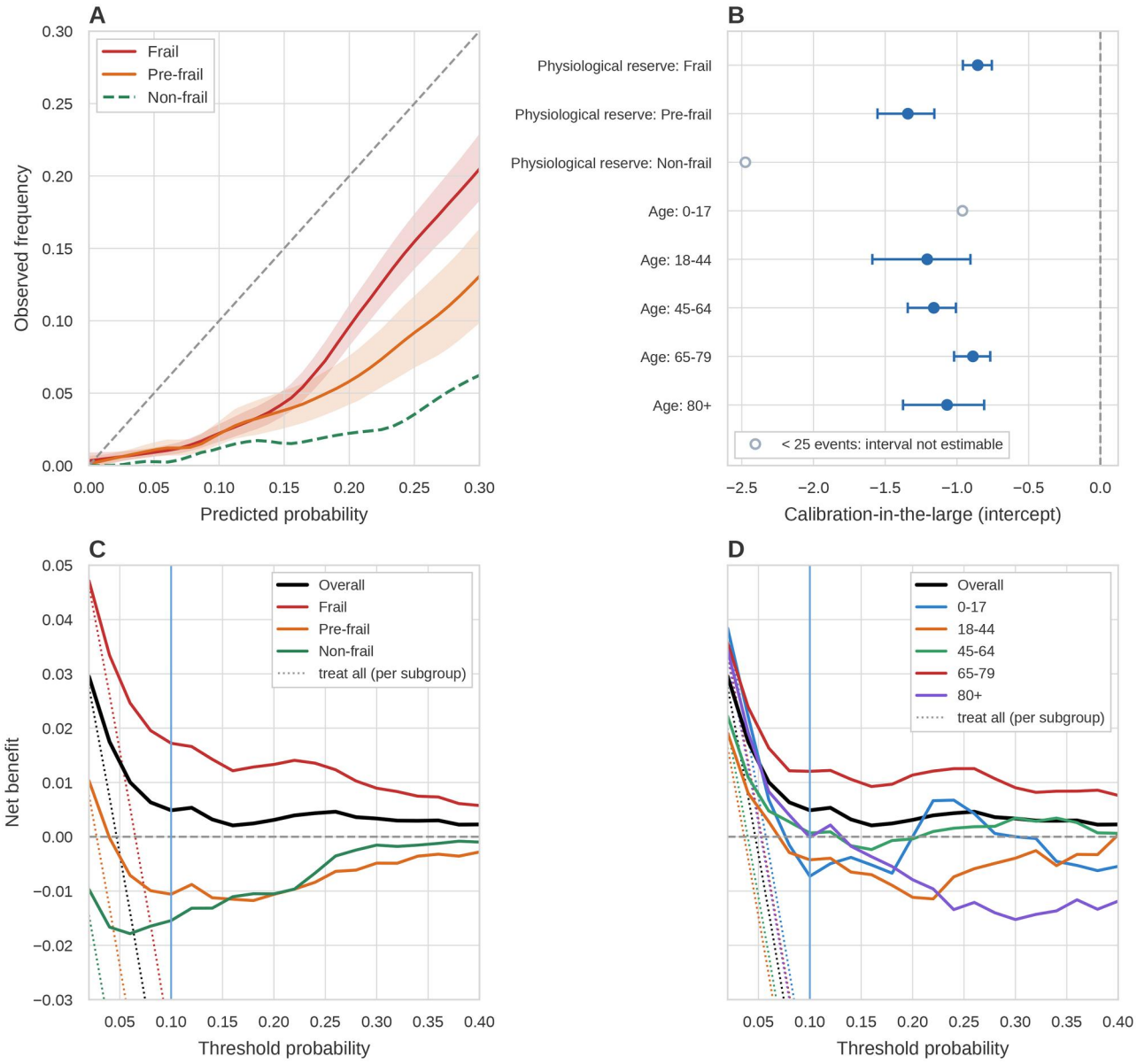

Figure 1. Subgroup calibration and clinical utility for the 2 axes that drive calibration heterogeneity (physiological reserve and age), audited cohort ($n = 10\,719$; 504 postoperative infection events). (A) LOESS-smoothed calibration curves with 95% bootstrap bands (2000 patient-level resamples) by physiological reserve (frail, prefrail, and nonfrail), over the predicted-probability range [0-0.30]; the dashed diagonal is perfect calibration and curves below indicate overestimation. (B) Calibration-in-the-large (fixed-slope logit intercept) with 95% bootstrap CIs for the physiological-reserve and age strata; 0 is perfect calibration and negative values indicate overestimation. Strata below the prespecified 25-event floor for calibration-intercept inference, nonfrail (6 events) and age 0-17 (21 events), are drawn with hollow markers and no interval and are shown descriptively; in panel A, the nonfrail curve is drawn dashed without a bootstrap band. (C, D) Decision curves of net benefit across $t = 0.02$ to 0.40 for (C) physiological reserve and (D) age. The $t = 0.10$ operating threshold (vertical blue line) is the risk level at which intensified monitoring would be triggered. "Treat all (per subgroup)" is shown separately for each subgroup using that subgroup's observed outcome prevalence; the zero line denotes "treat none." Here, intensified monitoring denotes closer clinical surveillance, for example, more frequent vital-sign and laboratory checks and earlier clinical review, rather than measurement with a specific device or treatment escalation. A useful model lies above both reference strategies.

descriptively rather than as primary deployment criteria. Sensitivity analyses evaluate net benefit at thresholds from 0.05 to 0.20 (Table S7), while Figure 1C and D shows the decision curves across the broader range from 0.02 to 0.40. We also reported sensitivity, specificity, and threshold-specific error-rate gaps, namely the equal-opportunity difference (sensitivity gap), the predictive-equality difference (false-positive-rate gap), the equalized-odds gap (the larger of the sensitivity and false-positive-rate gaps), and the statistical-parity difference (difference in the proportion of patients flagged) at this threshold.[31,32]

Calibration

Calibration assesses whether predicted probabilities correspond to observed risks, for example, whether patients assigned a 10%

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

risk develop infection about 10% of the time. We summarized it with calibration curves, calibration-in-the-large (the intercept), calibration slope, Brier score, expected calibration error (ECE), and the integrated calibration index (ICI).[33–36] Because overall calibration can conceal subgroup bias, each measure was computed within every subgroup.

Clinical utility

Clinical utility assesses whether acting on the score improves decisions at a chosen threshold. We assessed clinical utility using decision-curve net benefit[37] across thresholds from 0.02 to 0.40 and at the prespecified 0.10 operating threshold. Because net benefit depends on outcome prevalence, standardized net benefit was also reported.[23] Full metric definitions and estimator equations are provided in the Supplementary Methods.

Calibration within subgroups

Subgroup calibration was prespecified because overall calibration may conceal clinically relevant over- or underprediction. Each cohort-level calibration measure was estimated within every subgroup because deployment uses individual probabilities, not only ranks. Overestimation can trigger unnecessary monitoring, delayed discharge, or avoidable antibiotics; underestimation can miss patients needing surveillance. We assessed calibration visually, parametrically, and with aggregate scores.

Visual calibration curves

For each subgroup, we plotted the observed infection rate against predicted risk using LOESS smoothing, with 95% bootstrap bands. The 45° identity line represents perfect calibration; a curve below it indicates that the model overestimates risk in that subgroup, whereas a curve above it indicates underestimation.

Parametric summaries

We estimated calibration intercept and slope on the log-odds scale.[33] The intercept (calibration-in-the-large) measures systematic over- or underprediction: zero is unbiased; a negative value indicates that the model overestimates risk on average. The slope measures whether the spread of predicted risks is correctly scaled: 1 is ideal; below 1 indicates predictions that are too extreme, above 1 indicates predictions that are too compressed.

Aggregate scores

We calculated the Brier score, ECE (mean absolute predicted-observed gap across 10 equal-frequency bins), and ICI (the same gap evaluated patient-wise against the LOESS curve). Hosmer-Lemeshow statistics using 10 risk groups are reported per subgroup in Table S8 alongside, not in place of, ECE and ICI, with degrees of freedom, *P*-values, event counts, and the corresponding risk-group regression slope and intercept. Table S9 reports observed and expected events within each risk group.[38] Because model parameters were not estimated in the data under test, we used the number of risk groups ($g = 10$) as the degrees of freedom rather than $g - 2$. Given the statistic's dependence on sample size and event count, H-L results are interpreted descriptively and jointly with the complementary calibration measures. Estimator equations are given in the Supplementary Methods.

Clinical utility within subgroups

Clinical utility was prespecified because calibration alone does not show whether acting on the score improves decisions. We used subgroup decision-curve analysis across $t = 0.02$-0.40. Net benefit counts true positives and subtracts false positives, weighted by the threshold-implied trade-off; positive values improve over monitoring no one, and usefulness also requires comparison with monitor-all. Because net benefit depends on prevalence, we also report standardized net benefit, where 1 is a hypothetical perfect model and 0 is no better than monitoring no one. The primary clinical-utility gap was the maximum absolute pairwise subgroup difference in net benefit across thresholds, with bootstrap CIs.

Bias-mitigation strategies

Isotonic recalibration[39] was evaluated as a 3-step ladder comprising no recalibration, common cohort-wide recalibration, and group-specific recalibration by physiological reserve. To avoid in-sample evaluation of recalibrated probabilities, both common and group-specific isotonic transformations were fitted with 5-fold cross-fitting: each transformation was estimated on 4 folds and applied to the held-out fold, yielding one out-of-fold recalibrated prediction per procedure. Calibration-in-the-large, calibration slope, ECE, ICI, Brier score, net benefit, and subgroup calibration differences were compared across the 3 steps (Figure S1). Because isotonic regression can overfit sparse strata, post-recalibration calibration slope and subgroup event counts were also inspected. Wasserstein-barycenter postprocessing[40–42] was retained as a complementary score-distribution mitigation benchmark on the procedure-cluster axis, where distributional shift was most pronounced. The Wasserstein transformation was constrained to demographic parity; that is, it aligned the score distributions so that the same proportion of patients was flagged in every group, and was treated as a parity-focused comparator.[43] Mitigations were evaluated by their effect on discrimination, calibration, Brier score, net benefit, and subgroup gaps and were considered favorable only when the affected subgroup was improved, or calibration was corrected without degradation of absolute subgroup performance.

To assess whether class rebalancing accounted for the observed calibration pattern, the source modeling pipeline was additionally reproduced and refit without SMOTE. The published out-of-fold predictions were required to be recovered exactly by the reproduction arm before the no-SMOTE comparison was interpreted (Table S2). Both the overall calibration shift and changes in the frail-prefrail calibration-intercept difference were compared.

Statistical analysis and reporting

Confidence intervals were estimated from 2000 bootstrap resamples drawn by resampling patients within each subgroup, allowing subgroup event counts to vary across replicates so that intervals reflect sampling variability in observed prevalence. Gap statistics were recomputed within each replicate to preserve dependence between subgroup estimates, and the overall

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

cohort was resampled once per replicate so that a single overall interval applies across all subgroup axes. Cells in which more than 10% of replicates yielded nonestimable values (eg, zero-event replicates) are reported as not estimable rather than with an interval; replicate accounting is provided in Table S10. Confidence intervals for subgroup metrics are reported only where the subgroup carries at least the prespecified minimum number of events (AUROC 25; calibration intercept 25; calibration slope 100; net benefit 50; AUPRC, sensitivity, specificity, and predictive values follow the AUROC floor, and the Brier score the calibration-intercept floor). Below these floors, the point estimate is shown descriptively without an interval and is labeled not estimable where the table reports intervals for that metric. An earlier scheme that resampled cases and noncases separately, holding subgroup event counts fixed, understated interval width in sparse strata; it is retained in the software for reproducibility of the originally submitted analysis and its consequences are shown in Table S11. For sparse intersectional cells with fewer than 50 infection events, empirical-Bayes shrinkage was applied toward the relevant parent subgroup; cell counts, event counts, and shrinkage factors are reported in Table S12. Multiple testing was controlled using Benjamini-Hochberg false discovery rate correction within prespecified test families. Prespecified sex by physiological reserve, sex by ASA, and sex-by-comorbidity interaction tests were performed as secondary explanatory analyses; estimator details are provided in the Supplementary Methods, and the sex-by-comorbidity screen is reported in Table S13.

Temporal-holdout analysis

The source pipeline was refit on 2018-2020 data and evaluated on held-out 2021-2022 data, restricting the temporal experiment to the period in which the POA-coded endpoint was ascertainable. All preprocessing, feature selection, and SMOTE were fitted using development-period data only. The temporal experiment was analyzed under the same subgroup bootstrap and estimability rules as the primary audit. Case mix was compared between periods, with additional reweighting analyses reported in the Supplementary Materials.

Analyses were implemented in Python using the *isitfair* package developed for this study. The manuscript and supplement follow TRIPOD+AI reporting principles for subgroup performance, calibration, and intended use.

Results

Outcome ascertainment across the cohort window

Outcome ascertainment was not constant across the cohort window, and characterizing it is necessary for interpreting prevalence-dependent results. The outcome requires a POA flag to distinguish infection acquired after surgery from infection present at operation. Present-on-admission coding is effectively absent from this extract before 2018 and enters routine use thereafter (Figure S2; Table S14). Where the flag is absent, infection timing cannot be established and such cases were excluded from cohort construction rather than retained as negative outcomes. The 2014-2017 portion of the cohort therefore contributes 4668 procedures and 3 coded events, compared with an 8.28% coded event rate in 2018-2022. The primary audit nevertheless retains the full 2014-2022 cohort, with the same POA-based endpoint definition throughout, because it corresponds to the published model and its original out-of-fold predictions; key prevalence-dependent results are additionally reported for the ascertainable 2018-2022 period (Supplementary Tables 15-17).

Ascertainment also differed across physiological-reserve strata within 2018-2022. Among cohort-eligible patients carrying a bacterial infection code, the proportion recorded as a coded postoperative case was 15.0% (95% CI, 7.1-29.1) in nonfrail, 24.2% (20.2-28.6) in prefrail, and 33.5% (30.9-36.3) in frail patients (Table S18). Present-on-admission field coverage was 92.5%-98.0% across the same strata.

Cohort and subgroup distribution

The audited cohort (Table 1) comprised 10 719 surgical procedures with 504 postoperative infection events (4.7%). The recorded sex was 34.4% female ($n = 3689$) and 65.6% male ($n = 7030$). Physiological reserve classified 6139 patients (57.3%) as frail, 3558 (33.2%) as prefrail, and 1022 (9.5%) as nonfrail. Infection prevalence increased monotonically with reserve loss (0.6%, 2.8%, and 6.5%, respectively). By recorded sex, cardiac procedures accounted for 41.7% of operations in men vs 22.7% in women (Cramér's $V = 0.31$); men were more often classified as frail (63.6% vs 45.2%) and ASA 4 (57.5% vs 38.8%), whereas women were over-represented at both age extremes (17.1% aged 0-17 or 80+ vs 10.6% of men).

Discrimination was similar by sex but heterogeneous by age and physiological reserve

Crude infection prevalence was similar between sexes and marginally higher in women (5.2% vs 4.5%). As shown in Table 1, recorded sex was nonetheless correlated with age, physiological reserve, and procedure cluster. Age and physiological reserve were the main audited axes with deployment-relevant performance differences, while procedure cluster served as a case-mix reference axis.

Overall discrimination was strong and essentially identical by sex (Figure 2A and D; Figure S3). AUROC was 0.876 (95% CI, 0.859-0.892) overall, 0.878 (0.849-0.903) in women and 0.875 (0.854-0.895) in men; the paired female-minus-male difference was +0.004 (−0.030 to +0.038). At $t = 0.10$, sensitivity was 0.900 in women and 0.898 in men, and specificity was 0.640 vs 0.650. Sex-specific gaps were small: equal opportunity 0.002, predictive equality 0.010, equalized odds 0.010, and statistical parity 0.014 (Table 2). Discrimination varied more by age than by sex (Figure 2B, Table 2). AUROC was 0.807 at ages 0-17 and 0.805 at ≥80 years, vs 0.888-0.889 at ages 45-79. At $t = 0.10$, specificity was lowest at the age extremes (0.462 and 0.577), vs 0.654-0.682 at ages 18-79. The age equalized-odds gap was 0.220, mainly from specificity differences, so a single global threshold flagged more noninfected patients at the age extremes.

Discrimination was lowest in prefrail patients (Figure 2C, Table 2); the nonfrail estimate was imprecise because this

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

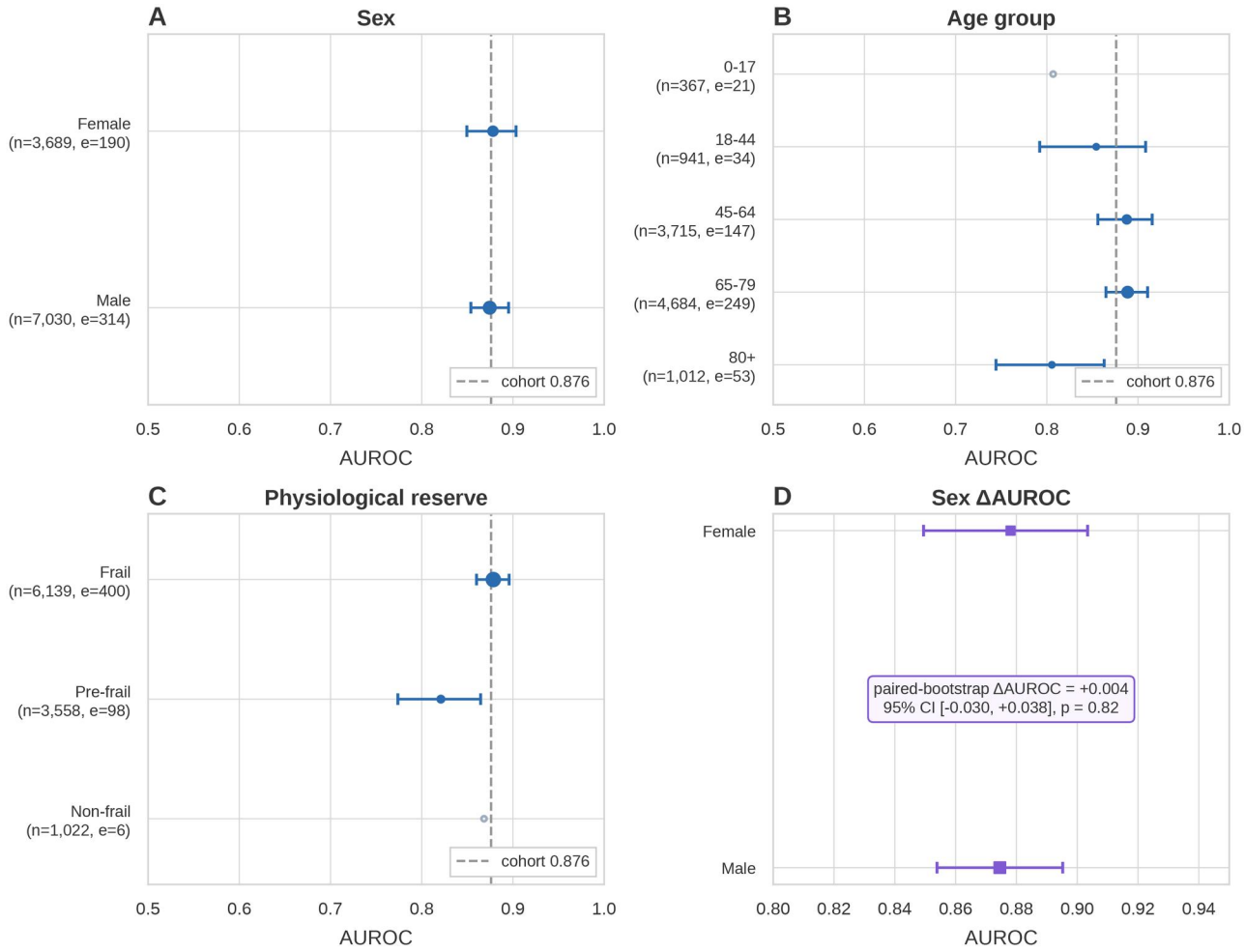

Figure 2. Subgroup discrimination. Panels A-C show within-subgroup AUROC estimates with 95% CIs from 2000 patient-level bootstrap resamples across (A) recorded sex, (B) age stratum, and (C) physiological reserve. Point sizes are proportional to within-subgroup event count; the vertical dashed line marks the overall cohort AUROC of 0.876 (95% CI 0.859-0.892). Estimates from strata below the prespecified 25-event AUROC floor, age 0-17 (21 events) and nonfrail (6 events), are drawn as hollow markers without an interval and are not interpreted inferentially. (D) Female and male AUROC estimates are shown on an expanded scale, with the paired-bootstrap sex contrast (female minus male: ΔAUROC +0.004, 95% CI −0.030 to +0.038; $P = .82$). Figure S3 shows the corresponding ROC curves.

stratum contained only 6 infection events. At $t = 0.10$, the share of true infections correctly flagged (sensitivity) was lowest in prefrail patients, where the model missed 21% of infections, compared with 7% in frail patients. The corresponding sensitivity gap (equal-opportunity difference) was 0.142, and the equalized-odds gap was 0.232. The procedure-cluster reference axis showed an even larger raw discrimination gap (equalized-odds gap 0.516; Table S19); 4 of the 12 procedure-cluster strata fell below the 25-event AUROC floor.

Calibration and clinical utility by subgroup

Calibration showed systematic overprediction of the coded endpoint across audited subgroups (Figure 1, Table 3). Calibration-in-the-large was −1.02 (95% CI, −1.12 to −0.94) overall, with greater overprediction across physiological-reserve strata: −0.86 (−0.96 to −0.76) in frail and −1.34 (−1.55 to −1.16) in prefrail patients, a difference of 0.49 (0.28-0.71). The nonfrail stratum contained only 6 events and was therefore not considered estimable. Sex calibration differences were small (maximum Δintercept 0.09, 0.00-0.28). This calibration pattern was reflected in clinical utility (Figure 1C and D). At $t = 0.10$, net benefit per 1000 patients was positive overall (+4.87, 95% CI, +0.81 to +8.84) and in frail patients (+17.21, +10.61 to +23.73), but negative in prefrail patients (−10.59, −15.77 to −5.37). Across the 7 thresholds from 0.05 to 0.20, prefrail net-benefit point estimates remained negative; the 95% CI included zero at 0.05 but excluded zero from 0.075 through 0.20 (Table S7). The nonfrail stratum remained below the prespecified 50-event floor for net-benefit inference.

Post hoc mitigation: recalibration ladder and no-SMOTE refit

Recalibration was evaluated as a three-step ladder of no recalibration, common cohort-wide recalibration, and group-specific recalibration, with all transformations cross-fitted (Table 4; Figure S1). Common recalibration corrected average calibration,

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

Table 2. Per-subgroup discrimination on the audited cohort (2014-2022; $n = 10\,719$; 504 postoperative infection events).

Subgroup	*n*	Events (%)	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Sex						
EOD 0.002, PE 0.010, EOG 0.010, SP 0.014						
Overall	10 719	504 (4.7)	0.876 (0.859-0.892)	0.376 (0.332-0.422)	0.899 (0.871-0.925)	0.647 (0.638-0.656)
Female	3689	190 (5.2)	0.878 (0.849-0.903)	0.408 (0.341-0.478)	0.900 (0.856-0.940)	0.640 (0.624-0.656)
Male	7030	314 (4.5)	0.875 (0.854-0.895)	0.356 (0.303-0.411)	0.898 (0.863-0.930)	0.650 (0.639-0.662)
Age group						
EOD 0.095, PE 0.220, EOG 0.220, SP 0.220						
Overall	10 719	504 (4.7)	0.876 (0.859-0.892)	0.376 (0.332-0.422)	0.899 (0.871-0.925)	0.647 (0.638-0.656)
0-17	367	21 (5.7)	0.807 (not estimable)	0.246 (not estimable)	0.857 (not estimable)	0.462 (not estimable)
18-44	941	34 (3.6)	0.854 (0.792-0.908)	0.329 (0.175-0.505)	0.824 (0.679-0.939)	0.682 (0.653-0.712)
45-64	3715	147 (4.0)	0.888 (0.856-0.916)	0.382 (0.303-0.468)	0.918 (0.872-0.959)	0.666 (0.650-0.682)
65-79	4684	249 (5.3)	0.889 (0.865-0.910)	0.438 (0.371-0.509)	0.912 (0.875-0.944)	0.654 (0.639-0.668)
80+	1012	53 (5.2)	0.805 (0.744-0.863)	0.227 (0.145-0.344)	0.849 (0.746-0.940)	0.577 (0.545-0.607)
Physiological reserve						
EOD 0.142, PE 0.232, EOG 0.232, SP 0.262						
Overall	10 719	504 (4.7)	0.876 (0.859-0.892)	0.376 (0.332-0.422)	0.899 (0.871-0.925)	0.647 (0.638-0.656)
Frail	6139	400 (6.5)	0.878 (0.860-0.896)	0.428 (0.379-0.481)	0.927 (0.901-0.951)	0.584 (0.571-0.597)
Prefrail	3558	98 (2.8)	0.821 (0.774-0.865)	0.190 (0.127-0.274)	0.786 (0.701-0.863)	0.702 (0.687-0.716)
Nonfrail	1022	6 (0.6)	0.868 (not estimable)	0.084 (not estimable)	0.833 (not estimable)	0.816 (not estimable)

Within-subgroup AUROC, AUPRC, sensitivity, and specificity at the $t = 0.10$ operating threshold, with 95% bootstrap CIs from 2000 patient-level resamples, subject to the prespecified event floors (see Methods). Gap statistics per axis are reported in the section header rows.
Abbreviations: EOD, equal-opportunity difference, the sensitivity gap; EOG, equalized-odds gap, the larger of the sensitivity and false-positive-rate gaps; PE, predictive-equality difference, the false-positive-rate gap; SP, statistical-parity difference, the difference in the proportion of patients flagged.

Table 3. Per-subgroup calibration on the audited cohort (2014-2022; $n = 10\,719$; 504 postoperative infection events).

Subgroup	*n*	Events (%)	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
Sex							
max ΔIntercept 0.09 (0.00, 0.28)							
max ΔSlope 0.01 (0.00, 0.29)							
Overall	10 719	504 (4.7)	−1.02 (−1.12, −0.94)	1.57 (1.45, 1.69)	0.041 (0.039, 0.043)	0.062	0.061
Female	3689	190 (5.2)	−0.96 (−1.11, −0.82)	1.56 (1.37, 1.76)	0.043 (0.040, 0.047)	0.061	0.06
Male	7030	314 (4.5)	−1.05 (−1.17, −0.94)	1.57 (1.42, 1.73)	0.040 (0.037, 0.043)	0.063	0.061
Age group							
max ΔIntercept 0.32 (not estimable)							
max ΔSlope 0.51 (not estimable)							
Overall	10 719	504 (4.7)	−1.02 (−1.12, −0.94)	1.57 (1.45, 1.69)	0.041 (0.039, 0.043)	0.062	0.061
0-17	367	21 (5.7)	−0.96 (not estimable)	1.54 (not estimable)	0.055 (not estimable)	0.073	0.071
18-44	941	34 (3.6)	−1.21 (−1.59, −0.91)	1.40 (not estimable)	0.036 (0.029, 0.044)	0.065	0.064
45-64	3715	147 (4.0)	−1.16 (−1.34, −1.01)	1.59 (1.39, 1.85)	0.036 (0.033, 0.040)	0.065	0.063
65-79	4684	249 (5.3)	−0.89 (−1.02, −0.77)	1.68 (1.50, 1.88)	0.042 (0.039, 0.046)	0.057	0.056
80+	1012	53 (5.2)	−1.07 (−1.38, −0.81)	1.16 (not estimable)	0.053 (0.046, 0.062)	0.074	0.075
Physiological reserve							
max ΔIntercept 1.62 (not estimable)							
max ΔSlope 0.31 (not estimable)							
Overall	10 719	504 (4.7)	−1.02 (−1.12, −0.94)	1.57 (1.45, 1.69)	0.041 (0.039, 0.043)	0.062	0.061
Frail	6139	400 (6.5)	−0.86 (−0.96, −0.76)	1.59 (1.46, 1.74)	0.052 (0.049, 0.055)	0.064	0.062
Nonfrail	1022	6 (0.6)	−2.47 (not estimable)	1.55 (not estimable)	0.011 (not estimable)	0.055	0.055
Prefrail	3558	98 (2.8)	−1.34 (−1.55, −1.16)	1.28 (not estimable)	0.031 (0.028, 0.034)	0.062	0.062

Within-subgroup calibration intercept (calibration-in-the-large; fixed-slope logit recalibration), calibration slope, Brier score, ECE, and ICI are reported. Confidence intervals were estimated from 2000 patient-level bootstrap resamples. Prespecified event floors for inferential reporting were 25 events for the calibration intercept and 100 events for the calibration slope. Point estimates below these floors are shown descriptively and labeled not estimable rather than assigned an inferential confidence interval. Maximum subgroup gaps are likewise labeled not estimable when either contributing stratum fails the corresponding event floor.

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

Table 4. Recalibration ladder: calibration and clinical utility by physiological-reserve stratum, audited cohort ($n = 10\,719$; 504 postoperative infection events).

Step	Reserve stratum	Events	Calibration-in-the-large (95% CI)	Slope	ECE	ICI	Brier	NB at $t = 0.10$ per 1000 (95% CI)
No recalibration	Overall	504	−1.020 (−1.116 to −0.938)	1.566	0.062	0.061	0.041	+4.87 (+0.81 to +8.84)
No recalibration	Frail	400	−0.855 (−0.959 to −0.757)	1.594	0.064	0.062	0.052	+17.21 (+10.61 to +23.73)
No recalibration	Prefrail	98	−1.342 (−1.554 to −1.158)	1.282	0.062	0.062	0.031	−10.59 (−15.77 to −5.37)
No recalibration	Nonfrail	6	−2.475 (not estimable)	1.546	0.055	0.055	0.011	−15.44 (not estimable)
Common (cohort-wide)	Overall	504	−0.007 (−0.116 to 0.091)	0.928	0.003	0.003	0.036	+21.76 (+18.42 to +25.06)
Common (cohort-wide)	Frail	400	0.100 (−0.023 to 0.217)	0.918	0.007	0.006	0.047	+34.99 (+29.57 to +40.47)
Common (cohort-wide)	Prefrail	98	−0.206 (−0.438 to 0.002)	0.827	0.01	0.008	0.025	+5.37 (+1.72 to +9.27)
Common (cohort-wide)	Nonfrail	6	−1.105 (not estimable)	0.974	0.011	0.011	0.006	−0.65 (not estimable)
Group-specific (by reserve)	Overall	504	−0.009 (−0.119 to 0.090)	0.85	0.005	0.004	0.036	+21.37 (+18.07 to +24.65)
Group-specific (by reserve)	Frail	400	−0.006 (−0.127 to 0.110)	0.881	0.005	0.005	0.047	+34.41 (+28.90 to +40.14)
Group-specific (by reserve)	Prefrail	98	−0.009 (−0.239 to 0.198)	0.794	0.006	0.005	0.025	+5.37 (+2.06 to +8.96)
Group-specific (by reserve)	Nonfrail	6	−0.146 (not estimable)	0.119	0.008	0.008	0.007	−1.20 (not estimable)

Calibration and clinical utility are shown with no recalibration, common cohort-wide isotonic recalibration, and group-specific isotonic recalibration by physiological reserve. Both recalibration transformations were cross-fitted on the published cross-validation folds. Prespecified event floors were retained for inferential reporting. The nonfrail stratum contains 6 events and is therefore below the floors for calibration-intercept, calibration-slope, and net-benefit inference. The prefrail stratum contains 98 events and is below the 100-event calibration-slope floor. Point estimates below these floors are shown descriptively but are not interpreted inferentially.

moving overall calibration-in-the-large from −1.020 to −0.007 (95% CI, −0.116 to 0.091), while the frail-prefrail intercept gap remained at 0.308 (0.069-0.554). Group-specific recalibration reduced this gap to 0.005 (−0.232 to 0.243). However, the residual gap after common recalibration was similar in magnitude to the estimated upper bound attributable to differential endpoint ascertainment in prefrail patients (0.327 logits; Table S18), and the nonfrail stratum was not estimable because it contained only 6 events. Separately, the nonfrail group–specific isotonic transform collapsed to a nearly constant mapping: its post-recalibration calibration slope was 0.119, and the fitted map assigned the 1022 nonfrail patients to 17 distinct probabilities (Figure S1). Because this stratum contained only 6 events and fell below every prespecified estimability floor, this diagnostic result is not interpreted inferentially.

A no-SMOTE refit was used to assess whether oversampling accounted for the observed miscalibration (Table S2). The reproduction arm recovered all 10 719 published out-of-fold predictions bit-identically. In the 2018-2022 evaluation window, removing SMOTE shifted overall calibration-in-the-large substantially but did not reduce the reserve gradient: the frail-prefrail intercept gap increased rather than decreased.

Intersectional and secondary interaction analyses

The secondary analyses were the prespecified model-based interaction tests: (1) sex by physiological reserve, (2) sex by ASA, and (3) a sex-by-comorbidity screen, allowing to examine whether covariate-outcome associations differed by sex. Pairwise intersectional analyses were limited by sparse cells (Table S12), but the patterns mirrored the age and physiological-reserve findings. Equalized-odds gaps were largest for age by physiological reserve. Intersectional analyses did not identify an additional pattern beyond the prespecified age and physiological-reserve margins.

The sex-by-physiological-reserve interaction term for infection was statistically significant but small (likelihood-ratio $P = .018$, $\Delta R^2 = 0.0008$). As Table 1 shows, sex was associated with physiological reserve and case mix, but infection prevalence itself was similar between sexes (5.2% vs 4.5%).

Secondary sex-stratified analyses are shown in Figure 3 and Table S13. Men and women differed substantially in procedure mix and baseline acuity (Figure 3A). However, comorbidity-specific infection associations did not show a consistent sex-specific direction (Figure 3B), and within-cluster AUROC differences varied in both directions rather than indicating systematic degradation for either sex (Figure 3C). Infection risk increased with ASA physical status in both sexes, with a statistically significant but small sex-by-ASA interaction (Figure 3D).

Temporal-holdout analysis

Refitting on 2018-2020 data (3847 procedures; 289 events) and evaluating the held-out 2021-2022 period (2204 procedures; 212 events) provided an out-of-period assessment (Table 5, Figure 4). AUROC was 0.875 (95% CI, 0.852-0.895) in development and 0.851 (0.823-0.876) in the holdout. The direction of overprediction persisted, with calibration-in-the-large of −0.881 (−1.003 to −0.768) and −0.711 (−0.851 to −0.574), respectively. The physiological-reserve calibration gradient persisted at reduced magnitude. Frail net benefit remained positive in both periods. Prefrail net benefit was −6.94 per 1000 in development, where 44 events were below the prespecified 50-event floor for an interval, and +14.72 per 1000 (−3.49 to +34.00) in the holdout. The nonfrail stratum was too sparse for interval-based inference, and other strata below the prespecified event floors were treated similarly.

A supplementary reproducibility analysis on the public SUPPORT2 dataset[44] is described in the Supplementary Materials.

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

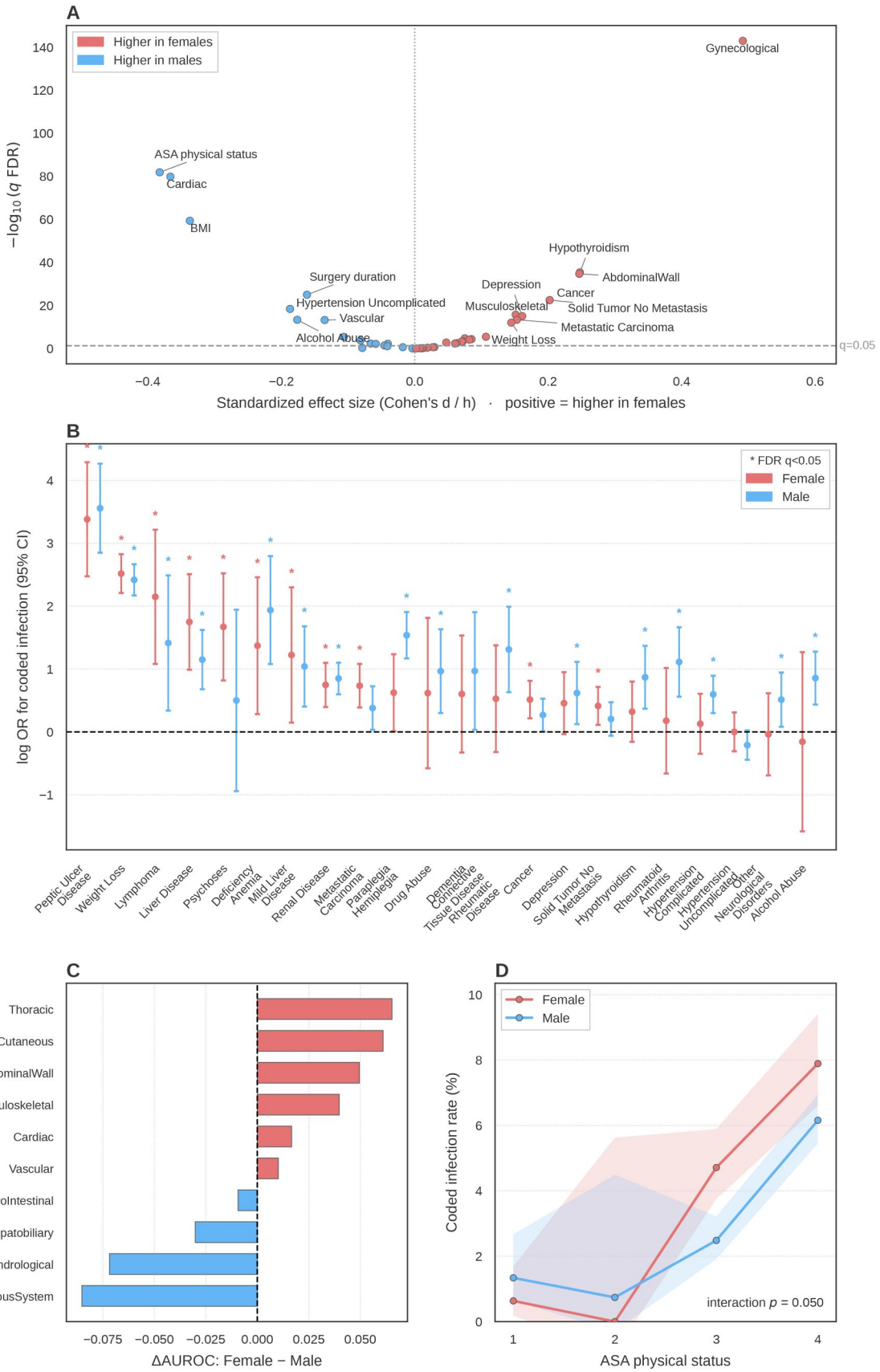

Figure 3. Targeted characterization of the sex axis. (A) Feature-by-sex effect-size volcano: standardized effect size (Cohen's *d* for continuous features, Cohen's *h* for binary features) against $-\log_{10}$(FDR-corrected *q*) for every candidate feature; positive effect = higher value or prevalence in females,

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

Discussion

This audit identified deployment-relevant subgroup heterogeneity that would have been missed by discrimination-only or sex-marginal evaluation. The principal finding was systematic overprediction with a physiological-reserve calibration gradient, accompanied by differences in clinical utility in the source-cohort audit. Because outcome ascertainment differed over time and across reserve strata, prevalence-dependent results are interpreted as performance against the POA-coded endpoint rather than true infection incidence. These findings support joint evaluation of subgroup discrimination, calibration, clinical utility, and relevant fairness criteria when assessing deployment readiness.

Discrimination, calibration, and clinical utility

Discrimination heterogeneity was relevant but less dominant than calibration. AUROC fell at both age extremes and in prefrail patients (Figure 2, Table 2). The larger raw gap on the procedure-cluster reference axis (Table S19) mainly reflects differing baseline infection risk across operations and sparse clusters, rather than a deployment-relevant subgroup performance failure. Because event prevalence differed across strata, AUROC alone can be misleading; per-subgroup results should be interpreted alongside prevalence, not ranking metrics alone.[45] A single operating threshold can behave differently when ranking performance and prevalence differ, and sparse-stratum threshold estimates are unstable. A safer deployment strategy is to recalibrate probabilities within clinically justified strata, then prospectively define subgroup-specific criteria before use.[21,22,43,46]

Calibration was critical because the end-of-surgery infection-risk score is intended to be used as an individualized probability at a predefined monitoring threshold. The physiological-reserve-stratified failure was mainly calibration-in-the-large, showing why subgroup intercepts and slopes should accompany binned and curve-based summaries.[21,33–36] Decision-curve analysis translated this miscalibration into deployment-relevant differences in clinical utility.[37,47–49] In the source-cohort audit, net benefit at $t = 0.10$ was positive in frail patients and negative in prefrail patients, whereas the nonfrail stratum was below the prespecified event floor for net-benefit inference. Because negative prefrail utility did not persist in the ascertainable-period and temporal analyses, it should not be interpreted as a stable subgroup effect.

Mitigation, leveling down, and practical deployment

Mitigation was evaluated under a prespecified leveling-down criterion, defined here as an apparent parity improvement that reduces subgroup gaps by worsening absolute performance in one or more subgroups. Smaller parity gaps are not desirable if they arise from degraded subgroup performance. Isotonic recalibration and Wasserstein-barycenter postprocessing are monotonic within subgroups and therefore largely preserve rank ordering,[40] but optimize different objectives. Isotonic recalibration corrects predicted-vs-observed mismatch, whereas Wasserstein postprocessing equalizes score distributions and can reduce a parity gap while worsening calibration (Table 4, Table S20). Thus, isotonic recalibration matched the observed calibration-driven failure. For deployment, it should be considered only when a subgroup has clear calibration error (intercept or slope materially departing from 0 or 1 with CIs excluding the target) and enough calibration-set events to fit the transform. In sparse strata, average calibration correction can coexist with unstable slopes, so parametric or hierarchical recalibration may be safer.[19,39,50–52] Neither postprocessing approach can repair true discrimination loss. The weaker separation in the youngest, oldest, and prefrail patients requires threshold review and larger validation cohorts.[21,33] The original model used SMOTE within training folds, which can shift predicted probabilities in low-prevalence strata.[53] In the 2018-2022 ascertainment-window comparison, removing SMOTE shifted overall calibration-in-the-large but did not reduce the physiological-reserve gradient; the frail-prefrail intercept gap increased rather than decreased (Table S2). Thus, SMOTE accounts for a substantial common calibration offset but does not explain the between-reserve calibration difference. The deployment implication is unchanged: probabilities require direct subgroup evaluation rather than inference from discrimination alone. Throughout these mitigation experiments the underlying prediction model was held fixed, and only post hoc transformations of its output probabilities were evaluated; at deployment, the model would therefore return an individual risk probability, optionally passed through a fixed, prespecified recalibration transform.

Intersectional and sex-composition findings

Pairwise intersectional analyses added little beyond the single-axis findings: the gaps mainly tracked the age and physiological-reserve margins, and the sex by physiological-reserve result reinforced the same point. Recorded sex was associated with

Figure 3. Continued
with direction indicated in the legend; the dashed line marks $q = 0.05$, and features with $|\text{effect}| \geq 0.15$ or strong significance are labeled. (B) Per-comorbidity odds ratio for postoperative bacterial infection by sex on the log scale with 95% CI, estimated separately within females and males as indicated in the legend; the 23 screened comorbidities are ordered by descending female OR, and * marks an FDR-significant odds ratio within that sex ($q < 0.05$); the sex × comorbidity interaction tests are reported separately in Table S13. (C) Per-cluster AUROC gap: difference in model discrimination (female − male AUROC) within each surgical procedure cluster from out-of-fold predictions, positive = better discrimination in females; the 10 clusters with both sexes represented are shown, excluding the Gynecological cluster (no male patients) and the unclassified bucket. (D) Sex × ASA interaction: observed postoperative infection rate by ASA physical-status score stratified by sex, with 95% CI bands and the likelihood-ratio interaction *P*-value ($P = .050$).

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

Table 5. Temporal-holdout analysis by subgroup.

Axis	Stratum	*n* (dev/ holdout)	Events (dev/ holdout)	AUROC, development	AUROC, holdout	Calibration-in-the-large, development	Calibration-in-the-large, holdout	Net benefit at *t* = 0.10 per 1000, development	Net benefit at *t* = 0.10 per 1000, holdout
Overall	Overall	3847/2204	289/212	0.875 (0.852 to 0.895)	0.851 (0.823 to 0.876)	−0.881 (−1.003 to −0.768)	−0.711 (−0.851 to −0.574)	+19.44 (+11.47 to +28.91)	+36.55 (+23.49 to +49.46)
Recorded sex	Female	1295/795	112/77	0.886 (0.850 to 0.917)	0.819 (0.765 to 0.866)	−0.761 (−0.955 to −0.579)	−0.779 (−1.021 to −0.547)	+33.46 (+17.24 to +50.11)	+33.40 (+12.30 to +56.46)
	Male	2552/1409	177/135	0.867 (0.839 to 0.895)	0.870 (0.841 to 0.899)	−0.949 (−1.108 to −0.806)	−0.672 (−0.850 to −0.513)	+12.32 (+2.22 to +22.60)	+38.32 (+22.24 to +55.13)
Age stratum	0-17	141/82	11/10	0.761 (not estimable)	0.890 (not estimable)	−0.872 (not estimable)	−0.243 (not estimable)	−1.58 (not estimable)	+63.69 (not estimable)
	18-44	317/201	21/13	0.888 (not estimable)	0.814 (not estimable)	−0.959 (not estimable)	−0.914 (not estimable)	+21.38 (not estimable)	+12.71 (not estimable)
	45-64	1311/738	72/74	0.882 (0.842 to 0.920)	0.849 (0.804 to 0.892)	−1.177 (−1.436 to −0.962)	−0.622 (−0.878 to −0.394)	+3.05 (−9.83 to +15.60)	+44.41 (+21.68 to +66.10)
	65-79	1726/972	154/93	0.889 (0.858 to 0.916)	0.873 (0.830 to 0.909)	−0.695 (−0.865 to −0.548)	−0.743 (−0.973 to −0.541)	+33.47 (+19.76 to +47.57)	+35.78 (+16.57 to +55.33)
	80+	352/211	31/22	0.783 (0.705 to 0.855)	0.786 (not estimable)	−0.883 (−1.308 to −0.520)	−0.891 (not estimable)	+18.31 (not estimable)	+24.75 (not estimable)
Physiological reserve	Frail	2401/1283	242/157	0.873 (0.848 to 0.897)	0.855 (0.821 to 0.886)	−0.731 (−0.875 to −0.610)	−0.604 (−0.768 to −0.436)	+37.35 (+24.99 to +50.35)	+56.38 (+38.19 to +76.04)
	Nonfrail	358/189	3/3	0.907 (not estimable)	0.909 (not estimable)	−2.417 (not estimable)	−1.808 (not estimable)	−20.48 (not estimable)	−13.52 (not estimable)
	Prefrail	1088/732	44/52	0.817 (0.745 to 0.876)	0.801 (0.740 to 0.854)	−1.254 (−1.610 to −0.973)	−0.866 (−1.179 to −0.595)	−6.94 (not estimable)	+14.72 (−3.49 to +34.00)

The source modeling pipeline was refit on the 2018-2020 development period (3847 procedures; 289 events) and evaluated on the held-out 2021-2022 period (2204 procedures; 212 events), using the same model architecture, hyperparameters, feature pipeline, and class-rebalancing procedure as the published model. All preprocessing, feature selection, and SMOTE were fitted using development-period data only. Results are shown for the prespecified subgroup axes; case-mix reference axes and pairwise intersections are reported in the Supplementary Materials. Estimates are shown with 95% CIs from 2000 patient-level bootstrap resamples under the same subgroup resampling and estimability rules as the primary audit. Prespecified event floors for inferential reporting were 25 events for AUROC and calibration-in-the-large and 50 events for net benefit. Estimates below these floors are shown descriptively and labeled not estimable, without inferential CIs. Interpretation of differences between periods is limited by the smaller development sample and by temporal differences in endpoint ascertainment. The proportion of eligible infected patients recorded as postoperative cases was 28.8% in development and 34.0% in the holdout; consequently, temporal changes in calibration and net benefit should not be interpreted as evidence of model improvement or deterioration alone. Calibration and net-benefit estimates describe performance against the POA-coded endpoint rather than true infection incidence.

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

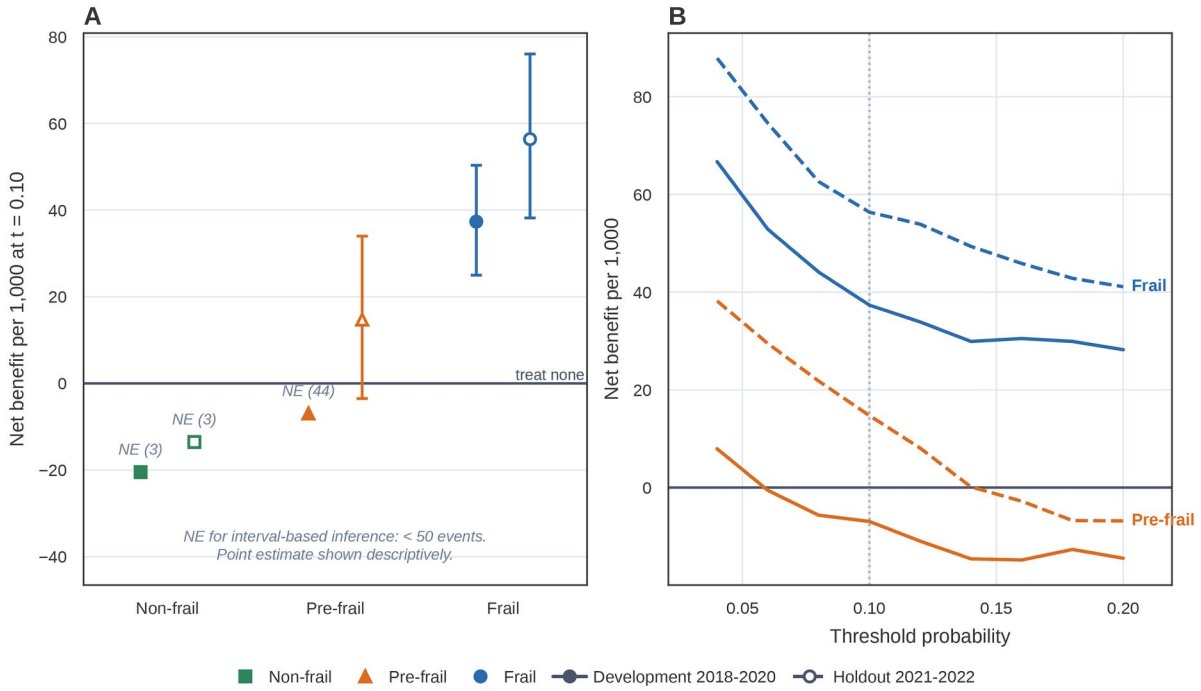

Figure 4. Subgroup net benefit by period, temporal-holdout analysis. The source pipeline was refit on the 2018-2020 development period (3847 procedures; 289 events) and evaluated on the held-out 2021-2022 period (2204 procedures; 212 events). (A) Net benefit per 1000 patients at the $t = 0.10$ operating threshold by physiological-reserve stratum, with 95% patient-level bootstrap intervals. Color and marker shape carry the stratum; filled markers are the development period and open markers are the holdout. Estimates below the prespecified 50-event floor for net-benefit inference are annotated NE with their event count and shown as point estimates without an interval: nonfrail in both periods (3 events each) and prefrail in development (44 events). (B) Net benefit across threshold probabilities from 0.04 to 0.20, solid for development and dashed for the holdout, with the dotted vertical line at $t = 0.10$; the nonfrail stratum is omitted from this panel as not estimable at any threshold. Zero is the treat-none reference in both panels. Interpretation is limited by the smaller development sample and by differential endpoint ascertainment between periods: the proportion of eligible infected patients recorded as postoperative cases was 28.8% in development and 34.0% in the holdout, so between-period differences should not be read as model improvement or deterioration alone.

physiological reserve and case mix, but sex-marginal performance was similar. An apparently reassuring sex-only audit can therefore conceal composition-mediated harm when sex correlates with the axis on which the model fails.[9,10] In this audit, common cohort-wide recalibration was favored for correcting the average calibration offset (Table 4). Although group-specific recalibration further reduced the frail-prefrail calibration-intercept difference, the residual difference after common recalibration was of the same magnitude as the upper bound attributable to differential endpoint ascertainment (Table S18), so we defer group-specific recalibration until ascertainment is comparable across reserve strata. Recorded sex should not be used as a proxy mitigation axis.[23,54]

Related work

These findings extend previous clinical AI fairness work showing that aggregate calibration can coexist with subgroup deterioration[9] and that postoperative prediction models require explicit fairness evaluation.[55] They also support a practical distinction between parity-focused and calibration-focused mitigation: threshold-based parity criteria remain useful for describing error-rate gaps,[31,32] but safe deployment requires predicted probabilities that match observed infection rates within each subgroup, not only balanced yes/no decisions.

Equity and representation

Subgroup performance auditing for deployment is also motivated by concerns about equity. Real-world EHR data can reveal deployment risks in the patient populations seen in practice, but equity-oriented evaluation requires systematic capture of protected and socially defined attributes. The absence of these variables limits the present audit, although these variables can be incorporated when they are available.

Limitations

Several limitations should guide interpretation. The primary cohort came from a single Swiss tertiary center. The supplementary SUPPORT2 analysis is a software reproducibility example, not external validation of the infection model. Tertiary referral likely enriched complex, higher-acuity cases, increasing the frail stratum and shrinking preserved-reserve strata. The nonfrail stratum contained only 6 infection events and therefore fell below the prespecified event floors for calibration and net-benefit inference; estimates from this stratum are shown descriptively

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

but are not interpreted inferentially. Sex was administratively recorded as binary and did not capture gender identity. Race, ethnicity, migrant status, language, socioeconomic disadvantage, and rurality were unavailable. Systematic capture of socially defined attributes is required for equitable prediction. The physiological-reserve construct is an EHR-derived proxy, not a validated frailty instrument, and depends partly on diagnosis coding; the mFI-4 hypertension item used ICD-10 codes without medication linkage, potentially overcounting hypertension. Postdischarge infections may be missed by the index admission ICD-10-GM outcome. Scores were out-of-fold cross-validation predictions rather than outputs from a locked deployment artifact, and mitigation was limited to post hoc score transformations rather than model retraining or prospective workflow redesign.

Outcome ascertainment is incomplete and differs across physiological-reserve strata; prevalence-dependent performance therefore describes fit to the POA-coded endpoint rather than true infection incidence. Present-on-admission field coverage was high in every reserve stratum (92.5%-98.0%; Table S18), so the differential is not explained by field availability alone; it may partly reflect genuine differences in infection susceptibility and timing across strata rather than documentation practice, and it is therefore treated as a property of the coded endpoint rather than attributed to coding behavior. The model was developed on pooled 2014-2022 data that included a pre-2018 period with outcome-dependent exclusion of infection cases lacking adequate POA adjudication. Temporal comparisons are additionally affected by documentation changes, including procedure-code completeness, and should not be interpreted as causal evidence of model drift. The temporal-holdout experiment provides out-of-period evidence but is not an external validation of the published fitted model. Prospective post-2022 validation with comparable outcome ascertainment across subgroups remains necessary before clinical deployment.

Conclusion

Strong average discrimination and reassuring sex-marginal metrics are insufficient evidence for deployment readiness across subgroups. In this audit, joint evaluation of discrimination, calibration, clinical utility, and formal fairness criteria identified subgroup-specific limitations that aggregate evaluation would have missed. Such audits can support fairness evaluation by identifying where model performance and clinical consequences differ across clinically relevant groups. More broadly, this audit illustrates a transparent approach to identifying deployment-relevant subgroup limitations, comparing mitigation strategies, and assessing deployment readiness in clinical prediction models.

Acknowledgments

We thank the members of the INFRA project, in particular Stefanie Marti. We thank Guy Maddern and Markus Trochsler for their valuable input.

Author contributions

Hugo Guillen-Ramirez (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Katharina Lucia Lucas (Investigation, Validation, Writing—review & editing), Yves Max Wintsch (Investigation, Validation, Writing—review & editing), Tobias Ueli Blatter (Data curation, Methodology, Writing—review & editing), Karen Triep (Conceptualization, Data curation, Writing—review & editing), Olga Endrich (Conceptualization, Data curation, Funding acquisition, Writing—review & editing), and Guido Beldi (Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing—review & editing)

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This research was supported by Swiss Personalized Health Network (SPHN) grant DEM-2022-08, the Multidisciplinary Center for Infectious Diseases (MCID) MA_18 of the University of Bern, and the Commission for Digitalization of the University of Bern.

Conflicts of interest

The authors declare no competing interests.

Data availability

The Bern cohort contains patient-level electronic health record data and cannot be shared publicly. The *isitfair* package, aggregated subgroup outputs, analysis code, package versions, and the SUPPORT2 reproducibility example are available at https://github.com/HugoGuillen/isitfair.

Ethics approval

The study was approved by the Cantonal Ethics Committee of Bern (KEK-ID: 2021-00965), with a waiver of informed consent due to the study's retrospective design

References

1. Badia JM, Casey AL, Petrosillo N, et al. Impact of surgical site infection on healthcare costs and patient outcomes: a systematic review in six European countries. *J Hosp Infect*. 2017;96:1-15. https://doi.org/10.1016/j.jhin.2017.03.004
2. O'Brien WJ, Gupta K, Itani KMF. Association of postoperative infection with risk of long-term infection and mortality. *JAMA Surg*. 2020;155:61. https://doi.org/10.1001/jamasurg.2019.4539

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

3. Dencker EE, Bonde A, Troelsen A, et al. Postoperative complications: an observational study of trends in the United States from 2012 to 2018. *BMC Surg*. 2021;21:393. https://doi.org/10.1186/s12893-021-01392-z
4. Adamina M, Steffen T, Tarantino I, et al. Meta-analysis of the predictive value of C-reactive protein for infectious complications in abdominal surgery. *Br J Surg*. 2015;102:590-598. https://doi.org/10.1002/bjs.9756
5. Guillen-Ramirez H, Sanchez-Taltavull D, Perrodin S, et al. Prediction of postoperative infections by strategic data imputation and explainable machine learning. *J Am Med Inform Assoc*. 2025;32:1706-1717. https://doi.org/10.1093/jamia/ocaf145
6. Blatter TU, Wintsch Y, Triep K, et al. End-of-surgery prediction of postoperative infectious complications from intraoperative vital-sign dynamics. *NPJ Digit Med*. 2026;9. https://doi.org/10.1038/s41746-026-02707-1
7. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447-453. https://doi.org/10.1126/science.aax2342
8. Seyyed-Kalantari L, Zhang H, McDermott MBA, et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27:2176-2182. https://doi.org/10.1038/s41591-021-01595-0
9. Davis SE, Dorn C, Park DJ, et al. Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. *J Am Med Inform Assoc*. 2025;32:845-854. https://doi.org/10.1093/jamia/ocaf039
10. Chakradeo K, Huynh I, Balaganeshan SB, et al. Navigating fairness aspects of clinical prediction models. *BMC Med*. 2025;23:567. https://doi.org/10.1186/s12916-025-04340-3
11. Gilbert T, Neuburger J, Kraindler J, et al. Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study. *Lancet*. 2018;391:1775-1782. https://doi.org/10.1016/S0140-6736(18)30668-8
12. Al-Sahab B, Leviton A, Loddenkemper T, et al. Biases in electronic health records data for generating real-world evidence: an overview. *J Healthc Inform Res*. 2024;8:121-139. https://doi.org/10.1007/s41666-023-00153-2
13. Abràmoff MD, Tarver ME, Loyo-Berrios N, et al.; Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, D.C. Considerations for addressing bias in artificial intelligence for health equity. *NPJ Digit Med*. 2023;6:170. https://doi.org/10.1038/s41746-023-00913-9
14. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. *PLOS Digit Health.* 2024;3:e0000651. https://doi.org/10.1371/journal.pdig.0000651
15. Arora A, Alderman JE, Palmer J, et al. The value of standards for health datasets in artificial intelligence-based applications. *Nat Med*. 2023;29:2929-2938. https://doi.org/10.1038/s41591-023-02608-w
16. Seki T, Kawazoe Y, Takiguchi T, et al. Sex differences in postnoncardiac surgery risks assessed using the revised cardiac risk index - a nationwide retrospective cohort study. *Circ J*. 2025;89:1547-1554. https://doi.org/10.1253/circj.CJ-24-0846
17. Wang H, Luo M, Yang Y, et al. Gender differences in postoperative pain, sleep quality, and recovery outcomes in patients undergoing visual thoracoscopic surgery. *Heliyon.* 2024;10:e39015. https://doi.org/10.1016/j.heliyon.2024.e39015
18. Bozic C, Strypet M, Mansvelder FJ, et al. Sex differences in preoperative risk profiles and 1-year mortality following elective cardiac surgery: a retrospective single-centre cohort study. *J Clin Med*. 2025;15:59. https://doi.org/10.3390/jcm15010059
19. Barda N, Yona G, Rothblum GN, et al. Addressing bias in prediction models by improving subpopulation calibration. *J Am Med Inform Assoc*. 2021;28:549-558. https://doi.org/10.1093/jamia/ocaa283
20. Ibrahim AM, Ghaferi AA, Thumma JR, et al. Variation in outcomes at bariatric surgery centers of excellence. *JAMA Surg*. 2017;152:629-636. https://doi.org/10.1001/jamasurg.2017.0542
21. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378
22. Benitez-Aurioles J, Joules A, Brusini I, et al. Understanding algorithmic fairness for clinical prediction in terms of subgroup net benefit and health equity. *Epidemiology*. 2026;37:386-396. https://doi.org/10.1097/EDE.0000000000001949
23. Naderalvojoud B, Curtin C, Asch SM, et al. Evaluating the impact of data biases on algorithmic fairness and clinical utility of machine learning models for prolonged opioid use prediction. *JAMIA Open.* 2025;8:ooaf115. https://doi.org/10.1093/jamiaopen/ooaf115
24. Dakshit S, Dakshit S, Khargonkar N, et al. Bias Analysis in Healthcare Time Series (BAHT) decision support systems from meta data. *J Healthc Inform Res*. 2023;7:225-253. https://doi.org/10.1007/s41666-023-00133-6
25. Zong Y, Yang Y, Hospedales T. MEDFAIR: benchmarking fairness for medical imaging. arXiv, arXiv:2210.01725, 2023. https://doi.org/10.48550/arXiv.2210.01725
26. Dou Q, Jin R, Li X, et al. FairMedFM: Fairness Benchmarking for Medical Imaging Foundation Models. In: *Advances in Neural Information Processing Systems 37 (NeurIPS)*. Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc.; 2024:111318-111357.
27. Bird S, Dudík M, Edgar R, et al. *Fairlearn: A Toolkit for Assessing and Improving Fairness in AI*. Microsoft; 2020. https://www.microsoft.com/en-us/research/wp-content/uploads/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf
28. Bellamy RKE, Dey K, Hind M, et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. 2018.
29. Subramaniam S, Aalberg JJ, Soriano RP, et al. New 5-factor modified frailty index using american college of surgeons NSQIP data. *J Am Coll Surg*. 2018;226:173-181.e8. https://doi.org/10.1016/j.jamcollsurg.2017.11.005
30. Guillen-Ramirez H, Triep K, Gaudet-Blavignac C, et al. *MAP-CARE: Enhancing Cross-Lingual Medical Intervention Terms Analysis Through LLM-supported Semantic Embeddings*. 2025.

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

31. Hardt M, Price E, Srebro N. *Equality of Opportunity in Supervised Learning*. 2016.
32. Corbett-Davies S, Pierson E, Feller A, et al. Algorithmic Decision Making and the Cost of Fairness. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM; 2017: 797-806.
33. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128-138. https://doi.org/10.1097/EDE.0b013e3181c30fb2
34. Van Calster B, McLernon DJ, van Smeden M, et al.; On behalf of Topic Group 'Evaluating Diagnostic Tests And Prediction Models' of the STRATOS Initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7
35. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med*. 2019;38:4051-4065. https://doi.org/10.1002/sim.8281
36. Huang Y, Li W, Macheret F, et al. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27:621-633. https://doi.org/10.1093/jamia/ocz228
37. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006; 26:565-574. https://doi.org/10.1177/0272989X06295361
38. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35:2052-2056. https://doi.org/10.1097/01.CCM.0000275267.64078.B0
39. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International conference on Machine Learning - ICML '05*. Bonn, Germany: ACM Press; 2005:625-32.
40. Chzhen E, Denis C, Hebiri M, et al. Fair regression with Wasserstein Barycenters. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Curran Associates, Inc.; 2020:7321–7331.
41. del Barrio E, Gamboa F, Gordaliza P, et al. *Obtaining fairness Using Optimal Transport Theory*. 2018.
42. Jiang R, Pacchiano A, Stepleton T, et al. Wasserstein fair classification. In: *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference. Proceedings of Machine Learning Research*, vol. 115. PMLR; 2020:862-872. https://proceedings.mlr.press/v115/jiang20a.html
43. Pleiss G, Raghavan M, Wu F, et al. On fairness and calibration. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc.; 2017.
44. Knaus W, Lynn J. *Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment (SUPPORT) and Hospitalized Elderly Longitudinal Project (HELP), 1989-1997: Version 2*; 2001.
45. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432
46. Muralidharan V, Burgart A, Daneshjou R, et al. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. *NPJ Digit Med*. 2023;6:166. https://doi.org/10.1038/s41746-023-00898-5
47. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making*. 2015;35:162-169. https://doi.org/10.1177/0272989X14547233
48. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6. https://doi.org/10.1136/bmj.i6
49. Tamma PD, Avdic E, Li DX, et al. Association of adverse events with antibiotic use in hospitalized patients. *JAMA Intern Med*. 2017;177:1308-1315. https://doi.org/10.1001/jamainternmed.2017.1938
50. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. 2021;40:4230-4251. https://doi.org/10.1002/sim.9025
51. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016;35:214-226. https://doi.org/10.1002/sim.6787
52. Steyerberg EW, Borsboom GJJM, Van Houwelingen HC, et al. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004;23:2567-2586. https://doi.org/10.1002/sim.1844
53. Van Den Goorbergh R, Van Smeden M, Timmerman D, et al. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc*. 2022;29:1525-1534. https://doi.org/10.1093/jamia/ocac093
54. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform*. 2021;113:103621. https://doi.org/10.1016/j.jbi.2020.103621
55. Davoudi A, Sajdeya R, Ison R, et al. Fairness in the prediction of acute postoperative pain using machine learning models. *Front Digit Health*. 2022;4:970281. https://doi.org/10.3389/fdgth.2022.970281

Downloaded from academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocag184/9005654 by guest on 02 November 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Journal of the American Medical Informatics Association, 2026, 00, 1–16
https://doi.org/10.1093/jamia/ocag184
Research and Applications