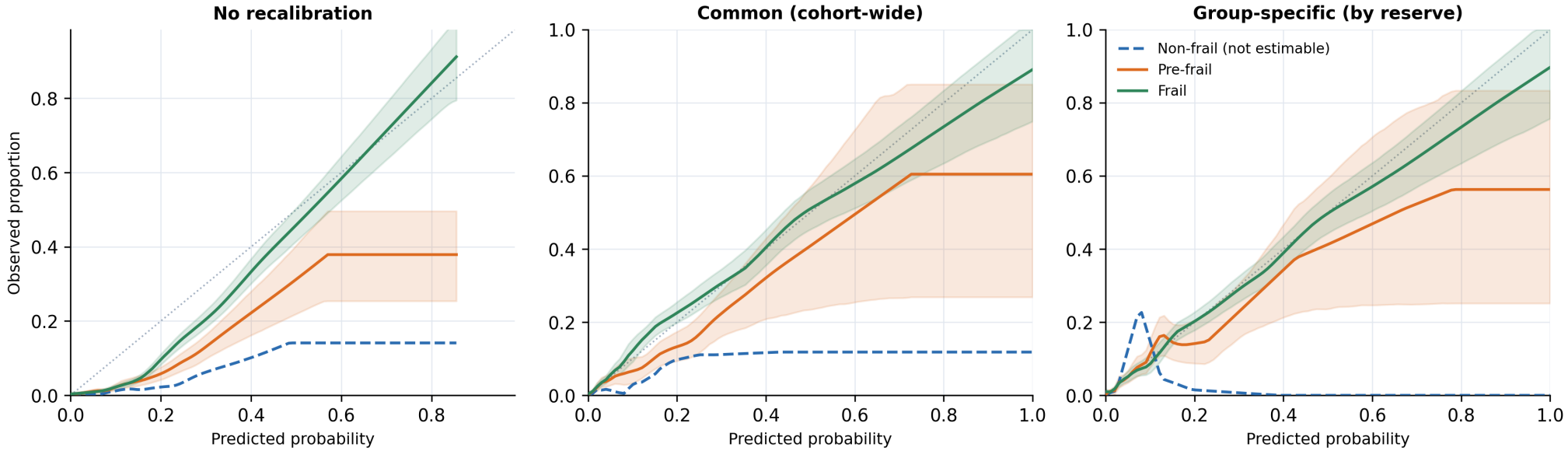
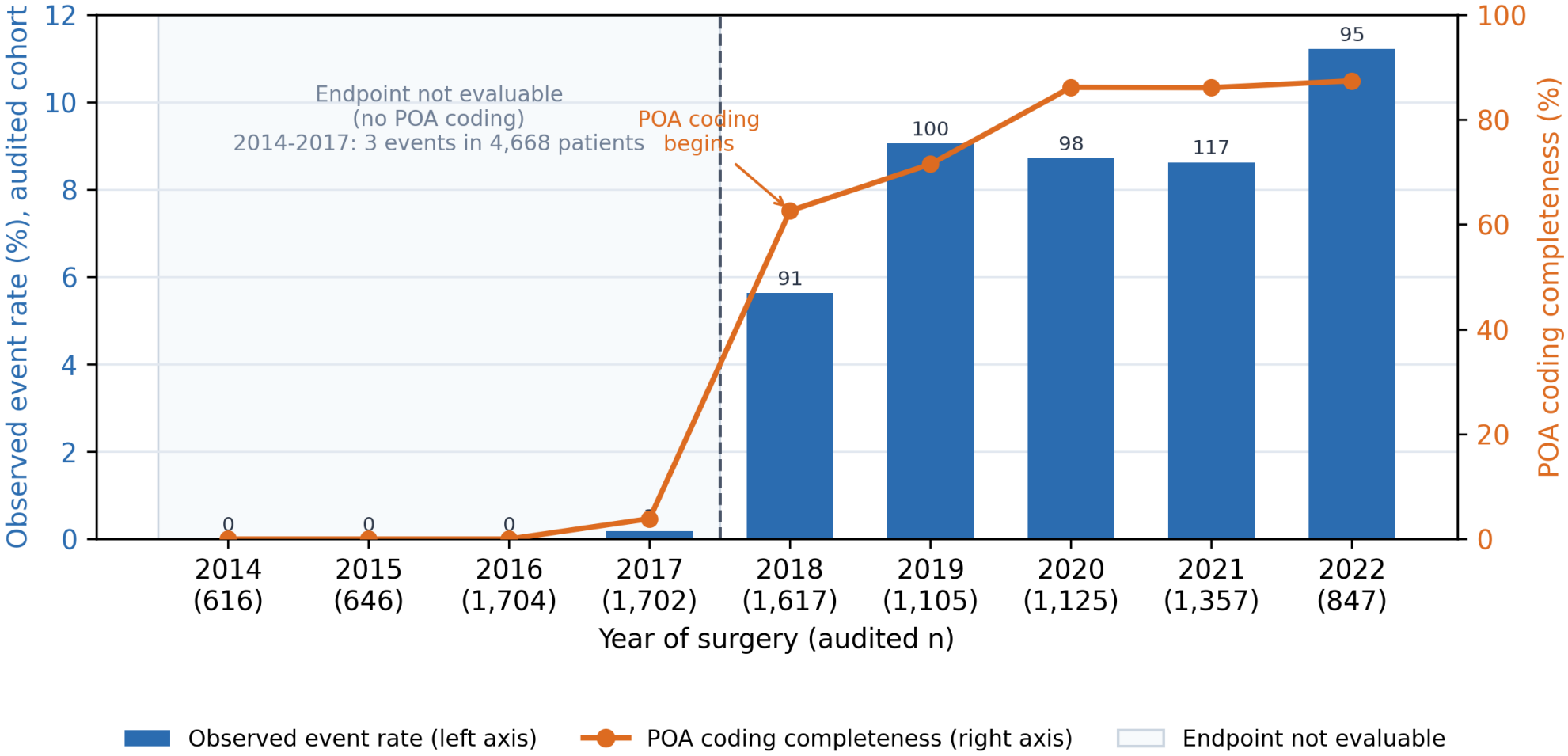
Supplementary Figures

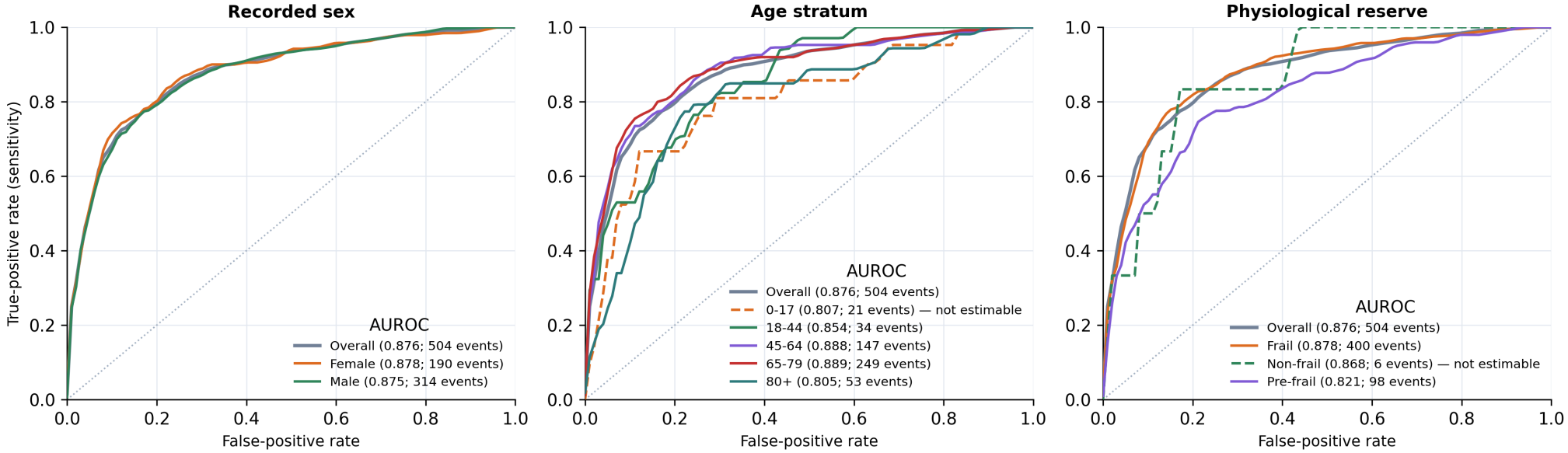
Supplementary Figure 1. Reliability by physiological reserve across the recalibration ladder, audited cohort (n = 10,719; 504 infection events). One panel per recalibration step (no recalibration, common (cohort-wide), and group-specific) with the frail, pre-frail and non-frail curves overlaid in each; the dotted diagonal is perfect calibration and curves below it indicate overestimation. Every transform is cross-fitted on the published cross-validation folds, so none is fitted and evaluated on the same patients; the calibrator family is isotonic throughout, matching the main-text recalibration ladder in Table 4. Each curve is clipped to its own stratum's predicted range so that interpolation beyond the observed range is not rendered as a fitted result. The non-frail stratum rests on 6 events, falls below every pre-registered event floor, and is drawn dashed with a hollow legend marker and labelled not estimable. Its shape under the group-specific step is the point of the panel: group-specific isotonic recalibration flattens it into the nearly constant transform described in Results, mapping the 1,022 non-frail patients to 17 distinct fitted probabilities and a post-recalibration calibration slope of 0.119.



Supplementary Figure 2. Outcome ascertainment by year of surgery, whole audited extract (2014-2022). Bars are the observed coded event rate per year against the left axis, each annotated with its event count; the line is Present-on-Admission (POA) coding completeness against the right axis; the x-axis labels carry each year's audited n, because the figure's claim is that the break is in the label and not in the surgery, which a reader can only check against the denominator. The shaded span marks the years in which the endpoint is not evaluable: POA coding is effectively absent from this extract before 2018, so 2014-2017 contribute 4,668 procedures and 3 coded events, against a coded event rate of 8.28% across the 6,051 procedures from 2018-2022. The dashed vertical line marks the 2018 discontinuity. Because a missing POA flag leaves an infection's postoperative timing undetermined, such cases were excluded at cohort construction rather than recorded as negative outcomes; uninfected patients entered regardless. Those years are therefore case-depleted, not mislabelled, and the cohort is selected on the outcome. This is the figure behind the ascertainment disclosure: the primary analysis remains the audited 2014-2022 cohort the deployed model was developed on, and every prevalence-dependent result is reported alongside its value in the 2018-2022 window. Ascertainment is incomplete and not constant within the retained window as well; the per-year figures are tabulated in Supplementary Table 14 and the by-stratum figures in Supplementary Table 18.



Supplementary Figure 3. Per-subgroup ROC curves, audited cohort (2014-2022; n = 10,719; 504 infection events). One panel per prespecified axis (recorded sex, age stratum, and the three-level physiological-reserve composite) drawn from the audited out-of-fold predictions. Each curve is labelled with its within-subgroup AUROC and its event count, because a curve and its AUROC are uninterpretable without the number of events behind them. Strata below the pre-registered AUROC event floor are drawn dashed and labelled not estimable rather than dropped, so a reader can see what was audited and found unsupportable rather than what was quietly omitted. The case-mix reference axes (surgical specialty, procedure cluster) are excluded: they are descriptive case-mix heterogeneity and too dense to read at supplement size, and their discrimination is tabulated in Supplementary Table 19. Precision-recall curves are not shown; AUPRC is reported numerically for every subgroup in Table 2 and Supplementary Table 19.



Supplementary Tables

Supplementary Table 1. Sex-stratified effect-size scan of the cohort characterization variables (audited cohort, n = 10,719; 504 postoperative infection events). For each of the eight variables tested in Table 1, the table reports the female-vs-male effect size (Cohen's d for continuous variables, Cramér's V for categorical variables, or Cohen's h for binary variables), its conventional magnitude class, the test p-value (Mann-Whitney U for continuous variables and χ^2 for categorical or binary variables), and the Benjamini-Hochberg FDR-adjusted q-value across the eight-variable family. Variables are ordered by ascending q-value.

Variable	Variable label	Kind	Effect kind	Effect size	Magnitude	P value	Q FDR	n
Surgical duration	Surgical duration	continuous	Cohen's d	-0.161	negligible	0.00100	0.00160	10,719
ASA physical status	ASA physical status	categorical	Cramer's V	0.191	small	0.00100	0.00160	10,719
Age stratum	Age stratum	categorical	Cramer's V	0.116	small	0.00100	0.00160	10,719
Physiological reserve	Physiological reserve (frailty)	categorical	Cramer's V	0.185	small	0.00100	0.00160	10,719
Procedure cluster	Procedure cluster	categorical	Cramer's V	0.306	medium	0.00100	0.00160	10,719
Emergency surgery	Emergency surgery rate	binary	Cohen's h	-0.0455	negligible	0.0280	0.0373	10,719
Postoperative infection	Infection prevalence	binary	Cohen's h	0.0320	negligible	0.123	0.141	10,719
Age	Age (continuous)	continuous	Cohen's d	-0.0785	negligible	0.467	0.467	10,719

Supplementary Table 2. No-SMOTE refit comparison. Subgroup calibration and net benefit under a 2 × 2 of training window (2014-2022 as published, and the POA 2018-2022 ascertainment window) and class rebalancing (SMOTE within training folds, and no SMOTE). All four arms are evaluated on the POA 2018-2022 window (501 coded events: 399 frail, 96 pre-frail, 6 non-frail), so calibration-in-the-large, slope, ECE, ICI, Brier score and net benefit at t = 0.10 per 1,000 patients are comparable across arms. The reproduction arm recovered the published out-of-fold predictions bit-identically before the comparison was interpreted. The non-frail stratum falls below every prespecified event floor and is labelled not estimable.

Arm	Reserve stratum	Events	Calibration-in-the-large (95% CI)	Slope	ECE	ICI	Brier	NB at t=0.10 per 1,000 (95% CI)
Published (2014-2022, SMOTE)	Overall	501	-0.579 (-0.673, -0.489)	1.44	0.0470	0.0470	0.0620	+32.70 (+25.74, +39.86)
Published (2014-2022, SMOTE)	Frail	399	-0.438 (-0.551, -0.340)	1.49	0.0440	0.0450	0.0750	+54.20 (+44.06, +64.33)
Published (2014-2022, SMOTE)	Pre-frail	96	-0.843 (-1.068, -0.653)	1.11	0.0550	0.0540	0.0500	+3.48 (-5.80, +13.55)
Published (2014-2022, SMOTE)	Non-frail	6	-1.996 (not estimable)	1.357	0.0590	0.0590	0.0170	-14.83 (not estimable)
No SMOTE (2014-2022)	Overall	501	0.358 (0.259, 0.452)	1.02	0.0210	0.0220	0.0620	+41.32 (+35.31, +47.45)
No SMOTE (2014-2022)	Frail	399	0.567 (0.448, 0.668)	1.02	0.0380	0.0390	0.0780	+62.22 (+53.17, +71.87)
No SMOTE (2014-2022)	Pre-frail	96	-0.011 (-0.243, 0.185)	0.805	0.0160	0.00900	0.0470	+12.45 (+5.00, +21.00)
No SMOTE (2014-2022)	Non-frail	6	-1.311 (not estimable)	0.886	0.0280	0.0280	0.0130	-3.45 (not estimable)
Corrected window (2018-2022, SMOTE)	Overall	501	-0.852 (-0.944, -0.763)	1.58	0.0770	0.0750	0.0660	+24.73 (+17.52, +32.23)
Corrected window (2018-2022, SMOTE)	Frail	399	-0.687 (-0.802, -0.589)	1.61	0.0720	0.0700	0.0780	+47.11 (+36.71, +57.43)
Corrected window (2018-2022, SMOTE)	Pre-frail	96	-1.161 (-1.381, -0.968)	1.22	0.0860	0.0840	0.0560	-3.79 (-13.92, +7.14)
Corrected window (2018-2022, SMOTE)	Non-frail	6	-2.358 (not estimable)	2.121	0.0860	0.0850	0.0220	-31.08 (not estimable)
No SMOTE, corrected window	Overall	501	-0.123 (-0.219, -0.034)	1.29	0.0200	0.0190	0.0610	+39.24 (+32.72, +46.07)
No SMOTE, corrected window	Frail	399	0.109 (-0.007, 0.211)	1.26	0.0220	0.0210	0.0750	+61.38 (+51.72, +71.21)
No SMOTE, corrected window	Pre-frail	96	-0.530 (-0.757, -0.339)	1.06	0.0310	0.0310	0.0480	+9.22 (+0.73, +18.93)
No SMOTE, corrected window	Non-frail	6	-1.875 (not estimable)	1.764	0.0520	0.0510	0.0150	-9.95 (not estimable)

Supplementary Table 3. Age-bin sensitivity, 2018-2022 ascertainment window (n = 6,051; 501 coded events). Per-stratum calibration-in-the-large and net benefit at t = 0.10 per 1,000 patients under three age binnings: the prespecified cut points (18/45/65/80), the institutional pediatric cut (≤15), and equal-frequency quintiles. The GAP row for each binning is the maximum between-stratum calibration-intercept gap. Prespecified event floors are retained and estimates below them are labelled not estimable.

Binning	Stratum	n	Events	Calibration-in-the-large (95% CI)	Net benefit at t=0.10 per 1,000 (95% CI)
Prespecified (18/45/65/80)	0-17	223	21	-0.463 (not estimable)	+24.41 (not estimable)
Prespecified (18/45/65/80)	18-44	518	34	-0.762 (-1.183, -0.441)	+14.59 (not estimable)
Prespecified (18/45/65/80)	45-64	2,049	146	-0.710 (-0.884, -0.555)	+25.16 (+13.83, +36.71)
Prespecified (18/45/65/80)	65-79	2,698	247	-0.467 (-0.601, -0.344)	+43.20 (+32.29, +54.36)
Prespecified (18/45/65/80)	80+	563	53	-0.604 (-0.914, -0.315)	+29.80 (+6.31, +55.07)
Prespecified (18/45/65/80)	GAP	--	--	0.298 (not estimable)	--
Institutional pediatric cut (≤15)	0-15	203	21	-0.385 (not estimable)	+30.10 (not estimable)
Institutional pediatric cut (≤15)	16-44	538	34	-0.797 (-1.183, -0.470)	+12.80 (not estimable)
Institutional pediatric cut (≤15)	45-64	2,049	146	-0.710 (-0.884, -0.555)	+25.16 (+13.83, +36.71)
Institutional pediatric cut (≤15)	65-79	2,698	247	-0.467 (-0.601, -0.344)	+43.20 (+32.29, +54.36)
Institutional pediatric cut (≤15)	80+	563	53	-0.604 (-0.914, -0.315)	+29.80 (+6.31, +55.07)
Institutional pediatric cut (≤15)	GAP	--	--	0.412 (not estimable)	--
Equal-frequency quintiles	0-52	1,223	91	-0.681 (-0.910, -0.469)	+22.80 (+8.63, +37.80)
Equal-frequency quintiles	52-62	1,210	86	-0.717 (-0.961, -0.510)	+24.33 (+10.19, +39.03)
Equal-frequency quintiles	62-69	1,264	101	-0.556 (-0.759, -0.378)	+35.34 (+20.04, +51.25)
Equal-frequency quintiles	69-75	1,192	113	-0.434 (-0.644, -0.244)	+42.13 (+25.63, +58.91)
Equal-frequency quintiles	75-99	1,162	110	-0.532 (-0.752, -0.347)	+39.30 (+22.18, +57.18)
Equal-frequency quintiles	GAP	--	--	0.282 (0.148, 0.616)	--

Supplementary Table 4. Code list and weights for the Hospital Frailty Risk Score (HFRS). HFRS codes and weights follow Gilbert et al. (Lancet 2018; 391:1775-82), Supplementary Table A2 (109 categories, version 5.6). The admission score is the sum of the weights of all distinct categories present. Risk bands: low (score < 5; contributes "non-frail"), intermediate (5 ≤ score ≤ 15; contributes "pre-frail"), high (score > 15; contributes "frail").

ICD-10 category	Weight	Condition
F00	7.10	Dementia in Alzheimer's disease
G81	4.40	Hemiplegia
G30	4	Alzheimer's disease
I69	3.70	Sequelae of cerebrovascular disease (secondary)
R29	3.60	Other symptoms involving nervous & musculoskeletal (R29.6 tendency to fall)
F05	3.20	Delirium, not induced by alcohol/psychoactive
N39	3.20	Other disorders of urinary system (incl. UTI, incontinence)
S00	3.20	Superficial injury of head
W19	3.20	Unspecified fall
R31	3	Unspecified haematuria
B96	2.90	Other bacterial agents (secondary)
R41	2.70	Other symptoms involving cognitive functions
I67	2.60	Other cerebrovascular diseases
R26	2.60	Abnormalities of gait and mobility
R56	2.60	Convulsions, NEC
R40	2.50	Somnolence, stupor and coma
S06	2.40	Intracranial injury
T83	2.40	Complications of genitourinary prosthetic devices
E86	2.30	Volume depletion
E87	2.30	Other disorders of fluid/electrolyte/acid-base
M25	2.30	Other joint disorders, NEC
S42	2.30	Fracture of shoulder/upper arm
R54	2.20	Senility
F03	2.10	Unspecified dementia
W18	2.10	Other fall on same level
Z50	2.10	Care involving rehabilitation procedures
F01	2	Vascular dementia
L03	2	Cellulitis
S80	2	Superficial injury of lower leg
Z75	2	Problems related to medical facilities
E53	1.90	Deficiency of other B group vitamins
H54	1.90	Blindness and low vision
G20	1.80	Parkinson's disease
K59	1.80	Other functional intestinal disorders
N17	1.80	Acute renal failure

ICD-10 category	Weight	Condition
R55	1.80	Syncope and collapse
S22	1.80	Fracture of rib(s)/sternum/thoracic spine
Z60	1.80	Problems related to social environment
B95	1.70	Streptococcus/staphylococcus as cause (secondary)
L89	1.70	Decubitus ulcer
Z22	1.70	Carrier of infectious disease
A41	1.60	Other septicaemia
I95	1.60	Hypotension
K26	1.60	Duodenal ulcer
L97	1.60	Ulcer of lower limb, NEC
N19	1.60	Unspecified renal failure
R44	1.60	Other symptoms involving general sensations
G40	1.50	Epilepsy
J96	1.50	Respiratory failure, NEC
M19	1.50	Other arthrosis
X59	1.50	Exposure to unspecified factor
Z87	1.50	Personal history of other diseases and conditions
E16	1.40	Other disorders of pancreatic internal secretion
M81	1.40	Osteoporosis without pathological fracture
N18	1.40	Chronic renal failure
R94	1.40	Abnormal results of function studies
S32	1.40	Fracture of lumbar spine and pelvis
S72	1.40	Fracture of femur
N28	1.30	Other disorders of kidney and ureter, NEC
R33	1.30	Retention of urine
R69	1.30	Unknown and unspecified causes of morbidity
G31	1.20	Other degenerative diseases of nervous system, NEC
G45	1.20	Transient cerebral ischaemic attacks
R32	1.20	Unspecified urinary incontinence
R45	1.20	Symptoms/signs involving emotional state
S09	1.20	Other and unspecified injuries of head
Y95	1.20	Nosocomial condition
A04	1.10	Other bacterial intestinal infections
A09	1.10	Diarrhoea and gastroenteritis of presumed infectious origin
J18	1.10	Pneumonia, organism unspecified
M79	1.10	Other soft tissue disorders, NEC
S01	1.10	Open wound of head
W06	1.10	Fall involving bed
Z74	1.10	Problems related to care-provider dependency

ICD-10 category	Weight	Condition
E55	1	Vitamin D deficiency
J69	1	Pneumonitis due to solids and liquids
R02	1	Gangrene, NEC
R47	1	Speech disturbances, NEC
Z93	1	Artificial opening status
E05	0.900	Thyrotoxicosis (hyperthyroidism)
H91	0.900	Other hearing loss
M41	0.900	Scoliosis
R63	0.900	Symptoms concerning food and fluid intake
W01	0.900	Fall on same level slipping/tripping
W10	0.900	Fall on/from stairs and steps
I63	0.800	Cerebral infarction
K92	0.800	Other diseases of digestive system
M80	0.800	Osteoporosis with pathological fracture
R13	0.800	Dysphagia
U80	0.800	Agent resistant to penicillin/related antibiotics
Z99	0.800	Dependence on enabling machines and devices
F10	0.700	Mental/behavioural disorders due to alcohol
J22	0.700	Unspecified acute lower respiratory infection
N20	0.700	Calculus of kidney and ureter
R00	0.700	Abnormalities of heart beat
Y84	0.700	Other medical procedures as cause of abnormal reaction
R79	0.600	Other abnormal findings of blood chemistry
Z73	0.600	Problems related to life-management difficulty
F32	0.500	Depressive episode
M48	0.500	Spinal stenosis (secondary code only)
S51	0.500	Open wound of forearm
Z91	0.500	Personal history of risk-factors, NEC
D64	0.400	Other anaemias
E83	0.400	Disorders of mineral metabolism
L08	0.400	Other local infections of skin and subcutaneous tissue
M15	0.400	Polyarthrosis
K52	0.300	Other noninfective gastroenteritis and colitis
R11	0.300	Nausea and vomiting
R50	0.100	Fever of unknown origin

Supplementary Table 5. Item-to-code map for the modified Frailty Index (mFI-4). The five-factor modified Frailty Index (mFI-5; Subramaniam et al., JAMA Surg 2018) includes functional status, which is not coded in our EHR. We therefore used the remaining four items (mFI-4). Each item scores 1 if any listed ICD-10 category is present; the item score is the sum (0-4).

mFI-4 item	ICD-10 categories	Note
Diabetes mellitus	E10, E11, E12, E13, E14	Any type
Chronic obstructive pulmonary disease	J41, J42, J43, J44, J47	Includes bronchiectasis (J47)
Congestive heart failure	I50	—
Hypertension (medication-requiring)	I10, I11, I12, I13, I15	Identified from ICD codes without medication linkage; may overcount this item (see Limitations / sensitivity analysis).

Supplementary Table 6. Composite physiological-reserve classification rule. A patient is assigned the highest reserve-loss level any single instrument supports. Missing inputs are conservatively ignored (a criterion can only raise the reserve-loss level if its input is present).

Level	Rule (any one criterion sufficient)
Frail	ASA-PS ≥ 4, or HFRS high (> 15), or mFI-4 ≥ 2
Pre-frail	(not frail) and: ASA-PS = 3, or HFRS intermediate (5–15), or mFI-4 = 1
Non-frail	none of the above (ASA-PS ≤ 2, HFRS low [< 5], mFI-4 = 0)

Supplementary Table 7. Threshold sensitivity of subgroup net benefit, audited cohort (n = 10,719; 504 postoperative infection events). Decision-curve net benefit per 1,000 patients, with 95% patient-level bootstrap intervals, at seven threshold probabilities from 0.05 to 0.20 across the three prespecified axes, together with the threshold at which each subgroup's net benefit crosses zero. The final column records whether the stratum met the prespecified 50-event floor for net-benefit inference; strata that did not are shown as point estimates labelled not estimable.

Axis	Subgroup	0.05	0.075	0.1	0.125	0.15	0.175	0.2	Zero crossing	Estimable
Recorded sex	Female	+17.68 (+10.56, +24.71)	+11.04 (+3.88, +17.94)	+8.43 (+1.23, +15.24)	+7.98 (+0.77, +14.64)	+6.47 (-0.70, +13.30)	+5.20 (-1.82, +12.21)	+6.03 (-1.15, +12.95)	none in range	yes
Recorded sex	Male	+11.10 (+6.23, +15.92)	+5.23 (+0.35, +10.11)	+3.00 (-2.06, +7.98)	+1.40 (-3.35, +6.38)	+0.33 (-4.30, +5.16)	+1.22 (-3.40, +6.11)	+1.56 (-3.06, +6.12)	none in range	yes
Age stratum	0-17	+15.63 (not estimable)	+2.36 (not estimable)	-7.27 (not estimable)	-7.40 (not estimable)	-1.76 (not estimable)	-5.95 (not estimable)	-0.00 (not estimable)	0.195	no
Age stratum	18-44	+5.48 (not estimable)	-0.03 (not estimable)	-4.25 (not estimable)	-5.47 (not estimable)	-7.81 (not estimable)	-8.79 (not estimable)	-11.16 (not estimable)	0.075	no
Age stratum	45-64	+7.18 (+0.78, +13.86)	+2.91 (-3.58, +9.76)	+0.69 (-5.77, +7.57)	-0.69 (-7.00, +6.00)	-2.11 (-8.15, +4.50)	-1.16 (-7.09, +5.47)	-0.40 (-6.46, +6.33)	0.198	yes
Age stratum	65-79	+19.64 (+13.49, +26.07)	+12.97 (+6.66, +19.11)	+12.05 (+5.65, +18.29)	+10.58 (+4.24, +16.93)	+9.81 (+3.62, +16.13)	+10.15 (+3.95, +16.49)	+11.37 (+5.44, +17.45)	none in range	yes
Age stratum	80+	+13.52 (+0.21, +27.56)	+5.02 (-7.99, +18.57)	-0.11 (-12.96, +13.50)	+0.14 (-12.85, +14.12)	-3.84 (-16.51, +10.17)	-4.97 (-17.79, +8.89)	-7.91 (-21.25, +5.44)	0.135	yes
Physiological reserve	Frail	+28.51 (+22.00, +34.83)	+20.93 (+14.43, +27.34)	+17.21 (+10.61, +23.73)	+14.64 (+8.12, +21.43)	+12.60 (+6.18, +19.10)	+12.96 (+6.70, +19.38)	+13.32 (+7.33, +19.55)	none in range	yes
Physiological reserve	Non-frail	-16.94 (not estimable)	-16.85 (not estimable)	-15.44 (not estimable)	-13.98 (not estimable)	-12.43 (not estimable)	-11.12 (not estimable)	-10.52 (not estimable)	none in range	no
Physiological reserve	Pre-frail	-4.05 (-9.38, +1.42)	-9.49 (-14.78, -4.25)	-10.59 (-15.77, -5.37)	-10.20 (-15.42, -5.02)	-10.80 (-15.79, -5.72)	-11.36 (-16.24, -6.17)	-10.61 (-15.25, -5.69)	none in range	yes

Supplementary Table 8. Per-subgroup Hosmer-Lemeshow goodness-of-fit on the audited cohort (n = 10,719; 504 postoperative infection events), for all 70 strata across the three prespecified axes, the two case-mix reference axes, and the three pairwise intersections. Patients were grouped into ten groups of approximately equal size by predicted risk, with cut points snapped to tied predicted probabilities so that identical predictions always share a group. For each stratum we report the Hosmer-Lemeshow statistic, its degrees of freedom, the corresponding p-value, and the number of non-empty groups actually used. Degrees of freedom equal the number of groups used: these are out-of-fold predictions from a model fitted on the remaining folds, so no degrees of freedom are spent estimating an intercept and slope on the data under test. Kramer & Zimmerman use g-2, which applies to development data where those two parameters are estimated on the sample under test; for out-of-fold predictions the validation convention g is the applicable one. Implementations differ on identical data, so the convention is stated here rather than left to be inferred. The decile-mean slope and intercept are the least-squares regression of observed on expected group event rates recommended by Kramer & Zimmerman (perfect calibration gives 1 and 0 respectively). Expected calibration error (ECE) and integrated calibration index (ICI) are reported alongside the statistic as the adjunct measures Kramer and Zimmerman [38] require; for the three prespecified axes these match main Table 3. All 70 strata used the full ten risk groups, so no stratum carries reduced degrees of freedom. Strata flagged ‡ used empirical-Bayes shrinkage toward the marginal rate (cell n events < 50) and their statistics are not interpretable; 22 of the 70 strata are so flagged. A dash in the statistic column marks a stratum containing a single outcome class, for which the statistic is undefined; its per-decile table of expected events is still reported. The Hosmer-Lemeshow statistic is sample-size dependent and is therefore not comparable across strata: at fixed lack of fit it grows with n, so large strata are flagged for deviations that are clinically irrelevant while small strata with worse fit are not.

Axis	Subgroup	EB shrinkage	n	Events	Hosmer-Lemeshow statistic	Hosmer-Lemeshow df	Hosmer-Lemeshow p value	N bins used	Bins reduced	Decile slope	Decile intercept	ECE	ICI	Gap max Hosmer-Lemeshow Gap
Recorded sex	Overall	No	10,719	504	531	10	9.90e-108	10	No	0.746	-0.0347	0.0624	0.0609	175
Recorded sex	female	No	3,689	190	178	10	6.82e-33	10	No	0.777	-0.0365	0.0614	0.0600	175
Recorded sex	male	No	7,030	314	353	10	9.60e-70	10	No	0.718	-0.0328	0.0629	0.0614	175
Age stratum	Overall	No	10,719	504	531	10	9.90e-108	10	No	0.746	-0.0347	0.0624	0.0609	198
Age stratum	0-17	No	367	21	22.5	10	0.0126	10	No	0.833	-0.0498	0.0733	0.0713	198
Age stratum	18-44	No	941	34	50.5	10	2.18e-07	10	No	0.532	-0.0178	0.0647	0.0636	198
Age stratum	45-64	No	3,715	147	202	10	6.91e-38	10	No	0.675	-0.0312	0.0646	0.0631	198
Age stratum	65-79	No	4,684	249	221	10	7.15e-42	10	No	0.884	-0.0437	0.0568	0.0560	198
Age stratum	80+	No	1,012	53	60.5	10	2.91e-09	10	No	0.535	-0.0155	0.0745	0.0748	198
Physiological reserve	Overall	No	10,719	504	531	10	9.90e-108	10	No	0.746	-0.0347	0.0624	0.0609	235
Physiological reserve	Frail	No	6,139	400	294	10	3.59e-57	10	No	0.860	-0.0461	0.0637	0.0620	235
Physiological reserve	Non-frail	No	1,022	6	58.1	10	8.32e-09	10	No	0.150	-0.00322	0.0550	0.0550	235
Physiological reserve	Pre-frail	No	3,558	98	188	10	5.79e-35	10	No	0.436	-0.0113	0.0623	0.0617	235
Surgical specialty	Overall	No	10,719	504	531	10	9.90e-108	10	No	0.746	-0.0347	0.0624	0.0609	255
Surgical specialty	ORL	No	233	15	22.4	10	0.0134	10	No	0.537	-0.0233	0.0984	0.0982	255
Surgical specialty	Gynaecology	No	452	24	13.6	10	0.194	10	No	0.783	-0.0212	0.0423	0.0450	255
Surgical specialty	Neurosurgery	No	1,117	26	50.7	10	1.97e-07	10	No	0.561	-0.0198	0.0522	0.0536	255
Surgical specialty	Orthopaedics	No	1,690	76	86.9	10	2.24e-14	10	No	0.593	-0.0206	0.0654	0.0649	255
Surgical specialty	Paediatrics	No	65	3	7.93	10	0.636	10	No	0.335	-0.00142	0.107	0.0887	255
Surgical specialty	Thoracic surgery	No	723	14	37.0	10	5.61e-05	10	No	0.477	-0.0164	0.0560	0.0552	255
Surgical specialty	Urology	No	349	19	17.8	10	0.0578	10	No	0.856	-0.0360	0.0546	0.0510	255
Surgical specialty	Vascular surgery	No	4,852	184	263	10	9.41e-51	10	No	0.701	-0.0339	0.0639	0.0624	255
Surgical specialty	Visceral surgery	No	1,238	143	66.6	10	2.02e-10	10	No	0.988	-0.0657	0.0713	0.0669	255
Procedure cluster	Overall	No	10,719	504	531	10	9.90e-108	10	No	0.746	-0.0347	0.0624	0.0609	204
Procedure cluster	chopcluster_1 (NervousSystem)	No	903	22	44.3	10	2.88e-06	10	No	0.575	-0.0218	0.0546	0.0551	204
Procedure cluster	chopcluster_10 (UroAndrological+Urologic)	No	346	28	15.6	10	0.110	10	No	1.12	-0.0559	0.0599	0.0516	204
Procedure cluster	chopcluster_11 (Gynecological)	No	303	9	13.3	10	0.208	10	No	0.339	0.00295	0.0533	0.0502	204
Procedure cluster	chopcluster_12 (Musculoskeletal)	No	1,623	62	86.6	10	2.51e-14	10	No	0.563	-0.0205	0.0658	0.0650	204
Procedure cluster	chopcluster_13 (Cutaneous)	No	213	43	5.89	10	0.824	10	No	1.24	-0.0699	0.0579	0.0323	204
Procedure cluster	chopcluster_3 (Thoracic)	No	807	22	42.7	10	5.70e-06	10	No	0.604	-0.0249	0.0587	0.0581	204
Procedure cluster	chopcluster_4 (Cardiac)	No	3,764	118	210	10	1.34e-39	10	No	0.608	-0.0273	0.0644	0.0630	204

Axis	Subgroup	EB shrinkage	n	Events	Hosmer-Lemeshow statistic	Hosmer-Lemeshow df	Hosmer-Lemeshow p value	N bins used	Bins reduced	Decile slope	Decile intercept	ECE	ICI	Gap max Hosmer-Lemeshow Gap
Procedure cluster	chopcluster_5 (Vascular)	No	769	30	46.2	10	1.33e-06	10	No	0.690	-0.0343	0.0684	0.0670	204
Procedure cluster	chopcluster_6 (AbdominalWall)	No	530	67	25.4	10	0.00465	10	No	1.08	-0.0738	0.0665	0.0646	204
Procedure cluster	chopcluster_7 (GastroIntestinal)	No	298	36	23.2	10	0.0101	10	No	0.990	-0.0917	0.0932	0.0928	204
Procedure cluster	chopcluster_8 (Hepatobiliary)	No	306	23	21.1	10	0.0207	10	No	0.653	-0.0226	0.0823	0.0739	204
Procedure cluster	Unclassified	No	857	44	43.1	10	4.73e-06	10	No	0.695	-0.0296	0.0644	0.0632	204
Recorded sex × age stratum	Overall	No	10,719	504	531	10	9.90e-108	10	No	0.746	-0.0347	0.0624	0.0609	146
Recorded sex × age stratum	female / 0-17	Yes	175	13	13.9	10	0.180	10	No	1.18	-0.0912	0.0771	0.0644	146
Recorded sex × age stratum	female / 18-44	Yes	408	13	27.5	10	0.00220	10	No	0.542	-0.0248	0.0731	0.0720	146
Recorded sex × age stratum	female / 45-64	No	1,179	52	60.9	10	2.45e-09	10	No	0.708	-0.0302	0.0606	0.0591	146
Recorded sex × age stratum	female / 65-79	No	1,470	93	61.6	10	1.82e-09	10	No	0.946	-0.0460	0.0519	0.0556	146
Recorded sex × age stratum	female / 80+	Yes	457	19	37.7	10	4.28e-05	10	No	0.466	-0.0172	0.0834	0.0839	146
Recorded sex × age stratum	male / 0-17	Yes	192	8	15.4	10	0.119	10	No	0.388	-0.00826	0.0859	0.0802	146
Recorded sex × age stratum	male / 18-44	Yes	533	21	24.0	10	0.00756	10	No	0.521	-0.0115	0.0582	0.0578	146
Recorded sex × age stratum	male / 45-64	No	2,536	95	144	10	7.33e-26	10	No	0.648	-0.0301	0.0666	0.0654	146
Recorded sex × age stratum	male / 65-79	No	3,214	156	160	10	3.25e-29	10	No	0.834	-0.0411	0.0590	0.0572	146
Recorded sex × age stratum	male / 80+	Yes	555	34	27.5	10	0.00218	10	No	0.590	-0.0138	0.0671	0.0676	146
Recorded sex × physiological reserve	Overall	No	10,719	504	531	10	9.90e-108	10	No	0.746	-0.0347	0.0624	0.0609	186
Recorded sex × physiological reserve	female / frail	No	1,666	129	95.1	10	5.07e-16	10	No	0.915	-0.0560	0.0683	0.0672	186
Recorded sex × physiological reserve	female / non_frail	Yes	522	1	37.1	10	5.47e-05	10	No	0.0773	-0.00315	0.0637	0.0638	186
Recorded sex × physiological reserve	female / pre_frail	No	1,501	60	58.0	10	8.67e-09	10	No	0.580	-0.0142	0.0531	0.0525	186
Recorded sex × physiological reserve	male / frail	No	4,473	271	207	10	4.52e-39	10	No	0.837	-0.0416	0.0620	0.0601	186
Recorded sex × physiological reserve	male / non_frail	Yes	500	5	21.8	10	0.0162	10	No	0.247	-0.00407	0.0460	0.0459	186
Recorded sex × physiological reserve	male / pre_frail	Yes	2,057	38	136	10	3.04e-24	10	No	0.312	-0.00876	0.0691	0.0686	186
Age stratum × physiological reserve	Overall	No	10,719	504	531	10	9.90e-108	10	No	0.746	-0.0347	0.0624	0.0609	128
Age stratum × physiological reserve	0-17 / frail	Yes	211	19	12.1	10	0.279	10	No	0.942	-0.0582	0.0705	0.0664	128
Age stratum × physiological reserve	0-17 / non_frail	Yes	45	0				10	No			0.0721	0.0721	128
Age stratum × physiological reserve	0-17 / pre_frail	Yes	111	2	11.8	10	0.299	10	No	-0.0195	0.0214	0.0927	0.0844	128
Age stratum × physiological reserve	18-44 / frail	Yes	300	21	19.6	10	0.0335	10	No	0.721	-0.0409	0.0827	0.0811	128
Age stratum × physiological reserve	18-44 / non_frail	Yes	331	2	20.3	10	0.0264	10	No	0.0785	0.00119	0.0561	0.0562	128
Age stratum × physiological reserve	18-44 / pre_frail	Yes	310	11	14.8	10	0.138	10	No	0.353	0.00367	0.0564	0.0567	128
Age stratum × physiological reserve	45-64 / frail	No	1,967	111	110	10	6.00e-19	10	No	0.803	-0.0437	0.0687	0.0667	128
Age stratum × physiological reserve	45-64 / non_frail	Yes	410	3	21.2	10	0.0197	10	No	0.315	-0.0109	0.0512	0.0510	128
Age stratum × physiological reserve	45-64 / pre_frail	Yes	1,338	33	77.0	10	1.92e-12	10	No	0.419	-0.0109	0.0625	0.0619	128
Age stratum × physiological reserve	65-79 / frail	No	3,032	204	140	10	4.51e-25	10	No	0.985	-0.0538	0.0568	0.0575	128
Age stratum × physiological reserve	65-79 / non_frail	Yes	218	1	13.5	10	0.197	10	No	-0.0155	0.00465	0.0552	0.0553	128

Axis	Subgroup	EB shrinkage	n	Events	Hosmer-Lemeshow statistic	Hosmer-Lemeshow df	Hosmer-Lemeshow p value	N bins used	Bins reduced	Decile slope	Decile intercept	ECE	ICI	Gap max Hosmer-Lemeshow Gap
Age stratum × physiological reserve	65-79 / pre_frail	Yes	1,434	44	71.1	10	2.77e-11	10	No	0.529	-0.0168	0.0593	0.0583	128
Age stratum × physiological reserve	80+ / frail	Yes	629	45	40.6	10	1.35e-05	10	No	0.592	-0.0156	0.0753	0.0763	128
Age stratum × physiological reserve	80+ / non_frail	Yes	18	0				10	No			0.0783	0.0783	128
Age stratum × physiological reserve	80+ / pre_frail	Yes	365	8	26.8	10	0.00283	10	No	0.257	-0.00178	0.0739	0.0726	128

Supplementary Table 9. Per-decile observed versus expected infection events by prespecified subgroup, audited cohort (n = 10,719; 504 infection events). Within each stratum, patients are ordered by predicted risk and split into ten groups of approximately equal size, with cut points snapped to tied predicted probabilities. For each group we report its size, the observed and expected number of events, both as rates, and their difference (observed minus expected; negative means the model predicted more events than occurred). This is the per-decile table reference [38] requires to be read alongside the Hosmer-Lemeshow statistic, which is reported with its adjunct calibration measures in Supplementary Table 8; neither table is interpretable without the other. Strata are the three prespecified axes only; the case-mix and intersectional axes are available in the machine-readable source, where almost every stratum falls below the pre-registered event floors. Two strata fall below the prespecified event floors and are marked †: non-frail (6 events) and age 0-17 (21 events). They are shown because the shape of their distribution across the ten groups is informative, not because their per-group rates are estimable. Expected counts are sums of predicted probabilities, so they are not integers.

Axis	Subgroup	Stratum events	Below event floor	Risk group	n	Observed	Expected	Observed rate	Expected rate	Difference
Recorded sex	Overall	504	No	1	1,227	4	11.4	0.00326	0.00926	-7.36
Recorded sex	Overall	504	No	2	833	4	18.9	0.00480	0.0227	-14.9
Recorded sex	Overall	504	No	3	1,322	10	46.0	0.00756	0.0348	-36.0
Recorded sex	Overall	504	No	4	741	6	35.1	0.00810	0.0473	-29.1
Recorded sex	Overall	504	No	5	1,325	13	82.4	0.00981	0.0622	-69.4
Recorded sex	Overall	504	No	6	1,024	13	84.4	0.0127	0.0824	-71.4
Recorded sex	Overall	504	No	7	1,050	20	112	0.0190	0.107	-92.3
Recorded sex	Overall	504	No	8	1,010	42	144	0.0416	0.143	-102
Recorded sex	Overall	504	No	9	1,120	78	231	0.0696	0.206	-153
Recorded sex	Overall	504	No	10	1,067	314	407	0.294	0.382	-93.3
Recorded sex	Female	190	No	1	436	3	4.12	0.00688	0.00946	-1.12
Recorded sex	Female	190	No	2	278	1	6.32	0.00360	0.0227	-5.32
Recorded sex	Female	190	No	3	427	3	14.8	0.00703	0.0348	-11.8
Recorded sex	Female	190	No	4	389	3	19.3	0.00771	0.0496	-16.3
Recorded sex	Female	190	No	5	328	5	21.2	0.0152	0.0645	-16.2
Recorded sex	Female	190	No	6	343	4	28.2	0.0117	0.0823	-24.2
Recorded sex	Female	190	No	7	366	5	39.1	0.0137	0.107	-34.1
Recorded sex	Female	190	No	8	393	19	57.9	0.0483	0.147	-38.9
Recorded sex	Female	190	No	9	367	31	80.5	0.0845	0.219	-49.5
Recorded sex	Female	190	No	10	362	116	145	0.320	0.401	-29.1
Recorded sex	Male	314	No	1	791	1	7.24	0.00126	0.00915	-6.24
Recorded sex	Male	314	No	2	555	3	12.6	0.00541	0.0226	-9.55
Recorded sex	Male	314	No	3	895	7	31.2	0.00782	0.0348	-24.2
Recorded sex	Male	314	No	4	471	5	22.3	0.0106	0.0473	-17.3
Recorded sex	Male	314	No	5	878	6	54.7	0.00683	0.0623	-48.7
Recorded sex	Male	314	No	6	681	9	56.1	0.0132	0.0824	-47.1
Recorded sex	Male	314	No	7	684	15	73.2	0.0219	0.107	-58.2
Recorded sex	Male	314	No	8	682	25	97.5	0.0367	0.143	-72.5
Recorded sex	Male	314	No	9	679	47	138	0.0692	0.203	-90.5
Recorded sex	Male	314	No	10	714	196	264	0.275	0.370	-68.0
Age stratum	0-17	21	Yes	1	37	0	0.935	0	0.0253	-0.935
Age stratum	0-17	21	Yes	2	36	1	1.72	0.0278	0.0476	-0.715
Age stratum	0-17	21	Yes	3	38	0	2.46	0	0.0649	-2.46
Age stratum	0-17	21	Yes	4	32	2	2.57	0.0625	0.0803	-0.570

Axis	Subgroup	Stratum events	Below event floor	Risk group	n	Observed	Expected	Observed rate	Expected rate	Difference
Age stratum	0-17	21	Yes	5	45	0	4.41	0	0.0980	-4.41
Age stratum	0-17	21	Yes	6	33	1	3.91	0.0303	0.118	-2.91
Age stratum	0-17	21	Yes	7	39	1	5.54	0.0256	0.142	-4.54
Age stratum	0-17	21	Yes	8	35	2	6.21	0.0571	0.177	-4.21
Age stratum	0-17	21	Yes	9	38	4	8.38	0.105	0.220	-4.38
Age stratum	0-17	21	Yes	10	34	10	11.8	0.294	0.347	-1.79
Age stratum	18-44	34	No	1	82	0	0.515	0	0.00628	-0.515
Age stratum	18-44	34	No	2	105	0	2.08	0	0.0198	-2.08
Age stratum	18-44	34	No	3	85	0	2.82	0	0.0331	-2.82
Age stratum	18-44	34	No	4	115	1	5.11	0.00870	0.0444	-4.11
Age stratum	18-44	34	No	5	86	0	4.93	0	0.0574	-4.93
Age stratum	18-44	34	No	6	100	4	7.30	0.0400	0.0730	-3.30
Age stratum	18-44	34	No	7	86	2	8.13	0.0233	0.0945	-6.13
Age stratum	18-44	34	No	8	98	4	12.6	0.0408	0.129	-8.60
Age stratum	18-44	34	No	9	92	5	18.0	0.0543	0.196	-13.0
Age stratum	18-44	34	No	10	92	18	33.4	0.196	0.363	-15.4
Age stratum	45-64	147	No	1	348	1	2.23	0.00287	0.00641	-1.23
Age stratum	45-64	147	No	2	473	2	9.55	0.00423	0.0202	-7.56
Age stratum	45-64	147	No	3	321	2	10.4	0.00623	0.0324	-8.42
Age stratum	45-64	147	No	4	404	2	18.1	0.00495	0.0448	-16.1
Age stratum	45-64	147	No	5	336	0	20.2	0	0.0601	-20.2
Age stratum	45-64	147	No	6	372	4	28.5	0.0108	0.0767	-24.5
Age stratum	45-64	147	No	7	362	6	36.8	0.0166	0.102	-30.8
Age stratum	45-64	147	No	8	370	16	50.9	0.0432	0.138	-34.9
Age stratum	45-64	147	No	9	363	17	72.2	0.0468	0.199	-55.2
Age stratum	45-64	147	No	10	366	97	138	0.265	0.377	-40.9
Age stratum	65-79	249	No	1	524	2	4.88	0.00382	0.00932	-2.88
Age stratum	65-79	249	No	2	380	2	8.57	0.00526	0.0226	-6.57
Age stratum	65-79	249	No	3	413	3	13.4	0.00726	0.0324	-10.4
Age stratum	65-79	249	No	4	498	5	22.2	0.0100	0.0446	-17.2
Age stratum	65-79	249	No	5	565	8	35.1	0.0142	0.0621	-27.1
Age stratum	65-79	249	No	6	458	1	37.9	0.00218	0.0828	-36.9
Age stratum	65-79	249	No	7	449	10	47.9	0.0223	0.107	-37.9
Age stratum	65-79	249	No	8	443	18	63.4	0.0406	0.143	-45.4
Age stratum	65-79	249	No	9	493	34	102	0.0690	0.207	-67.9
Age stratum	65-79	249	No	10	461	166	180	0.360	0.390	-13.7
Age stratum	80+	53	No	1	98	0	1.29	0	0.0132	-1.29
Age stratum	80+	53	No	2	105	2	3.15	0.0190	0.0300	-1.15
Age stratum	80+	53	No	3	109	1	4.93	0.00917	0.0453	-3.93
Age stratum	80+	53	No	4	99	3	5.90	0.0303	0.0596	-2.90
Age stratum	80+	53	No	5	106	2	8.14	0.0189	0.0768	-6.14
Age stratum	80+	53	No	6	100	0	9.92	0	0.0992	-9.92

Axis	Subgroup	Stratum events	Below event floor	Risk group	n	Observed	Expected	Observed rate	Expected rate	Difference
Age stratum	80+	53	No	7	93	3	11.8	0.0323	0.127	-8.84
Age stratum	80+	53	No	8	102	8	18.1	0.0784	0.177	-10.1
Age stratum	80+	53	No	9	97	14	24.1	0.144	0.248	-10.1
Age stratum	80+	53	No	10	103	20	41.0	0.194	0.398	-21.0
Physiological reserve	Frail	400	No	1	604	2	7.77	0.00331	0.0129	-5.77
Physiological reserve	Frail	400	No	2	688	7	20.8	0.0102	0.0302	-13.8
Physiological reserve	Frail	400	No	3	608	3	27.3	0.00493	0.0449	-24.3
Physiological reserve	Frail	400	No	4	577	5	34.6	0.00867	0.0600	-29.6
Physiological reserve	Frail	400	No	5	646	9	49.9	0.0139	0.0773	-40.9
Physiological reserve	Frail	400	No	6	595	12	59.2	0.0202	0.0995	-47.2
Physiological reserve	Frail	400	No	7	586	21	75.1	0.0358	0.128	-54.1
Physiological reserve	Frail	400	No	8	597	28	102	0.0469	0.171	-74.0
Physiological reserve	Frail	400	No	9	636	94	155	0.148	0.243	-60.7
Physiological reserve	Frail	400	No	10	602	219	260	0.364	0.431	-40.7
Physiological reserve	Non-frail	6	Yes	1	94	0	0.285	0	0.00303	-0.285
Physiological reserve	Non-frail	6	Yes	2	143	0	1.77	0	0.0124	-1.77
Physiological reserve	Non-frail	6	Yes	3	61	0	1.22	0	0.0200	-1.22
Physiological reserve	Non-frail	6	Yes	4	106	0	2.90	0	0.0273	-2.90
Physiological reserve	Non-frail	6	Yes	5	110	0	4.10	0	0.0373	-4.10
Physiological reserve	Non-frail	6	Yes	6	96	1	4.50	0.0104	0.0469	-3.50
Physiological reserve	Non-frail	6	Yes	7	113	0	6.73	0	0.0596	-6.73
Physiological reserve	Non-frail	6	Yes	8	97	0	7.72	0	0.0796	-7.72
Physiological reserve	Non-frail	6	Yes	9	98	2	11.1	0.0204	0.113	-9.11
Physiological reserve	Non-frail	6	Yes	10	104	3	21.9	0.0288	0.211	-18.9
Physiological reserve	Pre-frail	98	No	1	382	1	2.35	0.00262	0.00615	-1.35
Physiological reserve	Pre-frail	98	No	2	333	1	5.77	0.00300	0.0173	-4.77
Physiological reserve	Pre-frail	98	No	3	354	2	9.74	0.00565	0.0275	-7.75
Physiological reserve	Pre-frail	98	No	4	293	4	11.0	0.0137	0.0375	-6.99
Physiological reserve	Pre-frail	98	No	5	367	4	18.2	0.0109	0.0496	-14.2
Physiological reserve	Pre-frail	98	No	6	410	5	27.6	0.0122	0.0672	-22.6
Physiological reserve	Pre-frail	98	No	7	372	5	33.1	0.0134	0.0891	-28.1
Physiological reserve	Pre-frail	98	No	8	330	10	39.3	0.0303	0.119	-29.3
Physiological reserve	Pre-frail	98	No	9	368	16	63.2	0.0435	0.172	-47.2
Physiological reserve	Pre-frail	98	No	10	349	50	109	0.143	0.313	-59.3

Supplementary Table 10. Bootstrap replicate accounting and estimability, audited cohort, primary analysis. For every cell, the observed event count, the pre-registered event floor for that metric, the proportion of bootstrap replicates in which the metric was not estimable, and the verdict. Not estimable is the conservative union of the two rules: a cell below its floor is suppressed even when every replicate estimated cleanly, which is the pre-frail calibration slope's situation at 98 events against a floor of 100. A gap inherits the floor failure of any constituent stratum. The prespecified axes are shown in full (123 of 675 cells), of which 25 are not estimable; 287 cells across all axes are not estimable.

Axis	Stratum	Metric	*n*	Events	Event floor	Below floor	Non-estimable replicates (%)	Replicate rule flags	Not estimable
Recorded sex	Female	auroc	3,689	190	25	no	0	no	no
Recorded sex	Female	auprc	3,689	190	25	no	0	no	no
Recorded sex	Female	sensitivity	3,689	190	25	no	0	no	no
Recorded sex	Female	specificity	3,689	190	25	no	0	no	no
Recorded sex	Female	ppv	3,689	190	25	no	0	no	no
Recorded sex	Female	npv	3,689	190	25	no	0	no	no
Recorded sex	Male	auroc	7,030	314	25	no	0	no	no
Recorded sex	Male	auprc	7,030	314	25	no	0	no	no
Recorded sex	Male	sensitivity	7,030	314	25	no	0	no	no
Recorded sex	Male	specificity	7,030	314	25	no	0	no	no
Recorded sex	Male	ppv	7,030	314	25	no	0	no	no
Recorded sex	Male	npv	7,030	314	25	no	0	no	no
Recorded sex	Overall	auroc	10,719	504	25	no	0	no	no
Recorded sex	Overall	auprc	10,719	504	25	no	0	no	no
Recorded sex	Overall	sensitivity	10,719	504	25	no	0	no	no
Recorded sex	Overall	specificity	10,719	504	25	no	0	no	no
Recorded sex	Overall	ppv	10,719	504	25	no	0	no	no
Recorded sex	Overall	npv	10,719	504	25	no	0	no	no
Age stratum	0-17	auroc	367	21	25	**yes**	0	no	**yes**
Age stratum	0-17	auprc	367	21	25	**yes**	0	no	**yes**
Age stratum	0-17	sensitivity	367	21	25	**yes**	0	no	**yes**
Age stratum	0-17	specificity	367	21	25	**yes**	0	no	**yes**
Age stratum	0-17	ppv	367	21	25	**yes**	0	no	**yes**
Age stratum	0-17	npv	367	21	25	**yes**	0	no	**yes**
Age stratum	18-44	auroc	941	34	25	no	0	no	no
Age stratum	18-44	auprc	941	34	25	no	0	no	no
Age stratum	18-44	sensitivity	941	34	25	no	0	no	no
Age stratum	18-44	specificity	941	34	25	no	0	no	no
Age stratum	18-44	ppv	941	34	25	no	0	no	no
Age stratum	18-44	npv	941	34	25	no	0	no	no
Age stratum	45-64	auroc	3,715	147	25	no	0	no	no
Age stratum	45-64	auprc	3,715	147	25	no	0	no	no
Age stratum	45-64	sensitivity	3,715	147	25	no	0	no	no
Age stratum	45-64	specificity	3,715	147	25	no	0	no	no
Age stratum	45-64	ppv	3,715	147	25	no	0	no	no
Age stratum	45-64	npv	3,715	147	25	no	0	no	no
Age stratum	65-79	auroc	4,684	249	25	no	0	no	no

Axis	Stratum	Metric	*n*	Events	Event floor	Below floor	Non-estimable replicates (%)	Replicate rule flags	Not estimable
Age stratum	65-79	auprc	4,684	249	25	no	0	no	no
Age stratum	65-79	sensitivity	4,684	249	25	no	0	no	no
Age stratum	65-79	specificity	4,684	249	25	no	0	no	no
Age stratum	65-79	ppv	4,684	249	25	no	0	no	no
Age stratum	65-79	npv	4,684	249	25	no	0	no	no
Age stratum	80+	auroc	1,012	53	25	no	0	no	no
Age stratum	80+	auprc	1,012	53	25	no	0	no	no
Age stratum	80+	sensitivity	1,012	53	25	no	0	no	no
Age stratum	80+	specificity	1,012	53	25	no	0	no	no
Age stratum	80+	ppv	1,012	53	25	no	0	no	no
Age stratum	80+	npv	1,012	53	25	no	0	no	no
Age stratum	Overall	auroc	10,719	504	25	no	0	no	no
Age stratum	Overall	auprc	10,719	504	25	no	0	no	no
Age stratum	Overall	sensitivity	10,719	504	25	no	0	no	no
Age stratum	Overall	specificity	10,719	504	25	no	0	no	no
Age stratum	Overall	ppv	10,719	504	25	no	0	no	no
Age stratum	Overall	npv	10,719	504	25	no	0	no	no
Physiological reserve	Frail	auroc	6,139	400	25	no	0	no	no
Physiological reserve	Frail	auprc	6,139	400	25	no	0	no	no
Physiological reserve	Frail	sensitivity	6,139	400	25	no	0	no	no
Physiological reserve	Frail	specificity	6,139	400	25	no	0	no	no
Physiological reserve	Frail	ppv	6,139	400	25	no	0	no	no
Physiological reserve	Frail	npv	6,139	400	25	no	0	no	no
Physiological reserve	Non-frail	auroc	1,022	6	25	**yes**	0.400	no	**yes**
Physiological reserve	Non-frail	auprc	1,022	6	25	**yes**	0.400	no	**yes**
Physiological reserve	Non-frail	sensitivity	1,022	6	25	**yes**	0.400	no	**yes**
Physiological reserve	Non-frail	specificity	1,022	6	25	**yes**	0	no	**yes**
Physiological reserve	Non-frail	ppv	1,022	6	25	**yes**	0	no	**yes**
Physiological reserve	Non-frail	npv	1,022	6	25	**yes**	0	no	**yes**
Physiological reserve	Pre-frail	auroc	3,558	98	25	no	0	no	no
Physiological reserve	Pre-frail	auprc	3,558	98	25	no	0	no	no
Physiological reserve	Pre-frail	sensitivity	3,558	98	25	no	0	no	no
Physiological reserve	Pre-frail	specificity	3,558	98	25	no	0	no	no
Physiological reserve	Pre-frail	ppv	3,558	98	25	no	0	no	no
Physiological reserve	Pre-frail	npv	3,558	98	25	no	0	no	no
Physiological reserve	Overall	auroc	10,719	504	25	no	0	no	no
Physiological reserve	Overall	auprc	10,719	504	25	no	0	no	no
Physiological reserve	Overall	sensitivity	10,719	504	25	no	0	no	no
Physiological reserve	Overall	specificity	10,719	504	25	no	0	no	no
Physiological reserve	Overall	ppv	10,719	504	25	no	0	no	no
Physiological reserve	Overall	npv	10,719	504	25	no	0	no	no
Recorded sex	Female	calibration intercept	3,689	190	25	no	0	no	no

Axis	Stratum	Metric	*n*	Events	Event floor	Below floor	Non-estimable replicates (%)	Replicate rule flags	Not estimable
Recorded sex	Female	calibration slope	3,689	190	100	no	0	no	no
Recorded sex	Female	brier	3,689	190	25	no	0	no	no
Recorded sex	Male	calibration intercept	7,030	314	25	no	0	no	no
Recorded sex	Male	calibration slope	7,030	314	100	no	0	no	no
Recorded sex	Male	brier	7,030	314	25	no	0	no	no
Recorded sex	Overall	calibration intercept	10,719	504	25	no	0	no	no
Recorded sex	Overall	calibration slope	10,719	504	100	no	0	no	no
Recorded sex	Overall	brier	10,719	504	25	no	0	no	no
Recorded sex	GAP	max intercept gap	—	—	—	no	0	no	no
Recorded sex	GAP	max slope gap	—	—	—	no	0	no	no
Age stratum	0-17	calibration intercept	367	21	25	**yes**	0	no	**yes**
Age stratum	0-17	calibration slope	367	21	100	**yes**	0	no	**yes**
Age stratum	0-17	brier	367	21	25	**yes**	0	no	**yes**
Age stratum	18-44	calibration intercept	941	34	25	no	0	no	no
Age stratum	18-44	calibration slope	941	34	100	**yes**	0	no	**yes**
Age stratum	18-44	brier	941	34	25	no	0	no	no
Age stratum	45-64	calibration intercept	3,715	147	25	no	0	no	no
Age stratum	45-64	calibration slope	3,715	147	100	no	0	no	no
Age stratum	45-64	brier	3,715	147	25	no	0	no	no
Age stratum	65-79	calibration intercept	4,684	249	25	no	0	no	no
Age stratum	65-79	calibration slope	4,684	249	100	no	0	no	no
Age stratum	65-79	brier	4,684	249	25	no	0	no	no
Age stratum	80+	calibration intercept	1,012	53	25	no	0	no	no
Age stratum	80+	calibration slope	1,012	53	100	**yes**	0	no	**yes**
Age stratum	80+	brier	1,012	53	25	no	0	no	no
Age stratum	Overall	calibration intercept	10,719	504	25	no	0	no	no
Age stratum	Overall	calibration slope	10,719	504	100	no	0	no	no
Age stratum	Overall	brier	10,719	504	25	no	0	no	no
Age stratum	GAP	max intercept gap	—	—	25	**yes**	0	no	**yes**
Age stratum	GAP	max slope gap	—	—	100	**yes**	0	no	**yes**
Physiological reserve	Frail	calibration intercept	6,139	400	25	no	0	no	no
Physiological reserve	Frail	calibration slope	6,139	400	100	no	0	no	no
Physiological reserve	Frail	brier	6,139	400	25	no	0	no	no
Physiological reserve	Non-frail	calibration intercept	1,022	6	25	**yes**	0.400	no	**yes**
Physiological reserve	Non-frail	calibration slope	1,022	6	100	**yes**	0.400	no	**yes**
Physiological reserve	Non-frail	brier	1,022	6	25	**yes**	0	no	**yes**
Physiological reserve	Pre-frail	calibration intercept	3,558	98	25	no	0	no	no
Physiological reserve	Pre-frail	calibration slope	3,558	98	100	**yes**	0	no	**yes**
Physiological reserve	Pre-frail	brier	3,558	98	25	no	0	no	no
Physiological reserve	Overall	calibration intercept	10,719	504	25	no	0	no	no
Physiological reserve	Overall	calibration slope	10,719	504	100	no	0	no	no
Physiological reserve	Overall	brier	10,719	504	25	no	0	no	no

Axis	Stratum	Metric	*n*	Events	Event floor	Below floor	Non-estimable replicates (%)	Replicate rule flags	Not estimable
Physiological reserve	GAP	max intercept gap	—	—	25	**yes**	0	no	**yes**
Physiological reserve	GAP	max slope gap	—	—	100	**yes**	0	no	**yes**

Supplementary Table 11. Resampling-scheme comparison, case-control vs patient, full 2014-2022 cohort. The originally submitted analysis resampled cases and non-cases separately within each subgroup, which fixed the event count in every replicate and estimated uncertainty conditional on the observed subgroup prevalence. The revised estimator resamples patients within subgroup strata, allowing event counts to vary. Point estimates are unchanged; only the intervals move, and they move outward. The width ratio is the patient-level width over the case-control width, so a value above 1 is an interval the submitted analysis reported too narrow. The prespecified axes and the three headline metrics (AUROC, calibration intercept and calibration slope) are shown (63 of 3,772 recomputed cells); the complete frame, including every metric and the case-mix axes, is available with the aggregated subgroup outputs in the study repository. The case-control scheme cannot produce a zero-event replicate at all, so its non-estimable proportions read zero throughout: that is the defect, not a property of the data. Intervals are shown for every cell, including strata below the pre-registered event floors, because the comparison of interval widths is the object of this table; they are not used for inference.

Axis	Stratum	Metric	Estimate	Case-control 95% CI	Patient-level 95% CI	Width, case-control	Width, patient-level	Width ratio
Recorded sex	Overall	auroc	0.876	(0.860, 0.890)	(0.859, 0.892)	0.030	0.034	1.12
Recorded sex	Female	auroc	0.878	(0.850, 0.906)	(0.849, 0.903)	0.057	0.054	0.95
Recorded sex	Male	auroc	0.875	(0.854, 0.891)	(0.854, 0.895)	0.036	0.041	1.14
Recorded sex	GAP	auroc	—	—	—	—	—	—
Recorded sex	GAP	auroc	—	—	—	—	—	—
Recorded sex	GAP	auroc	—	—	—	—	—	—
Recorded sex	GAP	auroc	—	—	—	—	—	—
Age stratum	Overall	auroc	0.876	(0.860, 0.895)	(0.859, 0.892)	0.035	0.034	0.95
Age stratum	0-17	auroc	0.807	(0.694, 0.909)	(0.687, 0.901)	0.215	0.215	1.00
Age stratum	18-44	auroc	0.854	(0.802, 0.908)	(0.792, 0.908)	0.105	0.116	1.10
Age stratum	45-64	auroc	0.888	(0.853, 0.915)	(0.856, 0.916)	0.063	0.060	0.95
Age stratum	65-79	auroc	0.889	(0.866, 0.909)	(0.865, 0.910)	0.043	0.046	1.06
Age stratum	80+	auroc	0.805	(0.744, 0.865)	(0.744, 0.863)	0.120	0.119	0.99
Age stratum	GAP	auroc	—	—	—	—	—	—
Age stratum	GAP	auroc	—	—	—	—	—	—
Age stratum	GAP	auroc	—	—	—	—	—	—
Age stratum	GAP	auroc	—	—	—	—	—	—
Physiological reserve	Overall	auroc	0.876	(0.858, 0.893)	(0.859, 0.892)	0.035	0.034	0.96
Physiological reserve	Frail	auroc	0.878	(0.859, 0.895)	(0.860, 0.896)	0.036	0.036	1.00
Physiological reserve	Non-frail	auroc	0.868	(0.752, 0.969)	(0.728, 0.971)	0.216	0.243	1.12
Physiological reserve	Pre-frail	auroc	0.821	(0.770, 0.867)	(0.774, 0.865)	0.097	0.091	0.94
Physiological reserve	GAP	auroc	—	—	—	—	—	—
Physiological reserve	GAP	auroc	—	—	—	—	—	—
Physiological reserve	GAP	auroc	—	—	—	—	—	—
Physiological reserve	GAP	auroc	—	—	—	—	—	—
Recorded sex	Overall	calibration intercept	-1.020	(-1.042, -0.993)	(-1.116, -0.938)	0.049	0.179	3.68
Recorded sex	Female	calibration intercept	-0.963	(-1.005, -0.920)	(-1.112, -0.821)	0.085	0.291	3.42
Recorded sex	Male	calibration intercept	-1.053	(-1.088, -1.025)	(-1.171, -0.940)	0.063	0.232	3.69
Recorded sex	GAP	calibration intercept	—	—	—	—	—	—
Recorded sex	GAP	calibration intercept	—	—	—	—	—	—
Age stratum	Overall	calibration intercept	-1.020	(-1.047, -0.999)	(-1.116, -0.938)	0.048	0.179	3.74

Axis	Stratum	Metric	Estimate	Case-control 95% CI	Patient-level 95% CI	Width, case-control	Width, patient-level	Width ratio
Age stratum	0-17	calibration intercept	-0.962	(-1.053, -0.886)	(-1.476, -0.571)	0.166	0.904	5.43
Age stratum	18-44	calibration intercept	-1.207	(-1.287, -1.114)	(-1.590, -0.906)	0.173	0.684	3.95
Age stratum	45-64	calibration intercept	-1.162	(-1.206, -1.125)	(-1.343, -1.007)	0.081	0.336	4.14
Age stratum	65-79	calibration intercept	-0.889	(-0.925, -0.855)	(-1.020, -0.768)	0.069	0.252	3.64
Age stratum	80+	calibration intercept	-1.069	(-1.143, -0.991)	(-1.376, -0.810)	0.153	0.566	3.70
Age stratum	GAP	calibration intercept	—	—	—	—	—	—
Age stratum	GAP	calibration intercept	—	—	—	—	—	—
Physiological reserve	Overall	calibration intercept	-1.020	(-1.042, -0.994)	(-1.116, -0.938)	0.048	0.179	3.74
Physiological reserve	Frail	calibration intercept	-0.855	(-0.885, -0.825)	(-0.959, -0.757)	0.060	0.202	3.38
Physiological reserve	Non-frail	calibration intercept	-2.475	(-2.555, -2.386)	(-3.593, -1.831)	0.170	1.762	10.39
Physiological reserve	Pre-frail	calibration intercept	-1.342	(-1.384, -1.306)	(-1.554, -1.158)	0.078	0.396	5.08
Physiological reserve	GAP	calibration intercept	—	—	—	—	—	—
Physiological reserve	GAP	calibration intercept	—	—	—	—	—	—
Recorded sex	Overall	calibration slope	1.566	(1.459, 1.681)	(1.447, 1.688)	0.221	0.241	1.09
Recorded sex	Female	calibration slope	1.561	(1.384, 1.795)	(1.370, 1.762)	0.411	0.392	0.95
Recorded sex	Male	calibration slope	1.568	(1.432, 1.711)	(1.423, 1.730)	0.279	0.307	1.10
Recorded sex	GAP	calibration slope	—	—	—	—	—	—
Recorded sex	GAP	calibration slope	—	—	—	—	—	—
Age stratum	Overall	calibration slope	1.566	(1.465, 1.707)	(1.447, 1.688)	0.241	0.241	1.00
Age stratum	0-17	calibration slope	1.542	(0.924, 2.453)	(0.903, 2.400)	1.529	1.497	0.98
Age stratum	18-44	calibration slope	1.404	(1.048, 1.836)	(1.038, 1.843)	0.787	0.805	1.02
Age stratum	45-64	calibration slope	1.594	(1.392, 1.864)	(1.386, 1.845)	0.472	0.459	0.97
Age stratum	65-79	calibration slope	1.678	(1.505, 1.882)	(1.502, 1.876)	0.378	0.374	0.99
Age stratum	80+	calibration slope	1.164	(0.881, 1.509)	(0.886, 1.518)	0.628	0.632	1.01
Age stratum	GAP	calibration slope	—	—	—	—	—	—
Age stratum	GAP	calibration slope	—	—	—	—	—	—
Physiological reserve	Overall	calibration slope	1.566	(1.464, 1.709)	(1.447, 1.688)	0.245	0.241	0.99
Physiological reserve	Frail	calibration slope	1.594	(1.471, 1.733)	(1.463, 1.741)	0.262	0.278	1.06
Physiological reserve	Non-frail	calibration slope	1.546	(0.786, 2.906)	(0.720, 3.047)	2.120	2.327	1.10
Physiological reserve	Pre-frail	calibration slope	1.282	(1.090, 1.541)	(1.038, 1.567)	0.450	0.530	1.18
Physiological reserve	GAP	calibration slope	—	—	—	—	—	—
Physiological reserve	GAP	calibration slope	—	—	—	—	—	—

Supplementary Table 12. Two-way intersectional cell sparsity (audited cohort, n = 10,719; 504 postoperative infection events). For each of the three prespecified two-way intersections, every cell's size, outcome count, and empirical-Bayes (EB) shrinkage status is reported. Cells with fewer than 50 outcome events were estimated with EB shrinkage of the cell-level postoperative infection rate toward the marginal subgroup rate; these are flagged in the shrinkage column, and the shrinkage factor (0 = no shrinkage, 1 = full shrinkage to the margin) is shown for them. Of the 31 intersectional cells, 22 required shrinkage. The corresponding discrimination gaps for these intersections, including the equalized-odds gap, are reported in Supplementary Table 19.

Intersection	Cell	n	Events	EB shrinkage	Cell rate	Marginal rate	Shrinkage factor
Recorded sex × age stratum	female / 0-17	175	13	Yes	0.0743	0.0515	0.502
Recorded sex × age stratum	female / 18-44	408	13	Yes	0.0319	0.0515	0.302
Recorded sex × age stratum	female / 45-64	1,179	52	No	0.0441	0.0515	0.130
Recorded sex × age stratum	female / 65-79	1,470	93	No	0.0633	0.0515	0.107
Recorded sex × age stratum	female / 80+	457	19	Yes	0.0416	0.0515	0.278
Recorded sex × age stratum	male / 0-17	192	8	Yes	0.0417	0.0447	0.479
Recorded sex × age stratum	male / 18-44	533	21	Yes	0.0394	0.0447	0.249
Recorded sex × age stratum	male / 45-64	2,536	95	No	0.0375	0.0447	0.0650
Recorded sex × age stratum	male / 65-79	3,214	156	No	0.0485	0.0447	0.0520
Recorded sex × age stratum	male / 80+	555	34	Yes	0.0613	0.0447	0.241
Recorded sex × physiological reserve	female / frail	1,666	129	No	0.0774	0.0515	0.112
Recorded sex × physiological reserve	female / non_frail	522	1	Yes	0.00192	0.0515	0.287
Recorded sex × physiological reserve	female / pre_frail	1,501	60	No	0.0400	0.0515	0.123
Recorded sex × physiological reserve	male / frail	4,473	271	No	0.0606	0.0447	0.0448
Recorded sex × physiological reserve	male / non_frail	500	5	Yes	0.0100	0.0447	0.296
Recorded sex × physiological reserve	male / pre_frail	2,057	38	Yes	0.0185	0.0447	0.0926
Age stratum × physiological reserve	0-17 / frail	211	19	Yes	0.0900	0.0572	0.389
Age stratum × physiological reserve	0-17 / non_frail	45	0	Yes	0	0.0572	0.749
Age stratum × physiological reserve	0-17 / pre_frail	111	2	Yes	0.0180	0.0572	0.548
Age stratum × physiological reserve	18-44 / frail	300	21	Yes	0.0700	0.0361	0.309
Age stratum × physiological reserve	18-44 / non_frail	331	2	Yes	0.00604	0.0361	0.289
Age stratum × physiological reserve	18-44 / pre_frail	310	11	Yes	0.0355	0.0361	0.302
Age stratum × physiological reserve	45-64 / frail	1,967	111	No	0.0564	0.0396	0.0640
Age stratum × physiological reserve	45-64 / non_frail	410	3	Yes	0.00732	0.0396	0.247
Age stratum × physiological reserve	45-64 / pre_frail	1,338	33	Yes	0.0247	0.0396	0.0913
Age stratum × physiological reserve	65-79 / frail	3,032	204	No	0.0673	0.0532	0.0424
Age stratum × physiological reserve	65-79 / non_frail	218	1	Yes	0.00459	0.0532	0.381
Age stratum × physiological reserve	65-79 / pre_frail	1,434	44	Yes	0.0307	0.0532	0.0857
Age stratum × physiological reserve	80+ / frail	629	45	Yes	0.0715	0.0524	0.176
Age stratum × physiological reserve	80+ / non_frail	18	0	Yes	0	0.0524	0.882
Age stratum × physiological reserve	80+ / pre_frail	365	8	Yes	0.0219	0.0524	0.269

Supplementary Table 13. Sex × comorbidity interaction screen for the postoperative infection outcome (audited cohort, n = 10,719; 504 postoperative infection events). For each of the 23 screened comorbidities, the odds ratio for postoperative infection (95% Wald CI) and the absolute postoperative infection rate among exposed vs unexposed patients are reported separately within each sex (the data plotted in Figure 3B), followed by the sex × comorbidity interaction: a likelihood-ratio test of nested logistic-regression models (adjusted for sex, age, ASA, and the comorbidity main effect) of whether the comorbidity to postoperative infection association differs between sexes, its Cox–Snell pseudo-R^2 increment (ΔR^2), and the Benjamini-Hochberg FDR-adjusted q-value across the comorbidity family. Each interaction is a 1-degree-of-freedom likelihood-ratio test (binary comorbidity × binary sex). Rows are ordered by interaction q.

Comorbidity	Label	N exposed female	N exposed male	OR female (95% CI)	OR male (95% CI)	Postoperative infection rate female exposed	Postoperative infection rate female unexposed	Postoperative infection rate male exposed	Postoperative infection rate male unexposed	Interaction LRT stat	Interaction df	Interaction p	Interaction q FDR	Interaction delta r2
ParaplegiaHemiplegia	ParaplegiaHemiplegia	134	231	1.87 (1.01, 3.44)	4.65 (3.22, 6.72)	0.0896	0.0501	0.165	0.0406	9.62	1	0.00193	0.0443	0.000897
HypertensionComplicated	HypertensionComplicated	348	788	1.14 (0.706, 1.83)	1.82 (1.35, 2.44)	0.0575	0.0509	0.0723	0.0412	4.73	1	0.0296	0.328	0.000441
RheumatoidArthritis	RheumatoidArthritis	99	124	1.19 (0.516, 2.76)	3.04 (1.75, 5.28)	0.0606	0.0513	0.121	0.0433	4.10	1	0.0428	0.328	0.000383
OtherNeurologicalDisorders	OtherNeurologicalDisorders	201	341	0.962 (0.501, 1.85)	1.67 (1.09, 2.57)	0.0498	0.0516	0.0704	0.0434	2.64	1	0.104	0.399	0.000247
Hypothyroidism	Hypothyroidism	295	185	1.38 (0.854, 2.23)	2.38 (1.45, 3.93)	0.0678	0.0501	0.0973	0.0432	2.81	1	0.0937	0.399	0.000262
ConnectiveTissueDiseaseRheumaticDisease	ConnectiveTissueDiseaseRheumaticDisease	72	69	1.70 (0.726, 3.96)	3.71 (1.88, 7.33)	0.0833	0.0509	0.145	0.0437	2.65	1	0.104	0.399	0.000247
AlcoholAbuse	AlcoholAbuse	45	274	0.855 (0.206, 3.56)	2.35 (1.55, 3.59)	0.0444	0.0516	0.0949	0.0426	1.78	1	0.182	0.522	0.000166
Psychoses	Psychoses	32	28	5.32 (2.27, 12.5)	1.65 (0.390, 6.98)	0.219	0.0500	0.0714	0.0446	1.93	1	0.165	0.522	0.000180
SolidTumorNoMetastasis	SolidTumorNoMetastasis	1,078	1,442	1.51 (1.12, 2.05)	1.23 (0.941, 1.60)	0.0668	0.0452	0.0520	0.0428	0.515	1	0.473	0.603	4.80e-05
MetastaticCarcinoma	MetastaticCarcinoma	511	633	2.08 (1.47, 2.95)	1.46 (1.03, 2.07)	0.0900	0.0453	0.0616	0.0430	0.800	1	0.371	0.603	7.46e-05
RenalDisease	RenalDisease	481	1,117	2.11 (1.49, 3.00)	2.34 (1.82, 3.01)	0.0915	0.0455	0.0833	0.0374	0.865	1	0.352	0.603	8.07e-05
HypertensionUncomplicated	HypertensionUncomplicated	1,242	3,004	1.00 (0.735, 1.36)	0.810 (0.642, 1.02)	0.0515	0.0515	0.0396	0.0484	0.844	1	0.358	0.603	7.87e-05
LiverDisease	LiverDisease	39	170	5.75 (2.69, 12.3)	3.16 (1.97, 5.06)	0.231	0.0496	0.124	0.0427	1.18	1	0.278	0.603	0.000110
Cancer	Cancer	1,107	1,491	1.67 (1.24, 2.25)	1.31 (1.01, 1.70)	0.0705	0.0434	0.0543	0.0421	0.754	1	0.385	0.603	7.04e-05
PepticUlcerDisease	PepticUlcerDisease	20	33	29.4 (11.9, 72.9)	35.1 (17.3, 71.2)	0.600	0.0485	0.606	0.0420	0.660	1	0.417	0.603	6.16e-05
Dementia	Dementia	56	46	1.83 (0.721, 4.63)	2.63 (1.03, 6.71)	0.0893	0.0509	0.109	0.0442	0.458	1	0.498	0.603	4.28e-05
Depression	Depression	249	231	1.58 (0.965, 2.58)	1.86 (1.13, 3.05)	0.0763	0.0497	0.0779	0.0435	0.486	1	0.486	0.603	4.54e-05
DrugAbuse	DrugAbuse	33	93	1.86 (0.561, 6.13)	2.63 (1.35, 5.12)	0.0909	0.0511	0.108	0.0438	0.467	1	0.494	0.603	4.36e-05

Comorbidity	Label	N exposed female	N exposed male	OR female (95% CI)	OR male (95% CI)	Postoperative infection rate female exposed	Postoperative infection rate female unexposed	Postoperative infection rate male exposed	Postoperative infection rate male unexposed	Interaction LRT stat	Interaction df	Interaction p	Interaction q FDR	Interaction delta r2
DeficiencyAnemia	DeficiencyAnemia	23	29	3.94 (1.33, 11.7)	6.94 (2.94, 16.4)	0.174	0.0507	0.241	0.0439	1.19	1	0.275	0.603	0.000111
Lymphoma	Lymphoma	16	25	8.57 (2.95, 24.9)	4.11 (1.40, 12.1)	0.312	0.0504	0.160	0.0443	0.229	1	0.632	0.727	2.14e-05
MildLiverDisease	MildLiverDisease	26	96	3.40 (1.16, 9.96)	2.83 (1.50, 5.36)	0.154	0.0508	0.115	0.0437	0.109	1	0.741	0.812	1.02e-05
HIV	HIV	4		6.17 (0.638, 59.6)		0.250	0.0513			0.00740	1	0.931	0.931	6.90e-07
WeightLoss	WeightLoss	422	510	12.4 (9.11, 16.9)	11.2 (8.76, 14.4)	0.249	0.0260	0.249	0.0287	0.00951	1	0.922	0.931	8.87e-07

Supplementary Table 14. Outcome ascertainment by year of surgery, whole audited extract (2014-2022). For each operative year the table reports whether the coded endpoint is usable that year, the audited procedure count, coded event count and coded event rate, and Present-on-Admission (POA) field coverage in the audited cohort; alongside the corresponding figures in the raw extract: procedure count, patients carrying a bacterial infection code and their share, POA coverage overall and among infected patients, the number of patients labelled as postoperative cases, and the resulting ascertainment fraction. POA coding is effectively absent before 2018: the 2014-2017 years contribute 4,668 audited procedures and 3 coded events, against 6,051 procedures and 501 coded events (8.28%) in 2018-2022. This table supports Supplementary Figure 2.

Year of surgery	Endpoint usable	Audited n	Audited events	Audited event rate %	Audited POA any %	Raw n	Raw inf bacterial	Raw inf bacterial %	Raw POA n only %	Raw POA any %	Raw POA any % among infected	Labelled class1	Ascertainment fraction %
2,014	No	616	0	0	0	4,220	982	23.3	0	0	0	0	0
2,015	No	646	0	0	0	4,535	983	21.7	0	0	0	0	0
2,016	No	1,704	0	0	0	11,431	2,407	21.1	0.0300	0.0300	0	0	0
2,017	No	1,702	3	0.176	3.88	11,838	2,659	22.5	2.55	3.50	5.75	40	1.50
2,018	Yes	1,617	91	5.63	62.6	12,159	2,768	22.8	39.6	59.6	83.1	447	16.1
2,019	Yes	1,105	100	9.05	71.5	11,480	2,721	23.7	42.5	65.7	89.3	639	23.5
2,020	Yes	1,125	98	8.71	86.1	9,833	2,408	24.5	59.0	83.0	98.1	636	26.4
2,021	Yes	1,357	117	8.62	86.1	9,777	2,323	23.8	57.3	81.8	98.8	651	28.0
2,022	Yes	847	95	11.2	87.4	7,262	1,840	25.3	56.0	82.3	98.4	478	25.9

Supplementary Table 15. Ascertainment sensitivity: per-subgroup discrimination in the 2018-2022 window, in which the POA-coded endpoint is ascertainable (n = 6,051; 501 coded events). Within-subgroup AUROC, AUPRC, sensitivity, specificity, PPV and NPV at the t = 0.10 operating threshold, with 95% CIs from 2,000 patient-level bootstrap resamples. Gap statistics per axis are reported in the section header rows (EOD = equal-opportunity difference, PE = predictive-equality difference, EOG = equalized-odds gap, SP = statistical parity). Subgroups marked Yes in the EB shrinkage column used empirical-Bayes shrinkage toward the marginal rate (cell n events < 50). A subgroup estimate is labelled not estimable where the stratum falls below the pre-registered event floor for that metric (see Methods); its point estimate is shown and its interval withheld.

Axis	Subgroup	EB shrinkage	n	Events	Prevalence	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Gap equal opportunity difference	Gap predictive equality	Gap equalized odds Gap	Gap statistical parity
Recorded sex	Overall	No	6,051	501	0.0828	0.855 (0.836, 0.873)	0.440 (0.394, 0.486)	0.900 (0.873, 0.926)	0.590 (0.576, 0.602)	0.165 (0.152, 0.179)	0.985 (0.981, 0.989)	0.00733	0.00779	0.00779	0.0135
Recorded sex	female	No	2,090	189	0.0904	0.857 (0.826, 0.886)	0.456 (0.385, 0.530)	0.905 (0.862, 0.944)	0.584 (0.563, 0.606)	0.178 (0.154, 0.202)	0.984 (0.976, 0.991)	0.00733	0.00779	0.00779	0.0135
Recorded sex	male	No	3,961	312	0.0788	0.853 (0.830, 0.876)	0.430 (0.372, 0.487)	0.897 (0.862, 0.930)	0.592 (0.576, 0.608)	0.158 (0.142, 0.175)	0.985 (0.980, 0.990)	0.00733	0.00779	0.00779	0.0135
Age stratum	Overall	No	6,051	501	0.0828	0.855 (0.836, 0.873)	0.440 (0.394, 0.486)	0.900 (0.873, 0.926)	0.590 (0.576, 0.602)	0.165 (0.152, 0.179)	0.985 (0.981, 0.989)	0.0943	0.179	0.179	0.178
Age stratum	0-17	No	223	21	0.0942	0.803 (not estimable)	0.350 (not estimable)	0.857 (not estimable)	0.441 (not estimable)	0.137 (not estimable)	0.967 (not estimable)	0.0943	0.179	0.179	0.178
Age stratum	18-44	No	518	34	0.0656	0.823 (0.756, 0.884)	0.396 (0.243, 0.540)	0.824 (0.692, 0.938)	0.620 (0.574, 0.662)	0.132 (0.0874, 0.180)	0.980 (0.965, 0.994)	0.0943	0.179	0.179	0.178
Age stratum	45-64	No	2,049	146	0.0713	0.867 (0.831, 0.898)	0.427 (0.347, 0.511)	0.918 (0.872, 0.959)	0.610 (0.588, 0.632)	0.153 (0.131, 0.176)	0.990 (0.984, 0.995)	0.0943	0.179	0.179	0.178
Age stratum	65-79	No	2,698	247	0.0915	0.872 (0.845, 0.896)	0.515 (0.450, 0.580)	0.915 (0.878, 0.948)	0.598 (0.579, 0.617)	0.187 (0.165, 0.209)	0.986 (0.980, 0.991)	0.0943	0.179	0.179	0.178
Age stratum	80+	No	563	53	0.0941	0.768 (0.699, 0.829)	0.288 (0.191, 0.411)	0.849 (0.741, 0.938)	0.502 (0.457, 0.546)	0.151 (0.112, 0.195)	0.970 (0.947, 0.989)	0.0943	0.179	0.179	0.178
Physiological reserve	Overall	No	6,051	501	0.0828	0.855 (0.836, 0.873)	0.440 (0.394, 0.486)	0.900 (0.873, 0.926)	0.590 (0.576, 0.602)	0.165 (0.152, 0.179)	0.985 (0.981, 0.989)	0.136	0.249	0.249	0.292
Physiological reserve	Frail	No	3,684	399	0.108	0.860 (0.839, 0.879)	0.495 (0.442, 0.548)	0.927 (0.901, 0.952)	0.533 (0.516, 0.551)	0.194 (0.176, 0.212)	0.984 (0.978, 0.989)	0.136	0.249	0.249	0.292
Physiological reserve	Non-frail	No	547	6	0.0110	0.840 (not estimable)	0.0985 (not estimable)	0.833 (not estimable)	0.782 (not estimable)	0.0407 (not estimable)	0.998 (not estimable)	0.136	0.249	0.249	0.292
Physiological reserve	Pre-frail	No	1,820	96	0.0527	0.791 (0.737, 0.837)	0.241 (0.168, 0.332)	0.792 (0.701, 0.868)	0.636 (0.614, 0.659)	0.108 (0.0859, 0.130)	0.982 (0.974, 0.989)	0.136	0.249	0.249	0.292

Supplementary Table 16. Ascertainment sensitivity: per-subgroup calibration in the 2018-2022 window (n = 6,051; 501 coded events) for the three prespecified subgroup axes. Within-subgroup calibration intercept (calibration-in-the-large; fixed-slope logit recalibration) and calibration slope with 95% CIs from 2,000 patient-level bootstrap resamples, together with Brier score, expected calibration error (ECE; 10 quantile bins) and integrated calibration index (ICI). Per-axis maximum intercept- and slope-gaps are reported in the section header rows; a maximum gap is labelled not estimable where either contributing stratum falls below the pre-registered event floor, with the point estimate shown and the interval withheld. Subgroups marked Yes in the EB shrinkage column used empirical-Bayes shrinkage toward the marginal rate (cell n events < 50). Compare against Table 3, which reports the same analysis on the audited 2014-2022 cohort. A subgroup estimate is labelled not estimable where the stratum falls below the pre-registered event floor for that metric (see Methods); its point estimate is shown and its interval withheld.

Axis	Subgroup	EB shrinkage	n	Events	Calibration-in-the-large (95% CI)	Calibration slope (95% CI)	Brier (95% CI)	ECE	ICI	Gap max intercept Gap (95% CI)	Gap max slope Gap (95% CI)
Recorded sex	Overall	No	6,051	501	-0.579 (-0.673, -0.489)	1.44 (1.32, 1.58)	0.0620 (0.0582, 0.0659)	0.0472	0.0470	0.0612 (0.00371, 0.250)	0.0738 (0.00492, 0.328)
Recorded sex	female	No	2,090	189	-0.540 (-0.698, -0.390)	1.40 (1.21, 1.61)	0.0660 (0.0594, 0.0728)	0.0465	0.0453	0.0612 (0.00371, 0.250)	0.0738 (0.00492, 0.328)
Recorded sex	male	No	3,961	312	-0.601 (-0.719, -0.491)	1.47 (1.32, 1.64)	0.0599 (0.0552, 0.0645)	0.0479	0.0482	0.0612 (0.00371, 0.250)	0.0738 (0.00492, 0.328)
Age stratum	Overall	No	6,051	501	-0.579 (-0.673, -0.489)	1.44 (1.32, 1.58)	0.0620 (0.0582, 0.0659)	0.0472	0.0470	0.298 (not estimable)	0.600 (not estimable)
Age stratum	0-17	No	223	21	-0.463 (not estimable)	1.47 (not estimable)	0.0763 (not estimable)	0.0519	0.0469	0.298 (not estimable)	0.600 (not estimable)
Age stratum	18-44	No	518	34	-0.762 (-1.18, -0.441)	1.24 (not estimable)	0.0564 (0.0439, 0.0693)	0.0574	0.0547	0.298 (not estimable)	0.600 (not estimable)
Age stratum	45-64	No	2,049	146	-0.710 (-0.884, -0.555)	1.43 (1.22, 1.69)	0.0551 (0.0491, 0.0614)	0.0532	0.0514	0.298 (not estimable)	0.600 (not estimable)
Age stratum	65-79	No	2,698	247	-0.467 (-0.601, -0.344)	1.60 (1.41, 1.82)	0.0630 (0.0572, 0.0689)	0.0462	0.0478	0.298 (not estimable)	0.600 (not estimable)
Age stratum	80+	No	563	53	-0.604 (-0.914, -0.315)	0.997 (not estimable)	0.0817 (0.0683, 0.0971)	0.0567	0.0585	0.298 (not estimable)	0.600 (not estimable)
Physiological reserve	Overall	No	6,051	501	-0.579 (-0.673, -0.489)	1.44 (1.32, 1.58)	0.0620 (0.0582, 0.0659)	0.0472	0.0470	1.56 (not estimable)	0.378 (not estimable)
Physiological reserve	Frail	No	3,684	399	-0.438 (-0.551, -0.340)	1.49 (1.35, 1.65)	0.0748 (0.0694, 0.0799)	0.0441	0.0447	1.56 (not estimable)	0.378 (not estimable)
Physiological reserve	Non-frail	No	547	6	-2.00 (not estimable)	1.36 (not estimable)	0.0169 (not estimable)	0.0589	0.0590	1.56 (not estimable)	0.378 (not estimable)
Physiological reserve	Pre-frail	No	1,820	96	-0.843 (-1.07, -0.653)	1.11 (not estimable)	0.0497 (0.0433, 0.0567)	0.0548	0.0536	1.56 (not estimable)	0.378 (not estimable)

Supplementary Table 17. Ascertainment sensitivity: two-way intersectional cell sparsity, 2018-2022 window (n = 6,051; 501 infection events). For each of the three pre-specified two-way intersections, every cell's size, outcome count, and empirical-Bayes (EB) shrinkage status is reported. Cells with fewer than 50 outcome events were estimated with EB shrinkage of the cell-level coded-infection rate toward the marginal subgroup rate; these are flagged in the shrinkage column, and the shrinkage factor (0 = no shrinkage, 1 = full shrinkage to the margin) is shown for them. Of the 31 intersectional cells, 22 required shrinkage; the same cell structure as Supplementary Table 12, re-estimated on the restricted window. Ascertainment-sensitivity analysis restricted to 2018-2022, the window in which Present-on-Admission coding (which the outcome requires) is routinely recorded. Before 2018 an infection whose postoperative timing could not be established was excluded at cohort construction rather than recorded as a negative outcome, so the 4,668 procedures from 2014-2017 carry 3 coded events and the audited cohort's coded prevalence is 4.70% against 8.28% here.

Intersection	Cell	n	Events	EB shrinkage	Cell rate	Marginal rate	Shrinkage factor
Recorded sex × age stratum	female / 0-17	109	13	Yes	0.119	0.0904	0.617
Recorded sex × age stratum	female / 18-44	220	13	Yes	0.0591	0.0904	0.444
Recorded sex × age stratum	female / 45-64	683	52	No	0.0761	0.0904	0.204
Recorded sex × age stratum	female / 65-79	839	92	No	0.110	0.0904	0.173
Recorded sex × age stratum	female / 80+	239	19	Yes	0.0795	0.0904	0.423
Recorded sex × age stratum	male / 0-17	114	8	Yes	0.0702	0.0788	0.606
Recorded sex × age stratum	male / 18-44	298	21	Yes	0.0705	0.0788	0.370
Recorded sex × age stratum	male / 45-64	1,366	94	No	0.0688	0.0788	0.114
Recorded sex × age stratum	male / 65-79	1,859	155	No	0.0834	0.0788	0.0862
Recorded sex × age stratum	male / 80+	324	34	Yes	0.105	0.0788	0.351
Recorded sex × physiological reserve	female / frail	1,010	129	No	0.128	0.0904	0.171
Recorded sex × physiological reserve	female / non_frail	275	1	Yes	0.00364	0.0904	0.432
Recorded sex × physiological reserve	female / pre_frail	805	59	No	0.0733	0.0904	0.206
Recorded sex × physiological reserve	male / frail	2,674	270	No	0.101	0.0788	0.0724
Recorded sex × physiological reserve	male / non_frail	272	5	Yes	0.0184	0.0788	0.434
Recorded sex × physiological reserve	male / pre_frail	1,015	37	Yes	0.0365	0.0788	0.171
Age stratum × physiological reserve	0-17 / frail	128	19	Yes	0.148	0.0942	0.511
Age stratum × physiological reserve	0-17 / non_frail	28	0	Yes	0	0.0942	0.827
Age stratum × physiological reserve	0-17 / pre_frail	67	2	Yes	0.0299	0.0942	0.666
Age stratum × physiological reserve	18-44 / frail	183	21	Yes	0.115	0.0656	0.422
Age stratum × physiological reserve	18-44 / non_frail	182	2	Yes	0.0110	0.0656	0.423
Age stratum × physiological reserve	18-44 / pre_frail	153	11	Yes	0.0719	0.0656	0.466
Age stratum × physiological reserve	45-64 / frail	1,167	110	No	0.0943	0.0713	0.103
Age stratum × physiological reserve	45-64 / non_frail	210	3	Yes	0.0143	0.0713	0.389
Age stratum × physiological reserve	45-64 / pre_frail	672	33	Yes	0.0491	0.0713	0.166
Age stratum × physiological reserve	65-79 / frail	1,846	204	No	0.111	0.0915	0.0675
Age stratum × physiological reserve	65-79 / non_frail	118	1	Yes	0.00847	0.0915	0.531
Age stratum × physiological reserve	65-79 / pre_frail	734	42	Yes	0.0572	0.0915	0.154
Age stratum × physiological reserve	80+ / frail	360	45	Yes	0.125	0.0941	0.271
Age stratum × physiological reserve	80+ / non_frail	9	0	Yes	0	0.0941	0.937
Age stratum × physiological reserve	80+ / pre_frail	194	8	Yes	0.0412	0.0941	0.408

Supplementary Table 18. Case ascertainment by physiological-reserve and age stratum, 2018-2022 window. Among cohort-eligible patients carrying a bacterial infection code in 2018-2022, the proportion recorded as a coded case, by stratum, with 95% CIs, POA field coverage, coded prevalence and capture-equalized prevalence. The ordering is monotone in physiological reserve; not every adjacent pair separates. POA field coverage is near-universal across the strata, which locates the differential in adjudication rather than in data availability. The capture-equalised column raises each stratum to the best-ascertained stratum's capture and is an upper bound, not a corrected estimate: it assumes every unlabelled infection is a true postoperative case, which is the question Present-on-Admission coding exists to decide. Because physiological reserve is itself derived from diagnosis coding, residual confounding runs in both directions.

Axis	Stratum	Infected patients (cohort-eligible)	Coded cases	Ascertainment % (95% CI)	POA field coverage (%)	Coded prevalence (%)	Capture-equalised prevalence (%)
reserve	Non-frail	40	6	15.0 (7.1 to 29.1)	92.5	1.10	2.35
reserve	Pre-frail	397	96	24.2 (20.2 to 28.6)	95.5	5.27	7.16
reserve	Frail	1,190	399	33.5 (30.9 to 36.3)	98	10.8	10.8
Age stratum	0-17	33	21	63.6 (46.6 to 77.8)	100	9.42	9.42
Age stratum	18-44	135	34	25.2 (18.6 to 33.1)	94.1	6.56	15.1
Age stratum	45-64	487	146	30.0 (26.1 to 34.2)	97.3	7.13	14.0
Age stratum	65-79	764	247	32.3 (29.1 to 35.7)	97.5	9.15	16.6
Age stratum	80+	208	53	25.5 (20.0 to 31.8)	97.6	9.41	20.6

Supplementary Table 19. Per-subgroup discrimination on the audited cohort (n = 10,719; 504 infection events) for the case-mix reference axes (surgical specialty, procedure cluster) and the three prespecified pairwise intersections. Nine surgical specialties, eleven procedure clusters and an unclassified bucket, and the three prespecified pairwise intersections are shown. Within-subgroup AUROC, AUPRC, sensitivity, specificity, PPV and NPV at the t = 0.10 operating threshold are reported with 95% CIs from 2,000 patient-level bootstrap resamples drawn within subgroup strata. Gap statistics per axis are reported in the section header rows (EOD = equal-opportunity difference, PE = predictive-equality difference, EOG = equalized-odds gap, SP = statistical parity). Subgroups marked Yes in the EB shrinkage column used empirical-Bayes shrinkage toward the marginal rate (cell n events < 50); shrunk estimates carry no bootstrap interval, so a shrunk cell that meets its event floor is shown as a point estimate without an interval. The equalized-odds gap is 0.516 on the procedure-cluster axis and 0.373 on the surgical-specialty axis; among the intersections it is largest for age × physiological reserve (0.397). A subgroup estimate is labelled not estimable where the stratum falls below the pre-registered event floor for that metric (see Methods); its point estimate is shown and its interval withheld.

Axis	Subgroup	EB shrinkage	n	Events	Prevalence	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Gap equal opportunity difference	Gap predictive equality	Gap equalized odds Gap	Gap statistical parity
Surgical specialty	Overall	No	10,719	504	0.0470	0.876 (0.859, 0.892)	0.376 (0.332, 0.422)	0.899 (0.871, 0.925)	0.647 (0.638, 0.656)	0.112 (0.102, 0.121)	0.992 (0.990, 0.994)	0.333	0.373	0.373	0.386
Surgical specialty	ORL	No	233	15	0.0644	0.820 (not estimable)	0.221 (not estimable)	1 (not estimable)	0.408 (not estimable)	0.104 (not estimable)	1 (not estimable)	0.333	0.373	0.373	0.386
Surgical specialty	Gynaecology	No	452	24	0.0531	0.821 (not estimable)	0.395 (not estimable)	0.750 (not estimable)	0.689 (not estimable)	0.119 (not estimable)	0.980 (not estimable)	0.333	0.373	0.373	0.386
Surgical specialty	Neurosurgery	No	1,117	26	0.0233	0.871 (0.785, 0.939)	0.434 (0.240, 0.621)	0.808 (0.643, 0.955)	0.756 (0.731, 0.780)	0.0732 (0.0440, 0.107)	0.994 (0.988, 0.999)	0.333	0.373	0.373	0.386
Surgical specialty	Orthopaedics	No	1,690	76	0.0450	0.852 (0.808, 0.887)	0.232 (0.160, 0.332)	0.908 (0.835, 0.965)	0.633 (0.610, 0.657)	0.104 (0.0805, 0.127)	0.993 (0.987, 0.997)	0.333	0.373	0.373	0.386
Surgical specialty	Paediatrics	No	65	3	0.0462	0.653 (not estimable)	0.181 (not estimable)	0.667 (not estimable)	0.452 (not estimable)	0.0556 (not estimable)	0.966 (not estimable)	0.333	0.373	0.373	0.386
Surgical specialty	Thoracic surgery	No	723	14	0.0194	0.945 (not estimable)	0.411 (not estimable)	0.929 (not estimable)	0.781 (not estimable)	0.0774 (not estimable)	0.998 (not estimable)	0.333	0.373	0.373	0.386
Surgical specialty	Urology	No	349	19	0.0544	0.785 (not estimable)	0.331 (not estimable)	0.737 (not estimable)	0.624 (not estimable)	0.101 (not estimable)	0.976 (not estimable)	0.333	0.373	0.373	0.386
Surgical specialty	Vascular surgery	No	4,852	184	0.0379	0.866 (0.834, 0.896)	0.323 (0.257, 0.395)	0.886 (0.837, 0.929)	0.656 (0.643, 0.670)	0.0921 (0.0788, 0.106)	0.993 (0.990, 0.996)	0.333	0.373	0.373	0.386
Surgical specialty	Visceral surgery	No	1,238	143	0.116	0.883 (0.854, 0.909)	0.557 (0.472, 0.640)	0.965 (0.932, 0.993)	0.482 (0.453, 0.510)	0.196 (0.169, 0.226)	0.991 (0.981, 0.998)	0.333	0.373	0.373	0.386
Procedure cluster	Overall	No	10,719	504	0.0470	0.876 (0.859, 0.892)	0.376 (0.332, 0.422)	0.899 (0.871, 0.925)	0.647 (0.638, 0.656)	0.112 (0.102, 0.121)	0.992 (0.990, 0.994)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_1 (NervousSystem)	No	903	22	0.0244	0.879 (not estimable)	0.460 (not estimable)	0.818 (not estimable)	0.751 (not estimable)	0.0759 (not estimable)	0.994 (not estimable)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_10 (UroAndrological+Urologic)	No	346	28	0.0809	0.789 (0.673, 0.889)	0.383 (0.233, 0.594)	0.786 (0.625, 0.933)	0.560 (0.505, 0.614)	0.136 (0.0833, 0.193)	0.967 (0.938, 0.989)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_11 (Gynecological)	No	303	9	0.0297	0.761 (not estimable)	0.239 (not estimable)	0.556 (not estimable)	0.769 (not estimable)	0.0685 (not estimable)	0.983 (not estimable)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_12 (Musculoskeletal)	No	1,623	62	0.0382	0.856 (0.810, 0.894)	0.192 (0.135, 0.294)	0.919 (0.844, 0.981)	0.654 (0.631, 0.676)	0.0955 (0.0728, 0.119)	0.995 (0.990, 0.999)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_13 (Cutaneous)	No	213	43	0.202	0.796 (0.716, 0.871)	0.561 (0.419, 0.700)	0.930 (0.846, 1)	0.253 (0.192, 0.318)	0.240 (0.179, 0.301)	0.935 (0.857, 1)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_3 (Thoracic)	No	807	22	0.0273	0.941 (not estimable)	0.482 (not estimable)	0.955 (not estimable)	0.753 (not estimable)	0.0977 (not estimable)	0.998 (not estimable)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_4 (Cardiac)	No	3,764	118	0.0313	0.841 (0.793, 0.883)	0.251 (0.188, 0.338)	0.839 (0.767, 0.902)	0.673 (0.658, 0.687)	0.0767 (0.0623, 0.0918)	0.992 (0.989, 0.996)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_5 (Vascular)	No	769	30	0.0390	0.893 (0.827, 0.943)	0.392 (0.225, 0.565)	0.933 (0.823, 1)	0.636 (0.602, 0.672)	0.0943 (0.0625, 0.130)	0.996 (0.989, 1)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_6 (AbdominalWall)	No	530	67	0.126	0.889 (0.844, 0.927)	0.625 (0.516, 0.728)	0.955 (0.901, 1)	0.471 (0.425, 0.514)	0.207 (0.163, 0.253)	0.986 (0.969, 1)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_7 (GastroIntestinal)	No	298	36	0.121	0.864 (0.813, 0.913)	0.521 (0.365, 0.681)	1 (1, 1)	0.309 (0.256, 0.366)	0.166 (0.117, 0.216)	1 (1, 1)	0.444	0.516	0.516	0.543
Procedure cluster	chopcluster_8 (Hepatobiliary)	No	306	23	0.0752	0.802 (not estimable)	0.327 (not estimable)	0.870 (not estimable)	0.551 (not estimable)	0.136 (not estimable)	0.981 (not estimable)	0.444	0.516	0.516	0.543
Procedure cluster	Unclassified	No	857	44	0.0513	0.900 (0.866, 0.932)	0.388 (0.256, 0.531)	0.977 (0.927, 1)	0.625 (0.590, 0.658)	0.124 (0.0877, 0.159)	0.998 (0.994, 1)	0.444	0.516	0.516	0.543
Recorded sex × age stratum	Overall	No	10,719	504	0.0470	0.876 (0.859, 0.892)	0.376 (0.332, 0.422)	0.899 (0.871, 0.925)	0.647 (0.638, 0.656)	0.112 (0.102, 0.121)	0.992 (0.990, 0.994)	0.112	0.130	0.130	0.142

Axis	Subgroup	EB shrinkage	n	Events	Prevalence	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Gap equal opportunity difference	Gap predictive equality	Gap equalized odds Gap	Gap statistical parity
Recorded sex × age stratum	female / 0-17	Yes	175	13	0.0628	0.902 (not estimable)	0.408 (not estimable)	0.907 (not estimable)	0.549 (not estimable)	0.122 (not estimable)	0.994 (not estimable)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	female / 18-44	Yes	408	13	0.0378	0.873 (not estimable)	0.340 (not estimable)	0.896 (not estimable)	0.663 (not estimable)	0.101 (not estimable)	0.992 (not estimable)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	female / 45-64	No	1,179	52	0.0441	0.860 (0.788, 0.919)	0.377 (0.248, 0.517)	0.846 (0.738, 0.941)	0.674 (0.647, 0.702)	0.107 (0.0781, 0.139)	0.990 (0.982, 0.996)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	female / 65-79	No	1,470	93	0.0633	0.900 (0.865, 0.931)	0.545 (0.451, 0.640)	0.935 (0.882, 0.980)	0.645 (0.620, 0.670)	0.151 (0.120, 0.181)	0.993 (0.988, 0.998)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	female / 80+	Yes	457	19	0.0443	0.814 (not estimable)	0.166 (not estimable)	0.894 (not estimable)	0.596 (not estimable)	0.0983 (not estimable)	0.990 (not estimable)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	male / 0-17	Yes	192	8	0.0431	0.648 (not estimable)	0.119 (not estimable)	0.886 (not estimable)	0.560 (not estimable)	0.0859 (not estimable)	0.984 (not estimable)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	male / 18-44	Yes	533	21	0.0407	0.844 (not estimable)	0.329 (not estimable)	0.889 (not estimable)	0.679 (not estimable)	0.102 (not estimable)	0.990 (not estimable)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	male / 45-64	No	2,536	95	0.0375	0.903 (0.873, 0.929)	0.392 (0.301, 0.494)	0.958 (0.915, 0.991)	0.662 (0.644, 0.681)	0.0993 (0.0806, 0.119)	0.998 (0.995, 0.999)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	male / 65-79	No	3,214	156	0.0485	0.882 (0.848, 0.911)	0.378 (0.307, 0.463)	0.897 (0.843, 0.942)	0.658 (0.641, 0.675)	0.118 (0.100, 0.137)	0.992 (0.988, 0.996)	0.112	0.130	0.130	0.142
Recorded sex × age stratum	male / 80+	Yes	555	34	0.0573	0.798	0.269	0.891	0.595	0.112	0.987	0.112	0.130	0.130	0.142
Recorded sex × physiological reserve	Overall	No	10,719	504	0.0470	0.876 (0.859, 0.892)	0.376 (0.332, 0.422)	0.899 (0.871, 0.925)	0.647 (0.638, 0.656)	0.112 (0.102, 0.121)	0.992 (0.990, 0.994)	0.121	0.255	0.255	0.284
Recorded sex × physiological reserve	female / frail	No	1,666	129	0.0774	0.888 (0.855, 0.918)	0.492 (0.408, 0.577)	0.938 (0.892, 0.977)	0.530 (0.505, 0.556)	0.144 (0.121, 0.167)	0.990 (0.983, 0.996)	0.121	0.255	0.255	0.284
Recorded sex × physiological reserve	female / non_frail	Yes	522	1	0.0161	0.996 (not estimable)	0.333 (not estimable)	0.900 (not estimable)	0.748 (not estimable)	0.0816 (not estimable)	0.997 (not estimable)	0.121	0.255	0.255	0.284
Recorded sex × physiological reserve	female / pre_frail	No	1,501	60	0.0400	0.823 (0.761, 0.877)	0.226 (0.148, 0.345)	0.817 (0.710, 0.907)	0.703 (0.679, 0.726)	0.103 (0.0761, 0.131)	0.989 (0.983, 0.995)	0.121	0.255	0.255	0.284
Recorded sex × physiological reserve	male / frail	No	4,473	271	0.0606	0.874 (0.850, 0.896)	0.395 (0.335, 0.459)	0.923 (0.891, 0.954)	0.604 (0.589, 0.619)	0.130 (0.115, 0.145)	0.992 (0.988, 0.995)	0.121	0.255	0.255	0.284
Recorded sex × physiological reserve	male / non_frail	Yes	500	5	0.0203	0.857 (not estimable)	0.0707 (not estimable)	0.896 (not estimable)	0.785 (not estimable)	0.0908 (not estimable)	0.996 (not estimable)	0.121	0.255	0.255	0.284
Recorded sex × physiological reserve	male / pre_frail	Yes	2,057	38	0.0209	0.816	0.169	0.873	0.696	0.0600	0.993	0.121	0.255	0.255	0.284
Age stratum × physiological reserve	Overall	No	10,719	504	0.0470	0.876 (0.859, 0.892)	0.376 (0.332, 0.422)	0.899 (0.871, 0.925)	0.647 (0.638, 0.656)	0.112 (0.102, 0.121)	0.992 (0.990, 0.994)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	0-17 / frail	Yes	211	19	0.0773	0.791 (not estimable)	0.282 (not estimable)	0.862 (not estimable)	0.393 (not estimable)	0.104 (not estimable)	0.978 (not estimable)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	0-17 / non_frail	Yes	45	0	0.0429				0.711 (not estimable)	0 (not estimable)	1 (not estimable)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	0-17 / pre_frail	Yes	111	2	0.0395	0.539 (not estimable)	0.0425 (not estimable)	0.852 (not estimable)	0.510 (not estimable)	0.0705 (not estimable)	0.982 (not estimable)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	18-44 / frail	Yes	300	21	0.0595	0.864 (not estimable)	0.463 (not estimable)	0.835 (not estimable)	0.565 (not estimable)	0.106 (not estimable)	0.988 (not estimable)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	18-44 / non_frail	Yes	331	2	0.0147	0.869 (not estimable)	0.0314 (not estimable)	0.826 (not estimable)	0.789 (not estimable)	0.0727 (not estimable)	0.997 (not estimable)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	18-44 / pre_frail	Yes	310	11	0.0357	0.744 (not estimable)	0.141 (not estimable)	0.809 (not estimable)	0.680 (not estimable)	0.0796 (not estimable)	0.985 (not estimable)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	45-64 / frail	No	1,967	111	0.0564	0.898 (0.869, 0.926)	0.444 (0.354, 0.540)	0.964 (0.927, 0.992)	0.603 (0.581, 0.625)	0.127 (0.105, 0.149)	0.996 (0.993, 0.999)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	45-64 / non_frail	Yes	410	3	0.0153	0.971 (not estimable)	0.307 (not estimable)	0.920 (not estimable)	0.790 (not estimable)	0.0808 (not estimable)	0.999 (not estimable)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	45-64 / pre_frail	Yes	1,338	33	0.0260	0.808	0.175	0.887	0.701	0.0709	0.992	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	65-79 / frail	No	3,032	204	0.0673	0.889 (0.862, 0.914)	0.487 (0.416, 0.560)	0.922 (0.884, 0.958)	0.610 (0.592, 0.627)	0.146 (0.126, 0.165)	0.991 (0.986, 0.995)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	65-79 / non_frail	Yes	218	1	0.0231	0.592 (not estimable)	0.0104 (not estimable)	0.905 (not estimable)	0.742 (not estimable)	0.0971 (not estimable)	0.993 (not estimable)	0.155	0.397	0.397	0.416

Axis	Subgroup	EB shrinkage	n	Events	Prevalence	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Gap equal opportunity difference	Gap predictive equality	Gap equalized odds Gap	Gap statistical parity
Age stratum × physiological reserve	65-79 / pre_frail	Yes	1,434	44	0.0326	0.870	0.247	0.905	0.714	0.100	0.995	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	80+ / frail	Yes	629	45	0.0682	0.796	0.253	0.859	0.523	0.116	0.984	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	80+ / non_frail	Yes	18	0	0.0462				0.667 (not estimable)	0 (not estimable)	1 (not estimable)	0.155	0.397	0.397	0.416
Age stratum × physiological reserve	80+ / pre_frail	Yes	365	8	0.0301	0.740 (not estimable)	0.209 (not estimable)	0.836 (not estimable)	0.652 (not estimable)	0.0727 (not estimable)	0.987 (not estimable)	0.155	0.397	0.397	0.416

Supplementary Table 20. Wasserstein-barycenter score postprocessing of the end-of-surgery infection prediction model on the procedure-cluster axis, before vs after (audited cohort, n = 10,719; 504 postoperative infection events; nine clusters shown, together 9,253 patients and 428 events). Wasserstein postprocessing was fitted independently across the procedure-cluster (surgical group) axis under a demographic-parity constraint: each cluster's predicted-risk distribution is mapped, by a within-cluster quantile-to-quantile transformation, onto the group-size-weighted 1-D Wasserstein barycenter of the per-cluster calibration-set score distributions. The transform is monotone within each cluster, so within-cluster rank order (and AUROC) is preserved exactly; it equalizes the score distribution across clusters but does not target calibration. For each cluster the Brier score, calibration-in-the-large (intercept; 0 = perfectly calibrated), and calibration slope (1 = perfectly calibrated) are reported before and after postprocessing. The audit's leveling-down classifier compares the relative change in Brier score against a small tolerance band, classifying each cluster as leveled down, neutral, or improved; of the nine clusters shown, five are leveled down, three are neutral, and one (GastroIntestinal) improved. Low-event clusters 8 (Hepatobiliary) and 11 (Gynecological) and the unclassified bucket (together 1,466 patients, 76 events) did not meet the minimum per-group sample for a within-group fit; those patients retain their original predictions and are not shown.

Axis	Subgroup	n	Events	Brier before	Brier after	Intercept before	Intercept after	Slope before	Slope after	Levelling down
CHOP procedure cluster	NervousSystem (c1)	903	22	0.0231	0.0298	-1.32	-1.73	1.83	1.98	levelled_down
CHOP procedure cluster	Thoracic (c3)	807	22	0.0258	0.0301	-1.35	-1.61	1.94	2.38	levelled_down
CHOP procedure cluster	Cardiac (c4)	3,764	118	0.0320	0.0352	-1.27	-1.46	1.58	1.54	levelled_down
CHOP procedure cluster	Vascular (c5)	769	30	0.0364	0.0369	-1.20	-1.22	1.73	1.90	neutral
CHOP procedure cluster	AbdominalWall (c6)	530	67	0.0783	0.0815	-0.565	0.177	1.62	2.03	levelled_down
CHOP procedure cluster	GastroIntestinal (c7)	298	36	0.0895	0.0823	-0.788	0.111	1.75	1.66	improved
CHOP procedure cluster	UroAndrological (c10)	346	28	0.0639	0.0630	-0.493	-0.397	1.36	1.18	neutral
CHOP procedure cluster	Musculoskeletal (c12)	1,623	62	0.0400	0.0408	-1.17	-1.24	1.34	1.41	neutral
CHOP procedure cluster	Cutaneous (c13)	213	43	0.127	0.141	-0.112	0.820	1.38	1.22	levelled_down

Supplementary methods

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

Guillen-Ramirez, et al.

These Supplementary Methods give formal definitions, estimators, and equations underlying the subgroup deployment-readiness audit. They are the technical companion to the main-text Methods, which carries the design rationale and a plain-language description of each measure.

Notation

The audited cohort has n surgeries indexed by $i = 1, \ldots, n$. For each surgery:

- $Y_i \in \{0,1\}$ is the binary outcome (1 = bacterial postoperative infection within the index admission);
- $p_i \in [0,1]$ is the model's predicted probability that $Y_i = 1$, obtained as an out-of-fold prediction from 5-fold stratified cross-validation of the end-of-surgery model [1] , with SMOTE applied only within each training fold after fold assignment and never to held-out observations;
- $A_i \in G$ is the value of the subgroup axis — recorded sex, age stratum, or physiological reserve (e.g. {female, male} for sex; {frail, pre-frail, non-frail} for physiological reserve).

For a subgroup g, let $I_g = \{i: A_i = g\}$ be its surgeries, $n_g = | I_g |$ its size, n_g^+ its event count, $n_g^- = n_g - n_g^+$ its non-event count, and $\pi_g^+ = n_g^+ / n_g$ its outcome prevalence. Population-level quantities drop the g subscript. At a decision threshold $t \in [0,1]$ a surgery is flagged when $p_i \geq t$. Within subgroup g the four standard counts are true positives $\mathrm{TP}_g(t)$ (flagged events: $p_i \geq t, Y_i = 1$), false positives $\mathrm{FP}_g(t)$ (flagged non-events: $p_i \geq t, Y_i = 0$), false negatives $\mathrm{FN}_g(t)$ (missed events: $p_i < t, Y_i = 1$), and true negatives $\mathrm{TN}_g(t)$ (correctly unflagged: $p_i < t, Y_i = 0$).

Discrimination metrics

AUROC. The area under the receiver-operating-characteristic curve is the probability that the model assigns higher risk to a randomly chosen event case than to a randomly chosen non-event case in the same subgroup,

$$\mathrm{AUROC}_g = \Pr(p_i > p_j \mid i, j \in I_g, Y_i = 1, Y_j = 0),$$

estimated by the normalized Mann–Whitney U statistic restricted to I_g, with ties broken by averaging.

AUPRC. The area under the within-subgroup precision–recall curve, estimated as the trapezoidal sum over the operating points $\left(\mathrm{TPR}_g(t), \mathrm{PPV}_g(t)\right)$ swept across all observed thresholds. It is the natural complement to AUROC at low prevalence.

Per-threshold rates. At the operating threshold $t = 0.10$ for intensifying postoperative-infection surveillance, prespecified as a reference operating anchor rather than an evidence-derived optimum,

$$\mathrm{TPR}_g(t) = \frac{\mathrm{TP}_g(t)}{\mathrm{TP}_g(t) + \mathrm{FN}_g(t)}, \mathrm{FPR}_g(t) = \frac{\mathrm{FP}_g(t)}{\mathrm{FP}_g(t) + \mathrm{TN}_g(t)},$$

$$\mathrm{PPV}_g(t) = \frac{\mathrm{TP}_g(t)}{\mathrm{TP}_g(t) + \mathrm{FP}_g(t)}, \mathrm{NPV}_g(t) = \frac{\mathrm{TN}_g(t)}{\mathrm{TN}_g(t) + \mathrm{FN}_g(t)},$$

with specificity $\mathrm{TNR}_g(t) = 1 - \mathrm{FPR}_g(t)$. Sensitivity (TPR) and specificity (TNR) are reported in the main discrimination tables; PPV and NPV are provided in the per-subgroup supplementary data files. Net benefit is additionally evaluated at seven thresholds from 0.05 to 0.20 as a sensitivity analysis, together with the threshold at which each subgroup's net benefit crosses zero (Supplementary Table 7), while the decision curves are drawn across the wider 0.02 to 0.40 band.

Gap statistics. For a pair of subgroups g, g', all evaluated at the same t:

$$\Delta_{\mathrm{EOpp}} = \mathrm{TPR}_g - \mathrm{TPR}_{g'} \text{(equal-opportunity / sensitivity gap)},$$

$$\Delta_{\mathrm{PE}} = \mathrm{FPR}_g - \mathrm{FPR}_{g'} \text{(predictive-equality / false-alarm-rate gap)},$$

$$\Delta_{\mathrm{SP}} = \frac{\mathrm{TP}_g + \mathrm{FP}_g}{n_g} - \frac{\mathrm{TP}_{g'} + \mathrm{FP}_{g'}}{n_{g'}} \text{(demographic-parity / flag-rate gap)}.$$

The equalized-odds gap is the larger of the two error-rate gaps, $\Delta_{\mathrm{EO}} = \max\left(\left|\Delta_{EOpp}\right|, \left|\Delta_{PE}\right|\right)$. For attributes with more than two levels, each gap is summarized by its largest absolute value over all pairs of levels. These threshold-specific parity gaps are reported as descriptive characteristics of the decision rule at a fixed threshold. The primary deployment criteria are subgroup calibration and subgroup clinical utility; a parity gap is not treated as evidence of unfairness on its own, and raw differences across age, physiological reserve, and procedure-cluster strata are interpreted in the context of case mix and outcome prevalence.

Effect size. Each rate-based gap is reported with Cohen's h on the two underlying proportions $p_g, p_{g'}$,

$$h(p_g, p_{g'}) = 2\left(arcsin\sqrt{p}_g - arcsin\sqrt{p}_{g'}\right),$$

an effect-size magnitude less sensitive to the base rate than the raw difference.

Calibration metrics

A model is calibrated within subgroup g if, among surgeries assigned predicted risk p, the observed infection rate is also p, i.e. $E[\,Y \mid p, A = g\,] = p$. Calibration was assessed three ways.

(i) Calibration curves. The conditional outcome probability $E[\,Y \mid p, A = g\,]$ is estimated as a smooth function of p by LOESS, with 95% bootstrap bands. The identity line $y = p$ is perfect calibration; a curve below the line indicates overestimation of risk.

(ii) Calibration intercept and slope [2]. On the log-odds scale $l_i = \log\frac{p_i}{1-p_i}$, the intercept α_g (calibration-in-the-large) is the constant of a one-parameter logistic regression on I_g with the slope of l_i fixed to 1,

$$\text{logit}\,\Pr(Y_i = 1) = \alpha_g + l_i, i \in I_g.$$

$\alpha_g = 0$ is unbiased; $\alpha_g < 0$ indicates systematic over-prediction (as observed globally here, $\alpha = -1.02$). The slope β_g is the coefficient of l_i when both terms are free,

$$\text{logit}\,\Pr(Y_i = 1) = \alpha_g^{(2)} + \beta_g\, l_i, i \in I_g.$$

$\beta_g = 1$ means the predicted-risk gradient is correctly scaled in g; $\beta_g < 1$ means predictions are too spread out (over-extreme), $\beta_g > 1$ too compressed.

(iii) Brier score, ECE, and ICI. The Brier score is the mean squared distance between predicted probability and outcome,

$$\text{BS}_g = \frac{1}{n_g}\sum_{i \in I_g}(\,p_i - Y_i)^2.$$

Using $K = 10$ equal-frequency bins, with bin k containing $n_{g,k}$ predictions of mean predicted risk $p_{g,k}$ and observed event rate $y_{g,k}$, the expected calibration error is the prevalence-weighted mean absolute gap between predicted and observed risk,

$$\text{ECE}_g = \frac{1}{n_g}\sum_{k=1}^{K} n_{g,k}\ \left|p_{g,k} - y_{g,k}\right|.$$

The integrated calibration index evaluates the same surgery-by-surgery gap against the LOESS curve f_g from (i),

$$\text{ICI}_g = \frac{1}{n_g}\sum_{i \in I_g}\left|f_g(p_i) - p_i\right|.$$

(iv) Hosmer–Lemeshow goodness-of-fit. Within each subgroup, surgeries are ordered by predicted risk and partitioned into $g = 10$ groups of approximately equal size, with cut points snapped to tied predicted probabilities so that identical predictions always share a group. Writing O and E for the observed and expected event counts in a group, the statistic is the usual sum of $\frac{(O-E)^2}{E\left(1-\frac{E}{n}\right)}$ over groups. Because model parameters were not estimated on the data under test — the audited scores are out-of-

fold predictions from folds excluding each surgery — no degrees of freedom are spent recovering an intercept and slope on the evaluation sample, and the statistic is referred to a χ^2 distribution on g degrees of freedom rather than the standard $g - 2$ convention utilized in [38]. The statistic is sample-size dependent and is therefore not comparable across strata of different size: at fixed lack of fit it grows with n. It is interpreted descriptively and jointly with ECE and ICI, which are reported alongside it (Supplementary Table 8), and with the per-group table of observed versus expected events that [38] requires to be read with it (Supplementary Table 9). Strata estimated under empirical-Bayes shrinkage, and strata containing a single outcome class, do not carry an interpretable statistic.

Clinical-utility metrics

Decision curve analysis [4] was performed per subgroup across the threshold range $t \in [0.02,0.40]$, the plausible band for intensified postoperative monitoring. The net benefit of a "flag and monitor when $p_i \geq t$" policy is true positives gained minus false positives weighted by the harm ratio $t/(1-t)$ implied by the threshold itself,

$$\mathrm{NB}_g(t) = \frac{\mathrm{TP}_g(t)}{n_g} - \frac{\mathrm{FP}_g(t)}{n_g} \cdot \frac{t}{1-t}.$$

Because $\mathrm{NB}_g(t)$ scales with the subgroup prevalence π_g^+, the standardized net benefit [5]

$$\mathrm{sNB}_g(t) = \frac{\mathrm{NB}_g(t)}{\pi_g^+}$$

is also reported per subgroup, normalized so that $\mathrm{sNB}_g(t) = 1$ corresponds to a hypothetical perfect model on I_g and $\mathrm{sNB}_g(t) \leq 0$ to no benefit over the "treat-none" default. The primary clinical-utility gap statistic is the largest absolute net-benefit difference between any two subgroups, taken across the threshold band,

$$\Delta\mathrm{NB}_{\max} = \max_{t \in [0.02,\, 0.40]} \; \max_{g,g' \in G} \left| \mathrm{NB}_g(t) - \mathrm{NB}_{g'}(t) \right|,$$

with 95% percentile confidence intervals from 2,000 patient-level bootstrap resamples, and only for strata meeting the prespecified 50-event floor for net-benefit inference. Net benefit is interpreted against both the treat-none and the treat-all (monitor-all) default strategies.

Model-based interaction tests

Gap statistics describe what a fixed threshold sees on the calibrated probability scale; they do not, on their own, say whether a covariate's effect on the outcome differs by sensitive attribute. To complement them, likelihood-ratio tests compare nested logistic regressions. For a covariate X, sensitive attribute A (sex), and adjustment vector C (age and ASA, as appropriate):

$$M_0\colon \operatorname{logit} \Pr(Y = 1) = \beta_0 + \beta_X X + \beta_A A + \beta_C^{\top} C,$$

$$M_1\colon \operatorname{logit} \Pr(Y = 1) = \beta_0 + \beta_X X + \beta_A A + \beta_{XA}(X{\cdot}A) \; + \beta_C^{\top} C.$$

Under $H_0: \beta_{XA} = 0$, the test statistic

$$\mathrm{LR} = 2(\ell(M_1) - \ell(M_0)) \sim \chi^2_{\mathrm{df}},$$

with df equal to the number of interaction parameters. Prespecified pairs were sex × physiological reserve and sex × ASA; a screen of sex × comorbidity interactions covered 23 comorbidities and was Benjamini–Hochberg–adjusted across that family. Each test is reported with its p-value and the Cox–Snell pseudo-R^2 increment ΔR^2, which guards against the common pattern of a small p but clinically negligible effect that a sufficiently powered cohort produces by default.

For the sex × comorbidity screen, the odds ratio for postoperative infection associated with each comorbidity was additionally estimated separately within each sex from its 2×2 table, with a Wald 95% confidence interval (Figure 3B; Supplementary Table 13). A within-sex odds ratio is an expected association and is not, by itself, evidence of a sex difference; the deployment-relevant comparison is the interaction term above.

Bias-mitigation methods

Both mitigations are score-level postprocessors: they act on the model's predicted probabilities and require no access to the training pipeline. To avoid evaluating mitigated probabilities in-sample, each was fitted by 5-fold stratified cross-validation (within every fold the transform was estimated on four folds (the calibration set I_g^{cal}) and applied to the held-out fold) so that every surgery receives an out-of-fold mitigated prediction and the before-versus-after comparison is computed on the full cohort. Isotonic recalibration was evaluated as a three-rung ladder: no recalibration, common cohort-wide recalibration, and group-specific recalibration by physiological reserve. Because isotonic regression can overfit sparse strata, post-recalibration calibration slope and subgroup event counts were inspected at each rung. Wasserstein-barycenter postprocessing was applied to the procedure-cluster axis alone, where distributional shift was most pronounced, and is reported as a parity-focused comparator rather than a calibration-preserving mitigation.

Group-specific isotonic recalibration. For each subgroup g, a monotone non-decreasing step function $\iota_g: [0,1] \to [0,1]$ is fitted on the calibration set by minimizing the squared distance to the outcomes,

$$\iota_g = \arg \min_{\iota \text{ monotone}} \sum_{i \in I_g^{\mathrm{cal}}} (\iota(p_i) - Y_i)^2.$$

At test time the recalibrated probability for a surgery in g is $p_i^{\mathrm{iso}} = \iota_g(p_i)$. Because the map is monotone, the within-subgroup rank order, and hence within-subgroup AUROC, is preserved except for tie-induced rank collapses at very low event counts.

Wasserstein-barycenter score postprocessing. Let F_g denote the empirical CDF of $\{p_i : i \in I_g^{\mathrm{cal}}\}$. The one-dimensional Wasserstein barycenter, weighted by the group sizes π_g, is the distribution whose quantile function is the weighted average of the per-group quantile functions,

$$Q_{\mathrm{bary}}(u) = \sum_{g \in G} \pi_g \; F_g^{-1}(u), u \in [0,1].$$

A prediction p from a surgery in g is mapped to $p^{\mathrm{W}} = Q_{\mathrm{bary}}(F_g(p))$ and re-expressed as the value at its own within-group percentile, read off the common barycenter scale. The map is monotone within each subgroup, so within-group rank order and AUROC are preserved exactly; on the calibration set all per-subgroup score distributions coincide after the map, enforcing demographic parity at the score level. By the impossibility of jointly satisfying calibration and equalized odds when group base rates differ [6], score postprocessing was restricted to demographic parity, with equalized-odds and equal-opportunity gaps used only as evaluation diagnostics. Both outputs remain probabilities in $[0,1]$ and so stay usable by the calibration and decision-curve analyses [7–9], in contrast to hard-threshold schemes whose output is a decision rather than a probability [10,11].

No-SMOTE reproduction and refit. To test whether class rebalancing accounted for the observed calibration pattern, the source modelling pipeline was reproduced and refit without SMOTE. The reproduction arm was required to recover all 10,719 published out-of-fold predictions bit-identically before the no-SMOTE comparison was interpreted. Arms are crossed over training window (2014–2022 as published, and the corrected 2018–2022 ascertainment window) and class rebalancing (SMOTE within training folds, and none), and all four arms are evaluated on the 2018–2022 window so that calibration-in-the-large, slope, ECE, ICI, Brier score and net benefit at t = 0.10 are comparable across arms (Supplementary Table 2). Both the overall calibration shift and the change in the frail–pre-frail calibration-intercept difference are reported.

Temporal holdout

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

Uncertainty quantification and multiple comparisons

Bootstrap confidence intervals. All confidence intervals are 95% percentile intervals from a patient-level bootstrap (fixed seed), with $B = 2000$ resamples drawn by resampling patients within each subgroup stratum, allowing subgroup event counts to vary across replicates so that intervals reflect sampling variability in observed prevalence. The overall cohort is resampled once per replicate, so a single overall interval applies across all subgroup axes. Pairwise gap statistics reuse the same resampled distribution as their underlying within-group rates, so that the dependence between a gap and those rates is preserved.

Estimability rules and event floors. A confidence interval for a subgroup metric is reported only where the subgroup carries at least the prespecified minimum number of events: 25 for AUROC (and for AUPRC, sensitivity, specificity and predictive values), 25 for the calibration intercept (and the Brier score), 100 for the calibration slope, and 50 for net benefit; below these floors the point estimate is shown descriptively without an interval, and is labelled not estimable where the table reports intervals for that metric. Separately, a cell in which more than 10% of bootstrap replicates return a non-estimable value (for example, a replicate containing no events) is reported as not estimable rather than with an interval. The two rules are applied as a conservative union: a cell below its floor is suppressed even when every replicate estimated cleanly, and a gap statistic inherits the floor failure of either contributing

stratum. The rule was applied without exception, including where it suppressed our own estimates. Per-cell replicate accounting is given in Supplementary Table 10.

Resampling scheme. An earlier scheme resampled cases and non-cases separately within each subgroup, fixing the event count in every replicate and estimating uncertainty conditional on the observed subgroup prevalence. It cannot produce a zero-event replicate at all and understates interval width in sparse strata. It is retained in the software for reproducibility of the originally submitted analysis; point estimates are unchanged under the revised estimator and only the intervals move, outward. The comparison is given in Supplementary Table 11.

Paired bootstrap. Between-group differences of overall metrics, e.g. $\Delta\text{AUROC} = \text{AUROC}_g - \text{AUROC}_{g'}$ uses a paired bootstrap that resamples surgeries from the pooled cohort and re-stratifies by group, preserving the dependence between the two subgroup estimates. Letting Δ_b denote the difference computed on resample b, the two-sided percentile p-value with continuity correction is

$$p = \frac{1}{B+1}(1 + 2\ \min(\#\{b\colon \Delta_b \le 0\},\ \#\{b\colon \Delta_b \ge 0\})),$$

reported alongside the symmetric percentile interval $\left[\Delta_{\alpha/2},\ \Delta_{1-\alpha/2}\right]$ at $\alpha = 0.05$.

Empirical-Bayes shrinkage. For the intersectional analysis, two-way cells with fewer than 50 outcome events have their cell rate r_c shrunk toward the marginal subgroup rate r_0,

$$\hat{r}_c = (1 - w_c)\, r_c + w_c\, r_0, w_c = \frac{\tau}{n_c^+ + \tau},$$

where n_c^+ is the cell event count and τ is the shrinkage strength estimated by the method of moments from cross-cell variance. The pooling weight w_c (0 = no shrinkage, 1 = full shrinkage to the margin) is reported per cell (Supplementary Table 12) so that the degree of pooling is explicit.

Benjamini–Hochberg FDR. Multiple-comparison correction controls the false discovery rate at $q = 0.05$, applied independently within each of four test families: (a) the gap-statistical significance tests within each axis (discrimination, calibration, clinical utility); (b) the sex × comorbidity interaction screen; (c) the sex-stratified scan of cohort characterization variables (Table 1, Supplementary Table 1); and (d) the feature-by-sex effect-size scan (Figure 3A). Within a family of m tests, the ordered p-values $p_{(1)} \le \cdots \le p_{(m)}$ are compared with the linear thresholds $(k/m)\, q$; the largest $k^\star$ satisfying $p_{(k)} \le (k/m)\, q$ is identified, and all hypotheses with $p \le p_{(k^\star)}$ are rejected. Within-subgroup point estimates are reported without correction.

Effect sizes

The cohort characterization table and the per-subgroup gap annotations report an effect size matched to each comparison:

- **Cohen's** d for a continuous variable compared across two groups: $d = (\bar{x}_1 - \bar{x}_2)/s_p$, with s_p the pooled standard deviation.
- **Cohen's** h for a binary variable or proportion: $h = 2\left(\arcsin\sqrt{p_1} - \arcsin\sqrt{p_2}\right)$.

- **Cramér's** V for an unordered categorical variable in an $r \times c$ contingency table: $V = \sqrt{\chi^2/(n \cdot \min(r-1, c-1))}$.
- η^2 for a continuous variable stratified by an attribute with more than two levels: $\eta^2 = \mathrm{SS}_{\mathrm{between}}/\mathrm{SS}_{\mathrm{total}}$.

Each effect size is accompanied by its conventional magnitude class (negligible / small / medium / large). Where a comparison is composite (the equalized-odds gap) or operates on a non-rate metric (calibration intercept or slope, decision-curve net benefit), no single effect size applies, and the column is left blank rather than filled with a misleading number.

Outcome definition: ICD-10-GM code list

Bacterial postoperative infection within the index admission — the continuous inpatient stay containing the surgery, from admission to discharge — was defined by the presence of any code within the following 68 ICD-10-GM categories (280 codes), following the coding pipeline established for the source surgical cohort:

A02, A03, A04, A23, A26, A31, A32, A36–A42, A46, A48, A49, A54–A56, B95, B96, B98, G00, G01, G04, G05, H44, H60, H62, H66, I32, I41, J02, J03, J13–J17, J20, J86, K12, K35, K63, K65, K67, L00, L01, L03, L08, L51, M00, M01, M46, M63, M68, M72, M73, M86, N13, N49, N74, N76, U69, U80, U81, U82.

Present-on-admission (POA) flags were used to exclude infections coded as already present at admission. Where the POA flag was absent, postoperative timing could not be established and the infection case was excluded at cohort construction rather than retained as a negative outcome; uninfected patients entered regardless, so the cohort is selected on the outcome. POA coding is effectively absent from this extract before 2018: the 2014–2017 years contribute 4,668 procedures and 3 coded events, against 6,051 procedures and 501 coded events (8.28%) in 2018–2022. All results therefore describe performance against the POA-coded endpoint rather than true infection incidence, and prevalence-dependent results are additionally reported for the ascertainable 2018–2022 window (Supplementary Figure 2; Supplementary Tables 14–18). The model treats infection as binary, and this audit follows the same convention.

SUPPORT2 cohort

As a software reproducibility example, the identical audit was applied to the public SUPPORT2 cohort (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment; UCI Machine Learning Repository [12]), comprising 9,105 critically ill adults from five US medical centers. On this cohort the audited model is a logistic-regression classifier for in-hospital mortality, fitted on a training partition and audited on a held-out test partition ($n = 1{,}821$; mortality 25.9%). All metric definitions apply unchanged. The SUPPORT2 audit additionally exercises a race/ethnicity axis (sex and age stratum being shared with the primary audit), which the Swiss EHR does not record, and was run end-to-end from the example script shipped with the *isitfair* package under a fixed random seed. Group-specific isotonic recalibration was applied to the race/ethnicity axis. This secondary cohort is reported only as a software reproducibility example, not as external validation of the infection model, and is not pooled with the analysis from our cohort.

References

1 Blatter TU, Wintsch Y, Triep K, *et al.* End-of-surgery prediction of postoperative infectious complications from intraoperative vital-sign dynamics. *npj Digit Med*. Published Online First: 7 May 2026. doi: 10.1038/s41746-026-02707-1

2 Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*. 2010;21:128–38. doi: 10.1097/EDE.0b013e3181c30fb2

3 Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited*: *Critical Care Medicine*. 2007;35:2052–6. doi: 10.1097/01.CCM.0000275267.64078.B0

4 Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making*. 2006;26:565–74. doi: 10.1177/0272989X06295361

5 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. doi: 10.1093/jamiaopen/ooaf115

6 Pleiss G, Raghavan M, Wu F, *et al.* On Fairness and Calibration. *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 2017.

7 Chzhen E, Denis C, Hebiri M, *et al.* Fair Regression with Wasserstein Barycenters.

8 Jiang R, Pacchiano A, Stepleton T, *et al.* Wasserstein Fair Classification. Published Online First: 2019. doi: 10.48550/ARXIV.1907.12059

9 del Barrio E, Gamboa F, Gordaliza P, *et al.* Obtaining fairness using optimal transport theory. 2018.

10 Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. 2016.

11 Corbett-Davies S, Pierson E, Feller A, *et al.* Algorithmic Decision Making and the Cost of Fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM 2017:797–806.

12 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.

Subgroup Audit Reports

Perioperative infection prediction -- subgroup deployment-readiness audit (audited cohort (2014–2022))

Model	Random forest (M2, end of surgery), published out-of-fold predictions
Cohort	Inselspital Bern -- audited cohort (2014–2022): n=10,719; 504 events
Outcome	Present-on-Admission-coded postoperative bacterial infection
Validation	5-fold stratified CV with SMOTE inside training folds (OOF predictions)
Threshold	0.1 (reference operating anchor, not an elicited optimum)
Uncertainty	2,000 patient-level bootstrap resamples; pre-registered event floors applied (AUROC 25, calibration intercept 25, calibration slope 100, net benefit 50)
Mitigation	Not shown here. The mitigation analysis is the cross-fitted recalibration ladder reported in Table 4 and Supplementary Figure 1; a split-based before/after would be a different analysis.
Window	all years (2014-2022)

Audit summary

N: 10719	**Events:** 504	**Prevalence:** 0.047	**Threshold:** 0.100
Bootstrap resamples: 2000	**Resampling:** subgroup-stratified patient-level	**Attributes:** 9	

Cohort characterisation (Table 1)

6 of 8 variables differ significantly across 'sex' at FDR q<0.05 (Benjamini–Hochberg).

		Grouped by sex				
		Overall	Female	Male	P-Value	q_fdr
n		10719	3689	7030		
age, median [Q1,Q3]		66.0 [55.0,74.0]	66.0 [53.0,75.0]	66.0 [56.0,73.0]	0.467	0.467
age_group, n (%)	0-17	367 (3.4)	175 (4.7)	192 (2.7)	<0.001	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, n (%)	1	1227 (11.4)	630 (17.1)	597 (8.5)	<0.001	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)		
frailty, n (%)	frail	6139 (57.3)	1666 (45.2)	4473 (63.6)	<0.001	0.0016
	non_frail	1022 (9.5)	522 (14.2)	500 (7.1)		
	pre_frail	3558 (33.2)	1501 (40.7)	2057 (29.3)		
emergencysurgery, n (%)	0	9250 (86.3)	3221 (87.3)	6029 (85.8)	0.028	0.0373
	1	1469 (13.7)	468 (12.7)	1001 (14.2)		
duration_of_surg, median [Q1,Q3]		213.0 [147.0,281.0]	193.0 [135.0,267.0]	221.0 [156.0,285.0]	<0.001	0.0016
surgical_group, n (%)	chopcluster_1	903 (8.4)	365 (9.9)	538 (7.7)	<0.001	0.0016
	chopcluster_10	346 (3.2)	98 (2.7)	248 (3.5)		
	chopcluster_11	303 (2.8)	303 (8.2)	0 (0.0)		
	chopcluster_12	1623 (15.1)	711 (19.3)	912 (13.0)		
	chopcluster_13	213 (2.0)	95 (2.6)	118 (1.7)		
	chopcluster_3	807 (7.5)	317 (8.6)	490 (7.0)		
	chopcluster_4	3764 (35.1)	836 (22.7)	2928 (41.7)		
	chopcluster_5	769 (7.2)	189 (5.1)	580 (8.3)		
	chopcluster_6	530 (4.9)	249 (6.7)	281 (4.0)		

		Grouped by sex				
		Overall	Female	Male	P-Value	q_fdr
	chopcluster_7	298 (2.8)	105 (2.8)	193 (2.7)		
	chopcluster_8	306 (2.9)	95 (2.6)	211 (3.0)		
	none	857 (8.0)	326 (8.8)	531 (7.6)		
	0	10215 (95.3)	3499 (94.8)	6716 (95.5)	0.123	0.141
outcome, n (%)	1	504 (4.7)	190 (5.2)	314 (4.5)		

Per-variable effect sizes and BH-FDR

Variable	Kind	n	p-value	q (FDR)	Effect size	Kind	Magnitude
duration_of_surg	continuous	10719	0.001	0.0016	-0.161	d	negligible
frailty	categorical	10719	0.001	0.0016	+0.185	v	small
age_group	categorical	10719	0.001	0.0016	+0.116	v	small
asa	categorical	10719	0.001	0.0016	+0.191	v	small
surgical_group	categorical	10719	0.001	0.0016	+0.306	v	medium
emergencysurgery	binary	10719	0.028	0.0373	-0.046	h	negligible
outcome	binary	10719	0.123	0.1406	+0.032	h	negligible
age	continuous	10719	0.467	0.4670	-0.079	d	negligible

Discrimination

Attribute	Subgroup	n	Events	Prevalence	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
sex	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
sex	female	3689	190	0.052	0.878 (0.849–0.903)	0.408 (0.341–0.478)	0.900 (0.856–0.940)	0.640 (0.624–0.656)
sex	male	7030	314	0.045	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	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
age_group	0-17	367	21	0.057	0.807	0.246	0.857	0.462
age_group	18-44	941	34	0.036	0.854 (0.792–0.908)	0.329 (0.175–0.505)	0.824 (0.679–0.939)	0.682 (0.653–0.712)
age_group	45-64	3715	147	0.040	0.888 (0.856–0.916)	0.382 (0.303–0.468)	0.918 (0.872–0.959)	0.666 (0.650–0.682)
age_group	65-79	4684	249	0.053	0.889 (0.865–0.910)	0.438 (0.371–0.509)	0.912 (0.875–0.944)	0.654 (0.639–0.668)
age_group	80+	1012	53	0.052	0.805 (0.744–0.863)	0.227 (0.145–0.344)	0.849 (0.746–0.940)	0.577 (0.545–0.607)
frailty	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
frailty	frail	6139	400	0.065	0.878 (0.860–0.896)	0.428 (0.379–0.481)	0.927 (0.901–0.951)	0.584 (0.571–0.597)
frailty	non_frail	1022	6	0.006	0.868	0.084	0.833	0.816
frailty	pre_frail	3558	98	0.028	0.821 (0.774–0.865)	0.190 (0.127–0.274)	0.786 (0.701–0.863)	0.702 (0.687–0.716)
clinic	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
clinic	ORL	233	15	0.064	0.820	0.221	1.000	0.408
clinic	gynecology	452	24	0.053	0.821	0.395	0.750	0.689
clinic	neuro	1117	26	0.023	0.871 (0.785–0.939)	0.434 (0.240–0.621)	0.808 (0.643–0.955)	0.756 (0.731–0.780)
clinic	ortho	1690	76	0.045	0.852 (0.808–0.887)	0.232 (0.160–0.332)	0.908 (0.835–0.965)	0.633 (0.610–0.657)
clinic	pediatric	65	3	0.046	0.653	0.181	0.667	0.452
clinic	thoracic	723	14	0.019	0.945	0.411	0.929	0.781
clinic	urology	349	19	0.054	0.785	0.331	0.737	0.624
clinic	vascular	4852	184	0.038	0.866 (0.834–0.896)	0.323 (0.257–0.395)	0.886 (0.837–0.929)	0.656 (0.643–0.670)
clinic	visceral	1238	143	0.116	0.883 (0.854–0.909)	0.557 (0.472–0.640)	0.965 (0.932–0.993)	0.482 (0.453–0.510)
surgical_group	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
surgical_group	chopcluster_1	903	22	0.024	0.879	0.460	0.818	0.751
surgical_group	chopcluster_10	346	28	0.081	0.789 (0.673–0.889)	0.383 (0.233–0.594)	0.786 (0.625–0.933)	0.560 (0.505–0.614)
surgical_group	chopcluster_11	303	9	0.030	0.761	0.239	0.556	0.769
surgical_group	chopcluster_12	1623	62	0.038	0.856 (0.810–0.894)	0.192 (0.135–0.294)	0.919 (0.844–0.981)	0.654 (0.631–0.676)
surgical_group	chopcluster_13	213	43	0.202	0.796 (0.716–0.871)	0.561 (0.419–0.700)	0.930 (0.846–1.000)	0.253 (0.192–0.318)
surgical_group	chopcluster_3	807	22	0.027	0.941	0.482	0.955	0.753
surgical_group	chopcluster_4	3764	118	0.031	0.841 (0.793–0.883)	0.251 (0.188–0.338)	0.839 (0.767–0.902)	0.673 (0.658–0.687)

Attribute	Subgroup	n	Events	Prevalence	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
surgical_group	chopcluster_5	769	30	0.039	0.893 (0.827–0.943)	0.392 (0.225–0.565)	0.933 (0.823–1.000)	0.636 (0.602–0.672)
surgical_group	chopcluster_6	530	67	0.126	0.889 (0.844–0.927)	0.625 (0.516–0.728)	0.955 (0.901–1.000)	0.471 (0.425–0.514)
surgical_group	chopcluster_7	298	36	0.121	0.864 (0.813–0.913)	0.521 (0.365–0.681)	1.000 (1.000–1.000)	0.309 (0.256–0.366)
surgical_group	chopcluster_8	306	23	0.075	0.802	0.327	0.870	0.551
surgical_group	none	857	44	0.051	0.900 (0.866–0.932)	0.388 (0.256–0.531)	0.977 (0.927–1.000)	0.625 (0.590–0.658)
emergency	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
emergency	elective	9250	316	0.034	0.859 (0.837–0.879)	0.252 (0.209–0.303)	0.845 (0.808–0.883)	0.701 (0.692–0.711)
emergency	emergency	1469	188	0.128	0.858 (0.827–0.888)	0.554 (0.480–0.629)	0.989 (0.973–1.000)	0.268 (0.244–0.292)
sex × age_group	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
sex × age_group	*female / 0-17*	*175*	*13*	*0.063*	*0.902*	*0.408*	*0.907*	*0.549*
sex × age_group	*female / 18-44*	*408*	*13*	*0.038*	*0.873*	*0.340*	*0.896*	*0.663*
sex × age_group	female / 45-64	1179	52	0.044	0.860 (0.788–0.919)	0.377 (0.248–0.517)	0.846 (0.738–0.941)	0.674 (0.647–0.702)
sex × age_group	female / 65-79	1470	93	0.063	0.900 (0.865–0.931)	0.545 (0.451–0.640)	0.935 (0.882–0.980)	0.645 (0.620–0.670)
sex × age_group	*female / 80+*	*457*	*19*	*0.044*	*0.814*	*0.166*	*0.894*	*0.596*
sex × age_group	*male / 0-17*	*192*	*8*	*0.043*	*0.648*	*0.119*	*0.886*	*0.560*
sex × age_group	*male / 18-44*	*533*	*21*	*0.041*	*0.844*	*0.329*	*0.889*	*0.679*
sex × age_group	male / 45-64	2536	95	0.037	0.903 (0.873–0.929)	0.392 (0.301–0.494)	0.958 (0.915–0.991)	0.662 (0.644–0.681)
sex × age_group	male / 65-79	3214	156	0.049	0.882 (0.848–0.911)	0.378 (0.307–0.463)	0.897 (0.843–0.942)	0.658 (0.641–0.675)
sex × age_group	*male / 80+*	*555*	*34*	*0.057*	*0.798*	*0.269*	*0.891*	*0.595*
sex × frailty	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
sex × frailty	female / frail	1666	129	0.077	0.888 (0.855–0.918)	0.492 (0.408–0.577)	0.938 (0.892–0.977)	0.530 (0.505–0.556)
sex × frailty	*female / non_frail*	*522*	*1*	*0.016*	*0.996*	*0.333*	*0.900*	*0.748*
sex × frailty	female / pre_frail	1501	60	0.040	0.823 (0.761–0.877)	0.226 (0.148–0.345)	0.817 (0.710–0.907)	0.703 (0.679–0.726)
sex × frailty	male / frail	4473	271	0.061	0.874 (0.850–0.896)	0.395 (0.335–0.459)	0.923 (0.891–0.954)	0.604 (0.589–0.619)
sex × frailty	*male / non_frail*	*500*	*5*	*0.020*	*0.857*	*0.071*	*0.896*	*0.785*
sex × frailty	*male / pre_frail*	*2057*	*38*	*0.021*	*0.816*	*0.169*	*0.873*	*0.696*
age_group × frailty	Overall	10719	504	0.047	0.876 (0.859–0.892)	0.376 (0.332–0.422)	0.899 (0.871–0.925)	0.647 (0.638–0.656)
age_group × frailty	*0-17 / frail*	*211*	*19*	*0.077*	*0.791*	*0.282*	*0.862*	*0.393*
age_group × frailty	*0-17 / non_frail*	*45*	*0*	*0.043*	—	—	—	*0.711*
age_group × frailty	*0-17 / pre_frail*	*111*	*2*	*0.039*	*0.539*	*0.042*	*0.852*	*0.510*
age_group × frailty	*18-44 / frail*	*300*	*21*	*0.060*	*0.864*	*0.463*	*0.835*	*0.565*
age_group × frailty	*18-44 / non_frail*	*331*	*2*	*0.015*	*0.869*	*0.031*	*0.826*	*0.789*
age_group × frailty	*18-44 / pre_frail*	*310*	*11*	*0.036*	*0.744*	*0.141*	*0.809*	*0.680*

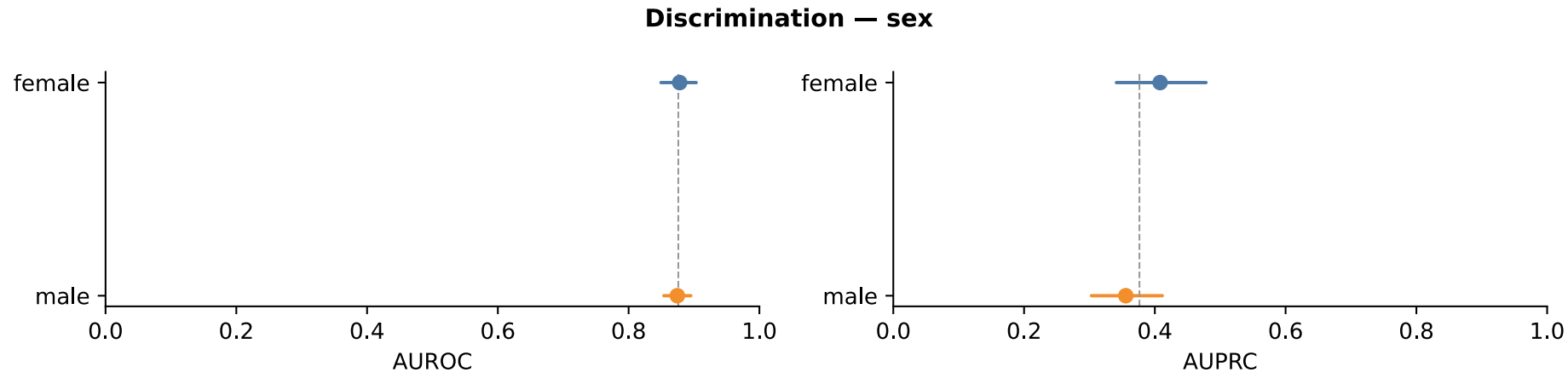
Attribute	Subgroup	n	Events	Prevalence	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
age_group × frailty	45-64 / frail	1967	111	0.056	0.898 (0.869–0.926)	0.444 (0.354–0.540)	0.964 (0.927–0.992)	0.603 (0.581–0.625)
age_group × frailty	*45-64 / non_frail*	*410*	*3*	*0.015*	*0.971*	*0.307*	*0.920*	*0.790*
age_group × frailty	*45-64 / pre_frail*	*1338*	*33*	*0.026*	*0.808*	*0.175*	*0.887*	*0.701*
age_group × frailty	65-79 / frail	3032	204	0.067	0.889 (0.862–0.914)	0.487 (0.416–0.560)	0.922 (0.884–0.958)	0.610 (0.592–0.627)
age_group × frailty	*65-79 / non_frail*	*218*	*1*	*0.023*	*0.592*	*0.010*	*0.905*	*0.742*
age_group × frailty	*65-79 / pre_frail*	*1434*	*44*	*0.033*	*0.870*	*0.247*	*0.905*	*0.714*
age_group × frailty	*80+ / frail*	*629*	*45*	*0.068*	*0.796*	*0.253*	*0.859*	*0.523*
age_group × frailty	*80+ / non_frail*	*18*	*0*	*0.046*	—	—	—	*0.667*
age_group × frailty	*80+ / pre_frail*	*365*	*8*	*0.030*	*0.740*	*0.209*	*0.836*	*0.652*

Gap statistics

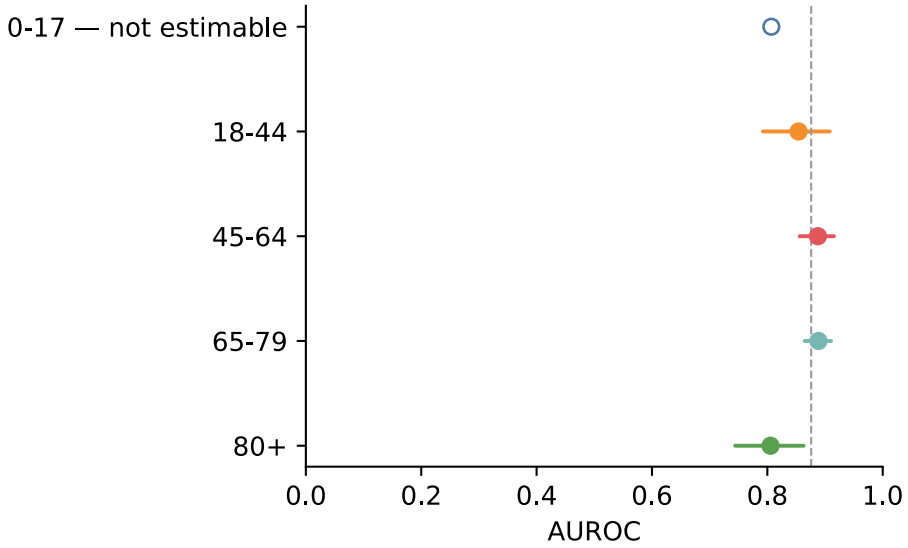
Effect-size column shows Cohen's *h* on rate-metric gaps for two-group attributes; multi-group attributes and ranking metrics (AUROC / AUPRC) show "—".

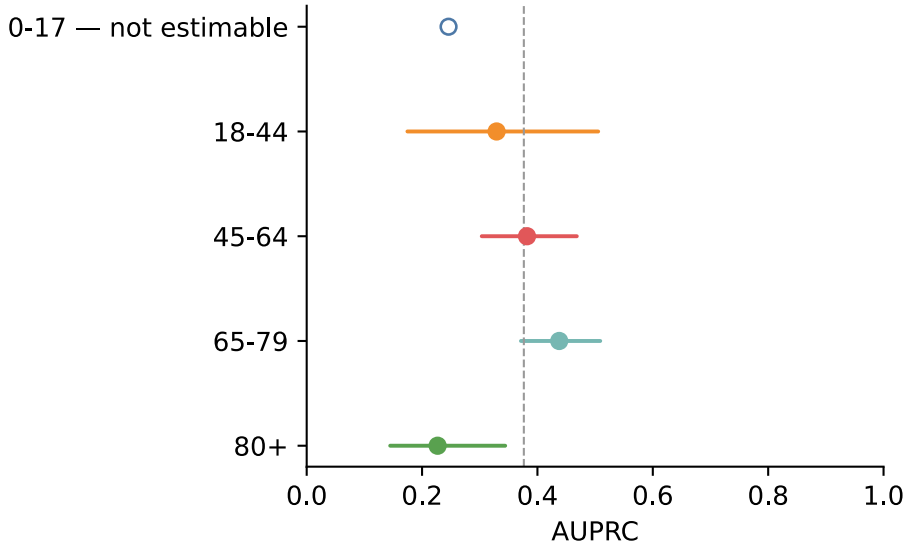
Attribute	Metric	Value	Effect size
sex	equal_opportunity_difference	0.0019	h=+0.01 (negligible)
sex	predictive_equality	0.0102	h=-0.02 (negligible)
sex	equalized_odds_gap	0.0102	—
sex	statistical_parity	0.0135	—
age_group	equal_opportunity_difference	0.0948	—
age_group	predictive_equality	0.2200	—
age_group	equalized_odds_gap	0.2200	—
age_group	statistical_parity	0.2200	—
frailty	equal_opportunity_difference	0.1418	—
frailty	predictive_equality	0.2320	—
frailty	equalized_odds_gap	0.2320	—
frailty	statistical_parity	0.2616	—
clinic	equal_opportunity_difference	0.3333	—
clinic	predictive_equality	0.3731	—
clinic	equalized_odds_gap	0.3731	—
clinic	statistical_parity	0.3857	—
surgical_group	equal_opportunity_difference	0.4444	—
surgical_group	predictive_equality	0.5158	—
surgical_group	equalized_odds_gap	0.5158	—

Attribute	Metric	Value	Effect size
surgical_group	statistical_parity	0.5431	—
emergency	equal_opportunity_difference	0.1444	h=-0.60 (medium)
emergency	predictive_equality	0.4335	h=+0.90 (large)
emergency	equalized_odds_gap	0.4335	—
emergency	statistical_parity	0.4477	—
sex × age_group	equal_opportunity_difference	0.1117	—
sex × age_group	predictive_equality	0.1300	—
sex × age_group	equalized_odds_gap	0.1300	—
sex × age_group	statistical_parity	0.1415	—
sex × frailty	equal_opportunity_difference	0.1213	—
sex × frailty	predictive_equality	0.2550	—
sex × frailty	equalized_odds_gap	0.2550	—
sex × frailty	statistical_parity	0.2844	—
age_group × frailty	equal_opportunity_difference	0.1546	—
age_group × frailty	predictive_equality	0.3970	—
age_group × frailty	equalized_odds_gap	0.3970	—
age_group × frailty	statistical_parity	0.4158	—

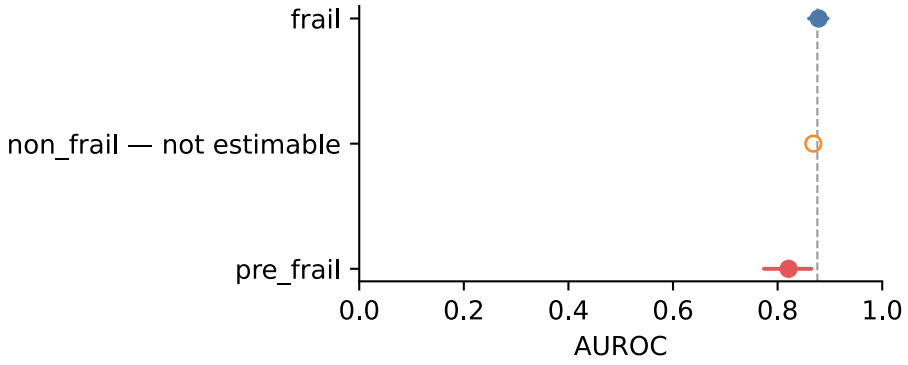


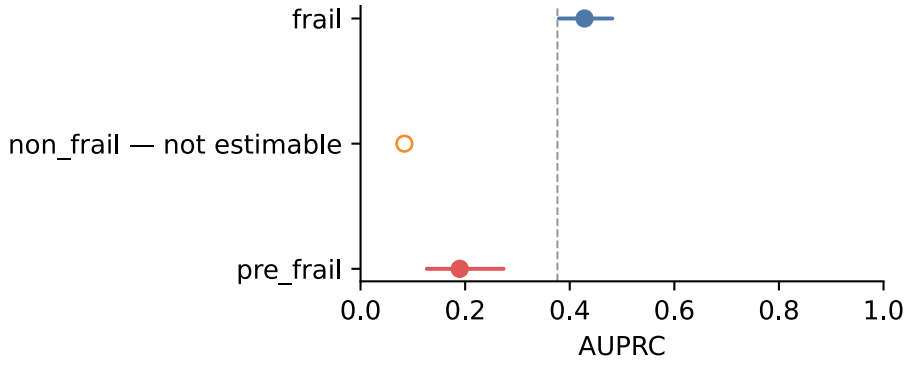
Discrimination — age_group

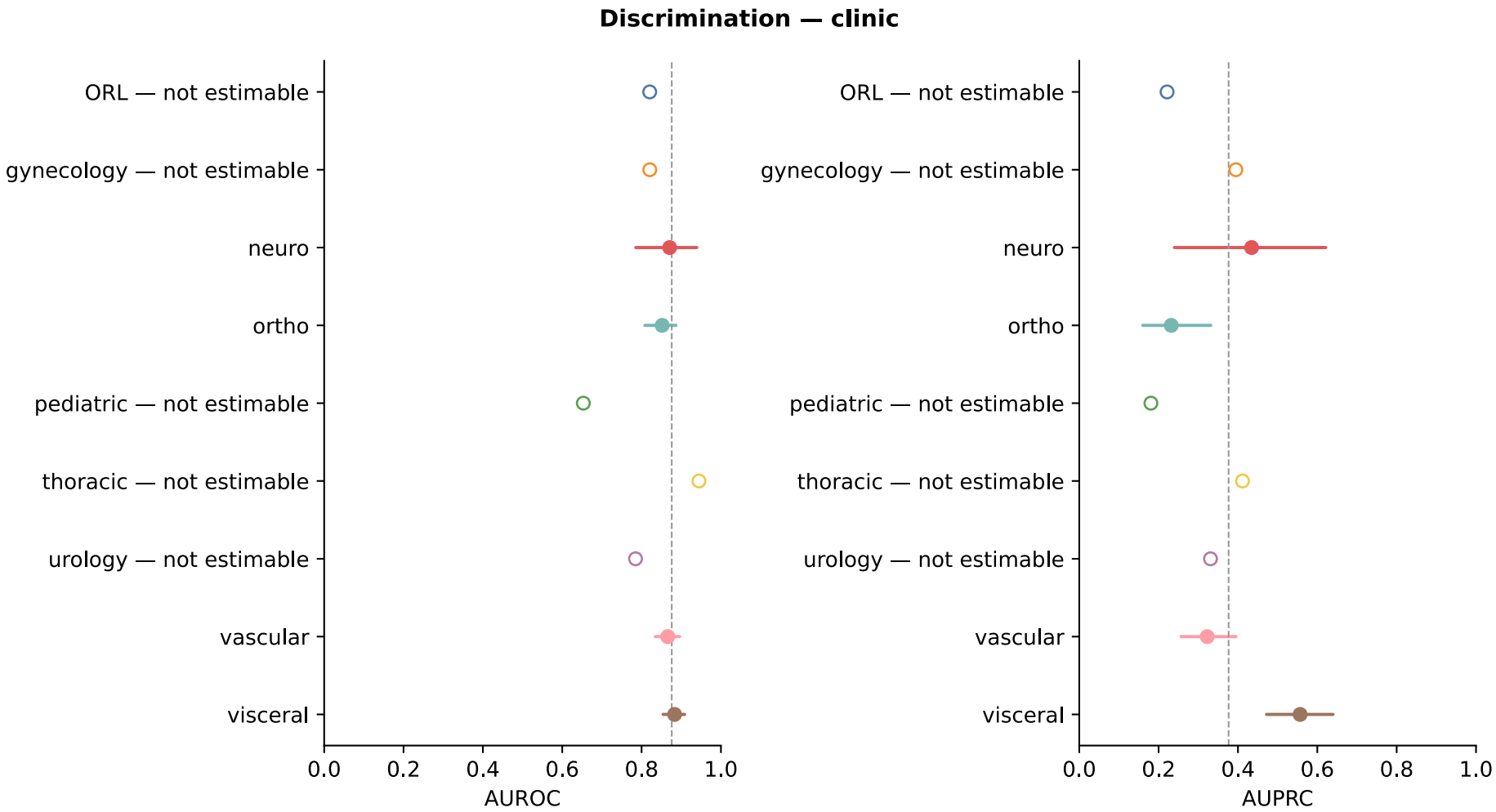


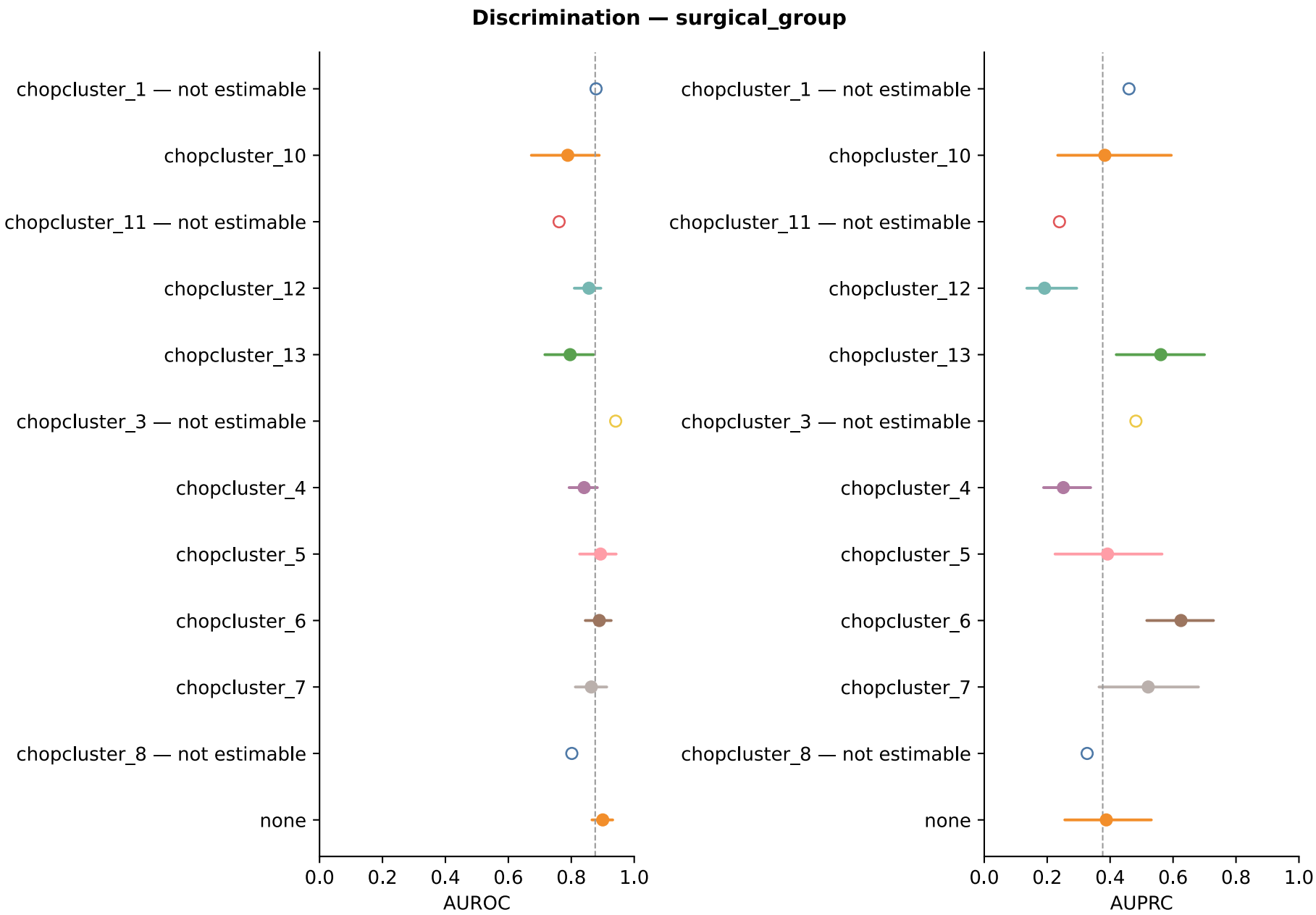


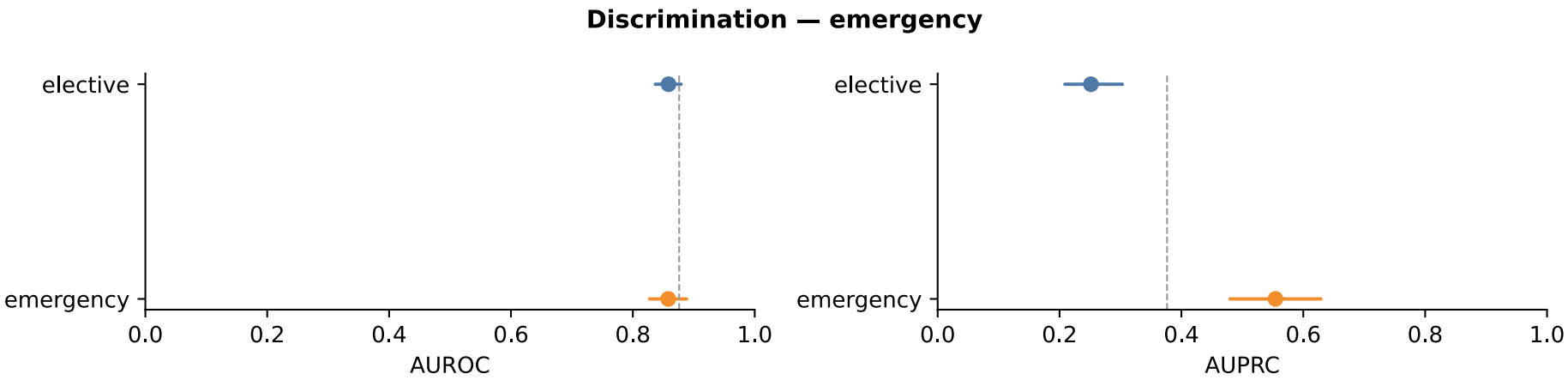
Discrimination — frailty

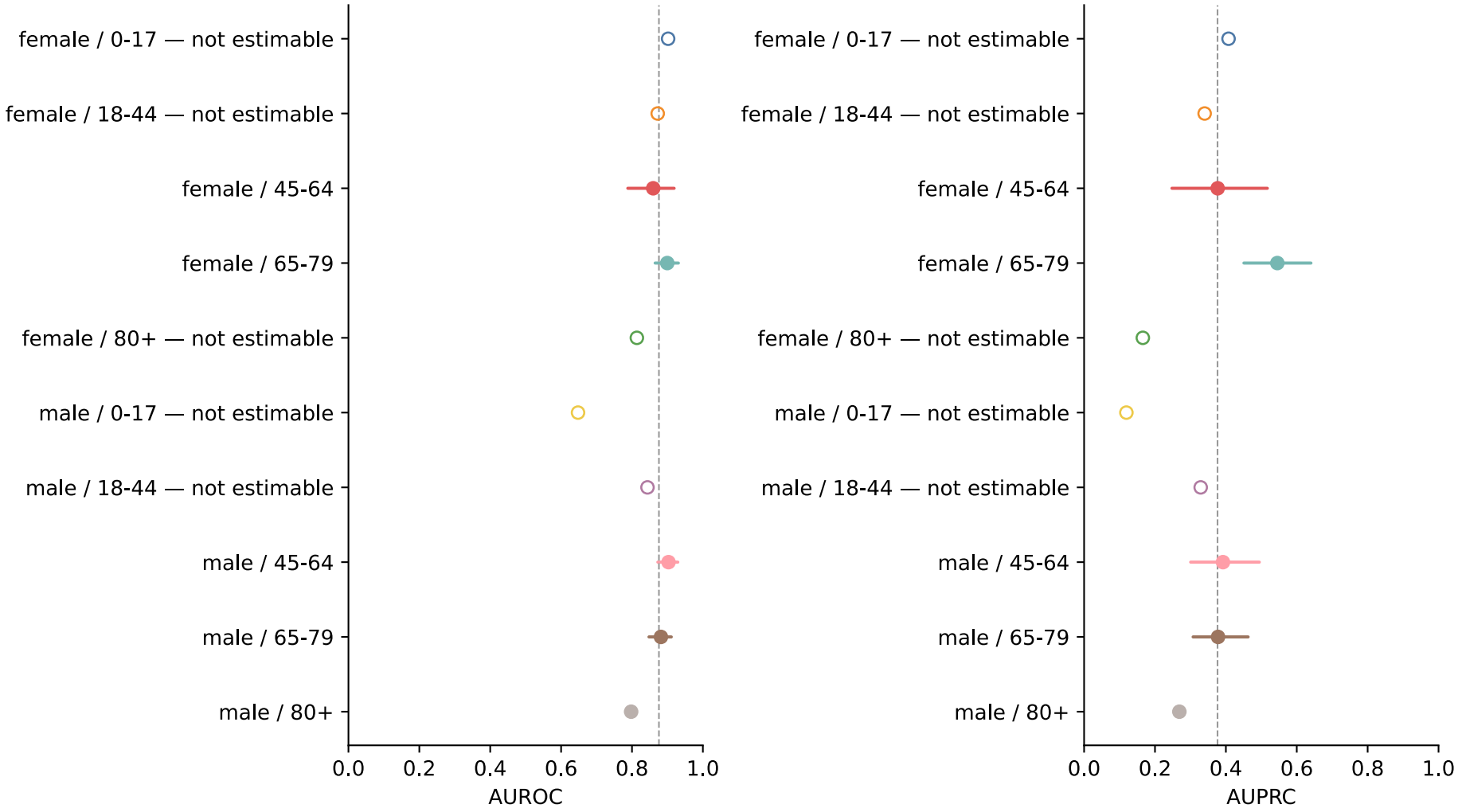


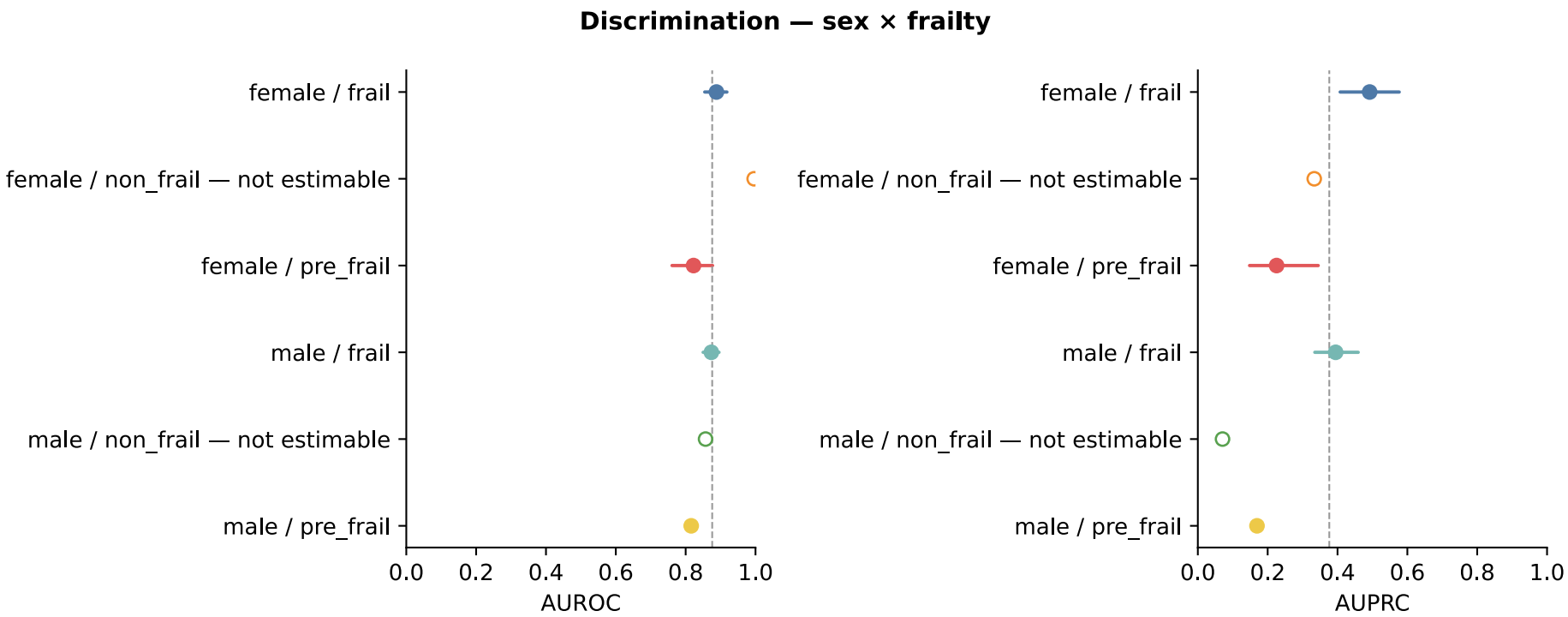


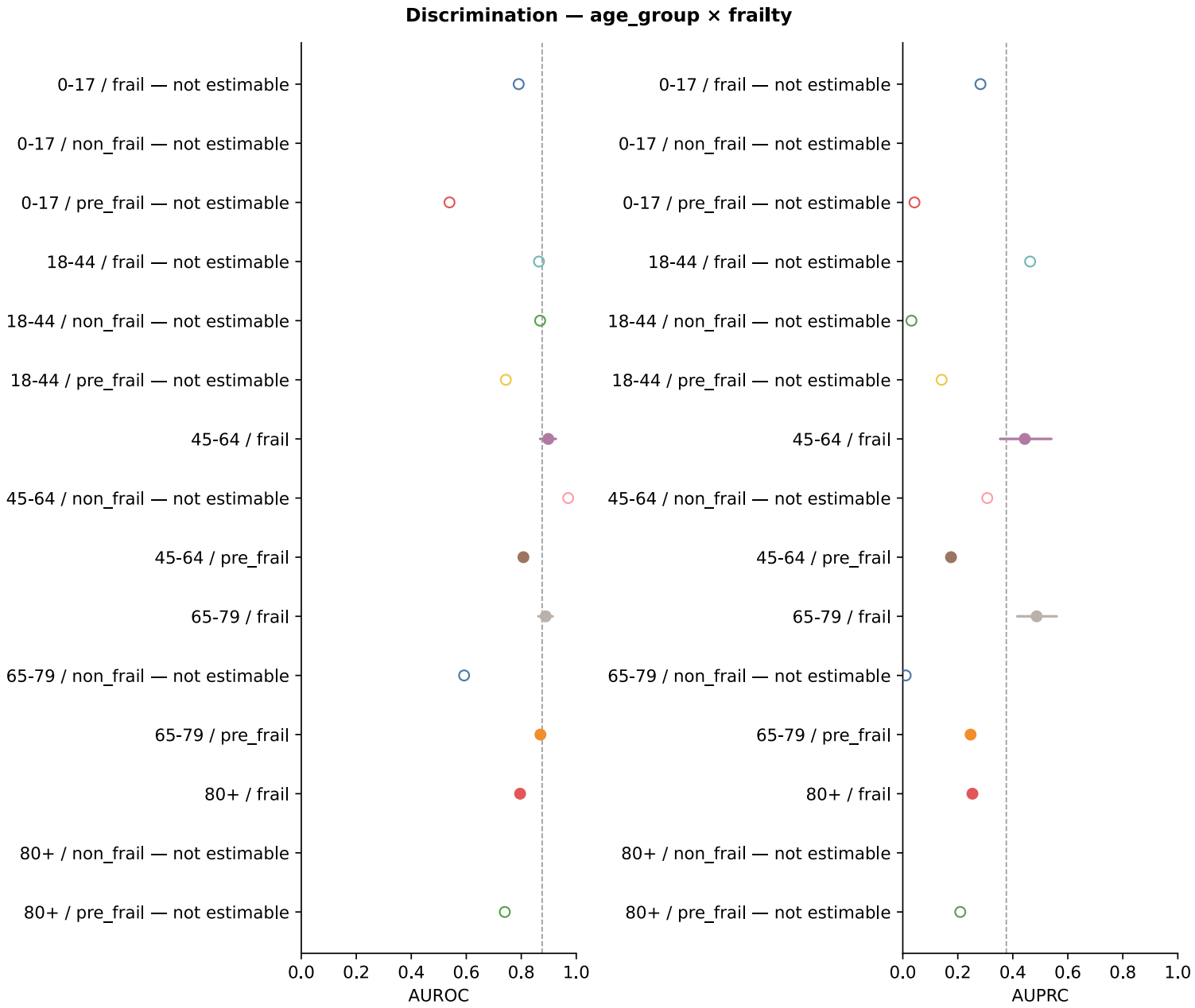

Discrimination — clinic
ORL — not estimable
gynecology — not estimable
neuro
ortho
pediatric — not estimable
thoracic — not estimable
urology — not estimable
vascular
visceral
0.0
0.2
0.4
0.6
0.8
1.0
AUROC
AUPRC


Discrimination — surgical_group
chopcluster_1 — not estimable
chopcluster_10
chopcluster_11 — not estimable
chopcluster_12
chopcluster_13
chopcluster_3 — not estimable
chopcluster_4
chopcluster_5
chopcluster_6
chopcluster_7
chopcluster_8 — not estimable
none
0.0
0.2
0.4
0.6
0.8
1.0
AUROC
AUPRC


Discrimination — emergency
elective
emergency
0.0
0.2
0.4
0.6
0.8
1.0
AUROC
elective
emergency
0.0
0.2
0.4
0.6
0.8
1.0
AUPRC

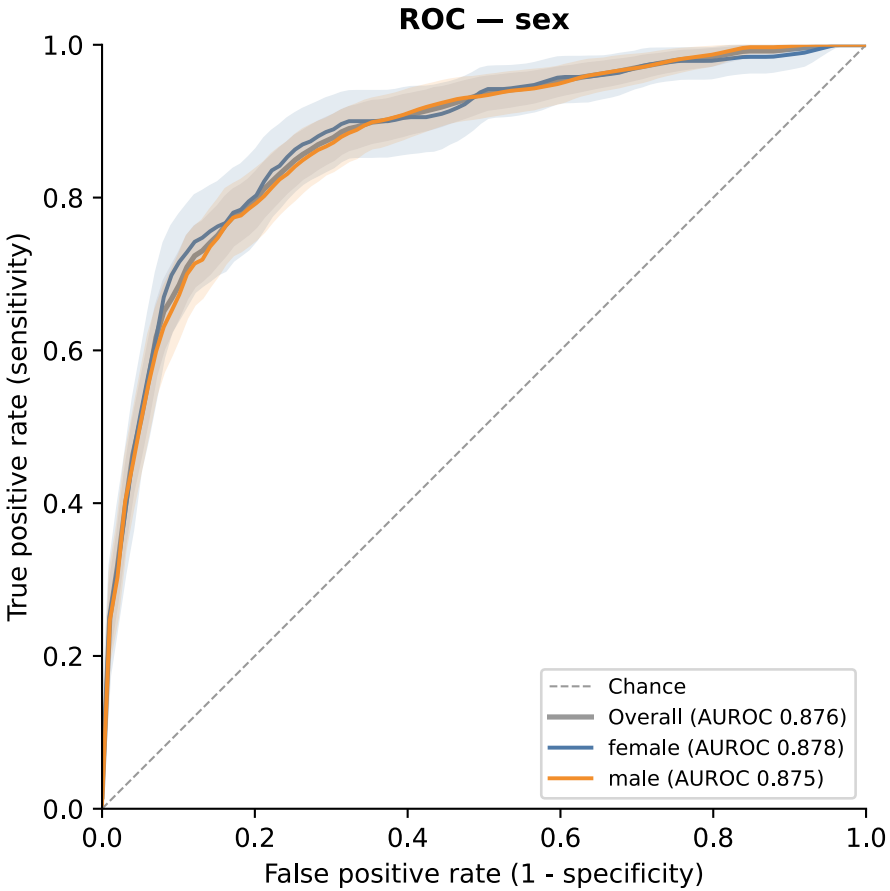

Discrimination — sex × age_group
female / 0-17 — not estimable
female / 18-44 — not estimable
female / 45-64
female / 65-79
female / 80+ — not estimable
male / 0-17 — not estimable
male / 18-44 — not estimable
male / 45-64
male / 65-79
male / 80+
0.0
0.2
0.4
0.6
0.8
1.0
AUROC
AUPRC

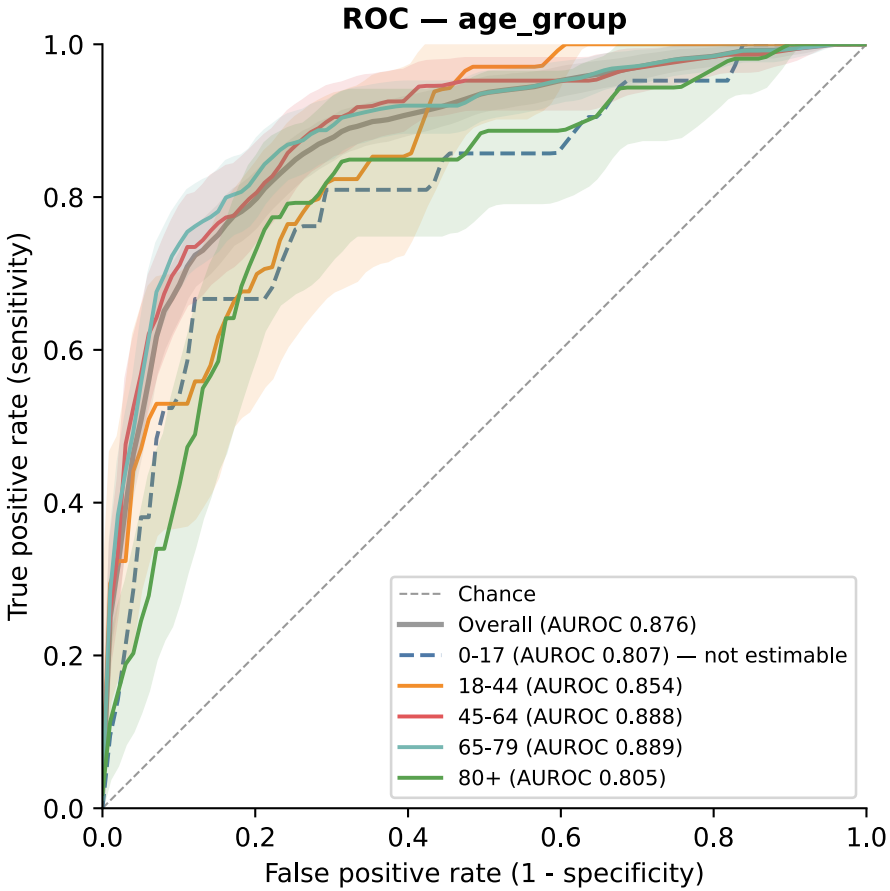

Discrimination — sex × frailty
female / frail
female / non_frail — not estimable
female / pre_frail
male / frail
male / non_frail — not estimable
male / pre_frail
0.0
0.2
0.4
0.6
0.8
1.0
AUROC
AUPRC

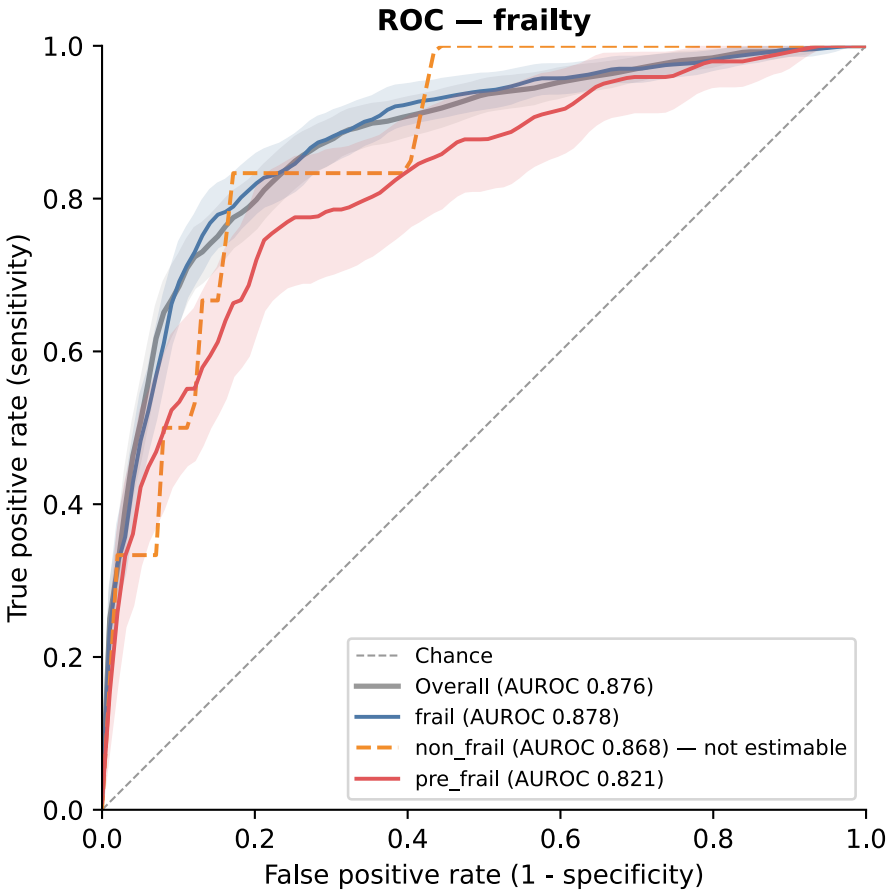

Discrimination — age_group × frailty
0-17 / frail — not estimable
0-17 / non_frail — not estimable
0-17 / pre_frail — not estimable
18-44 / frail — not estimable
18-44 / non_frail — not estimable
18-44 / pre_frail — not estimable
45-64 / frail
45-64 / non_frail — not estimable
45-64 / pre_frail
65-79 / frail
65-79 / non_frail — not estimable
65-79 / pre_frail
80+ / frail
80+ / non_frail — not estimable
80+ / pre_frail — not estimable
0.0 0.2 0.4 0.6 0.8 1.0
AUROC
AUPRC

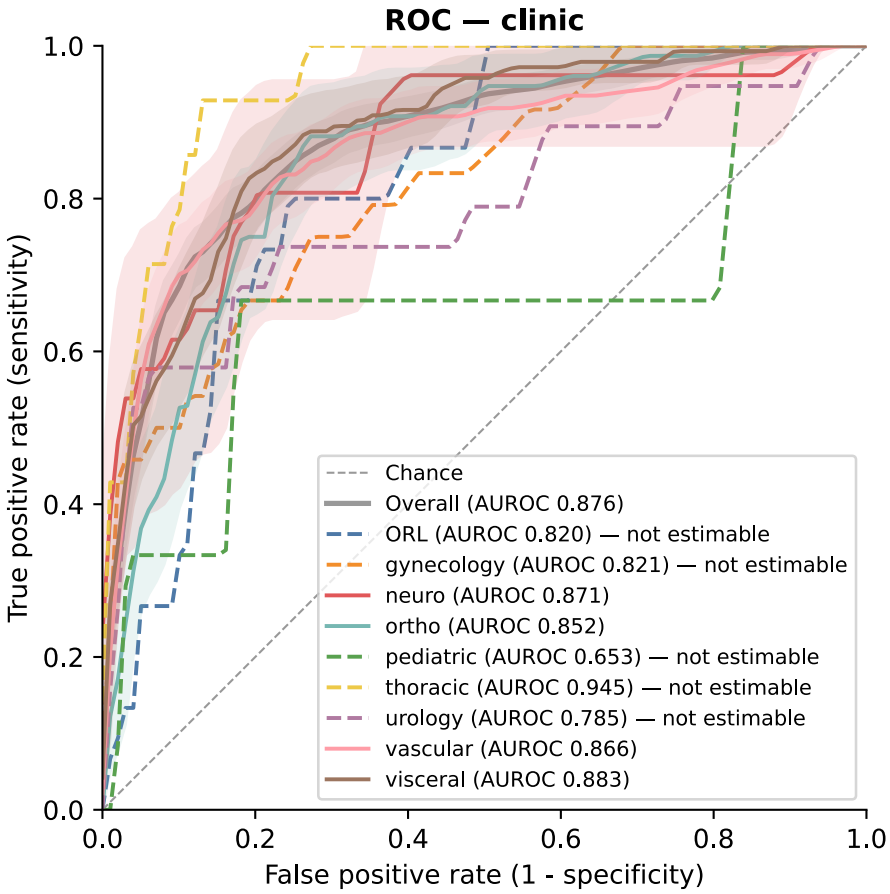
ROC curves

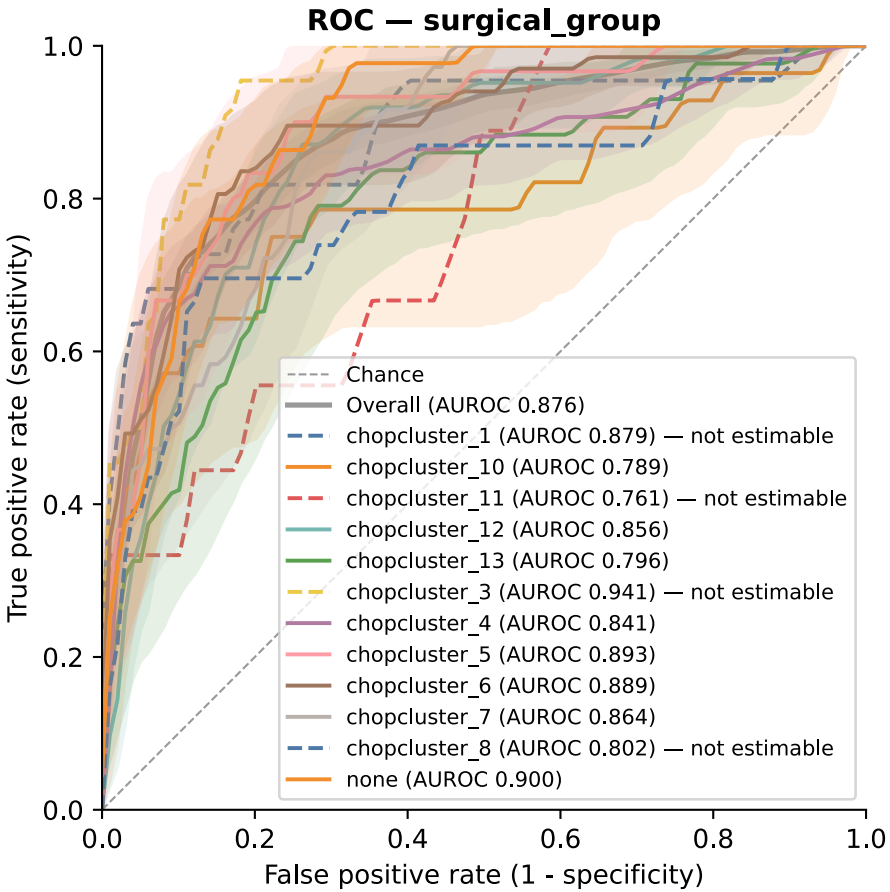
Per-subgroup ROC curves with pointwise 95% bootstrap bands. AUROC values in the legend are taken from the table above, not integrated from the drawn curve. Curves that cross indicate that no single subgroup dominates at every operating point — the ranking of subgroups depends on the threshold chosen.

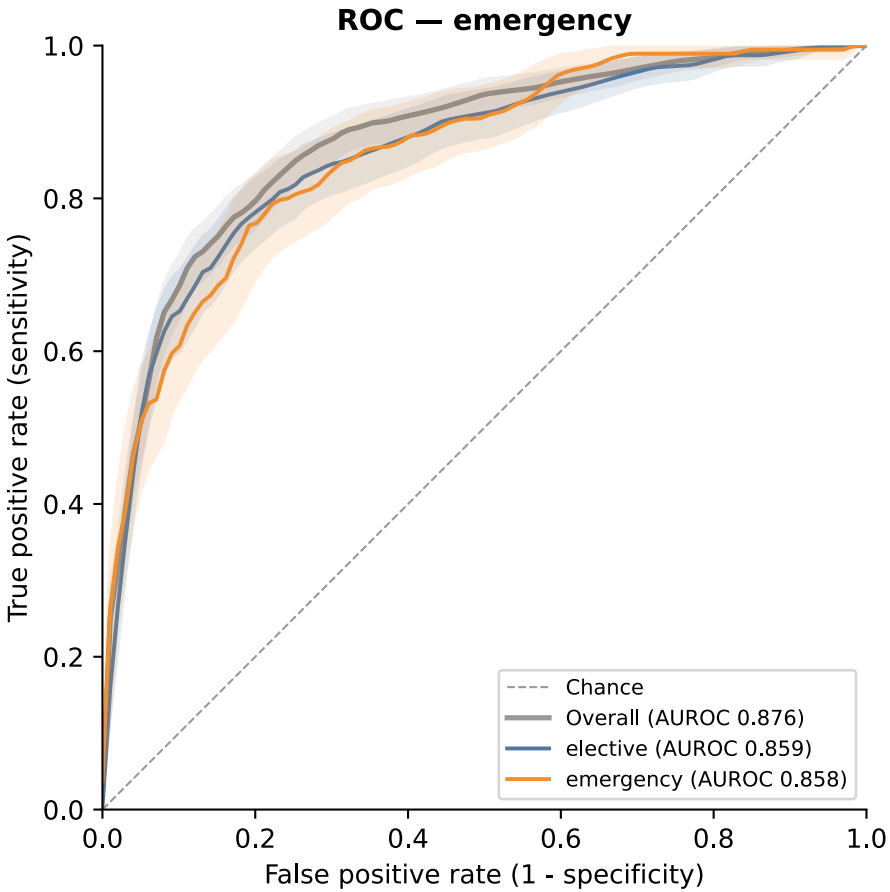


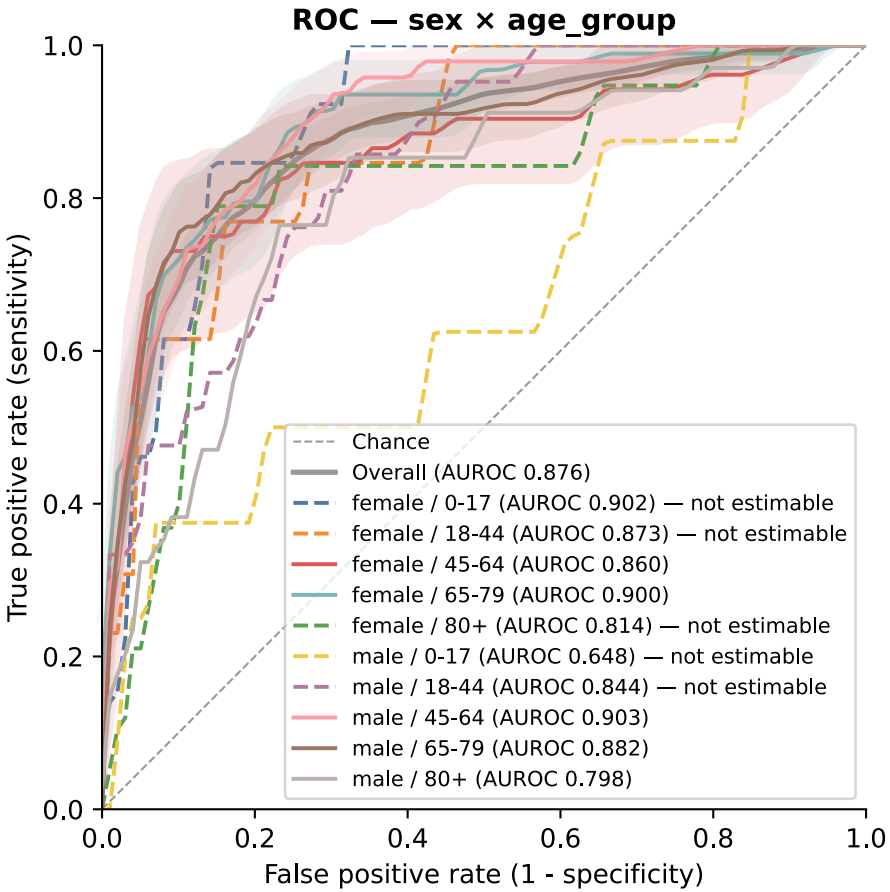


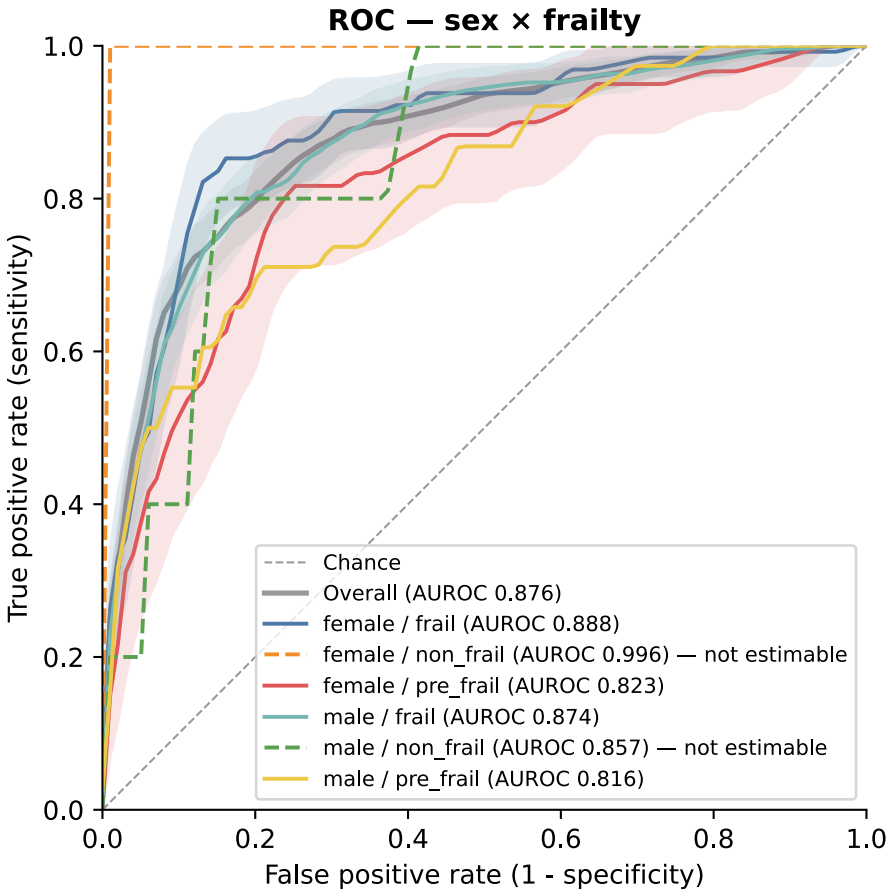

ROC — frailty
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
frail (AUROC 0.878)
non_frail (AUROC 0.868) — not estimable
pre_frail (AUROC 0.821)

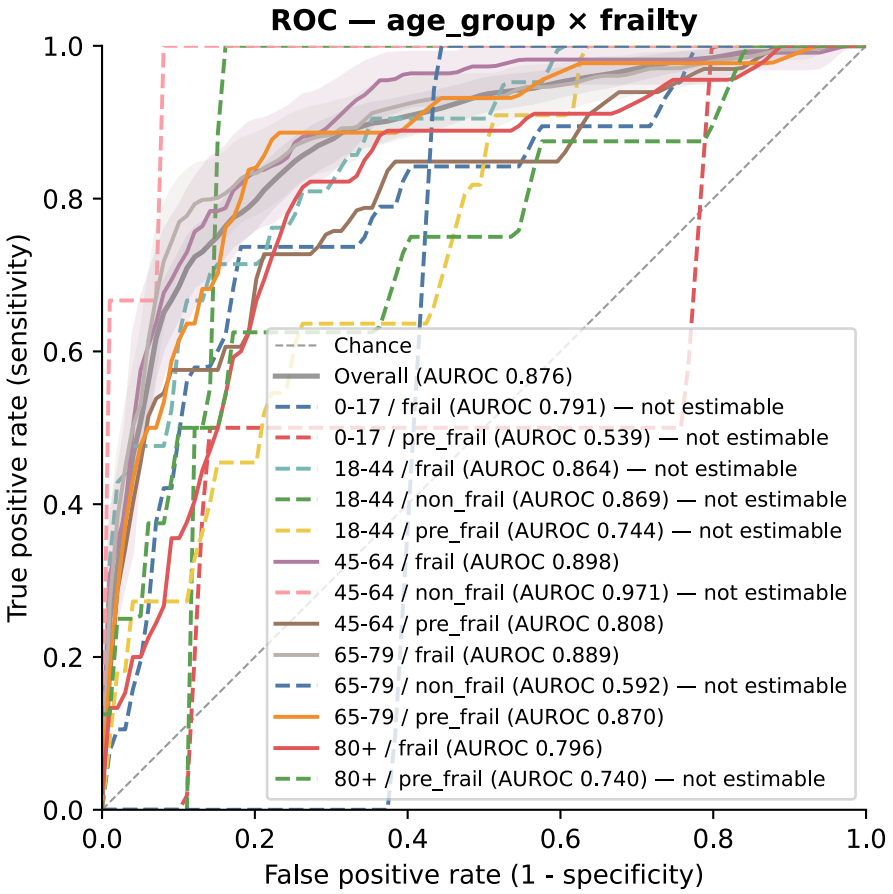

ROC — clinic
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
ORL (AUROC 0.820) — not estimable
gynecology (AUROC 0.821) — not estimable
neuro (AUROC 0.871)
ortho (AUROC 0.852)
pediatric (AUROC 0.653) — not estimable
thoracic (AUROC 0.945) — not estimable
urology (AUROC 0.785) — not estimable
vascular (AUROC 0.866)
visceral (AUROC 0.883)


ROC — surgical_group
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
chopcluster_1 (AUROC 0.879) — not estimable
chopcluster_10 (AUROC 0.789)
chopcluster_11 (AUROC 0.761) — not estimable
chopcluster_12 (AUROC 0.856)
chopcluster_13 (AUROC 0.796)
chopcluster_3 (AUROC 0.941) — not estimable
chopcluster_4 (AUROC 0.841)
chopcluster_5 (AUROC 0.893)
chopcluster_6 (AUROC 0.889)
chopcluster_7 (AUROC 0.864)
chopcluster_8 (AUROC 0.802) — not estimable
none (AUROC 0.900)


ROC — emergency
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
elective (AUROC 0.859)
emergency (AUROC 0.858)


ROC — sex × age_group
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
female / 0-17 (AUROC 0.902) — not estimable
female / 18-44 (AUROC 0.873) — not estimable
female / 45-64 (AUROC 0.860)
female / 65-79 (AUROC 0.900)
female / 80+ (AUROC 0.814) — not estimable
male / 0-17 (AUROC 0.648) — not estimable
male / 18-44 (AUROC 0.844) — not estimable
male / 45-64 (AUROC 0.903)
male / 65-79 (AUROC 0.882)
male / 80+ (AUROC 0.798)


ROC — sex × frailty
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
female / frail (AUROC 0.888)
female / non_frail (AUROC 0.996) — not estimable
female / pre_frail (AUROC 0.823)
male / frail (AUROC 0.874)
male / non_frail (AUROC 0.857) — not estimable
male / pre_frail (AUROC 0.816)


ROC — age_group × frailty
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
0-17 / frail (AUROC 0.791) — not estimable
0-17 / pre_frail (AUROC 0.539) — not estimable
18-44 / frail (AUROC 0.864) — not estimable
18-44 / non_frail (AUROC 0.869) — not estimable
18-44 / pre_frail (AUROC 0.744) — not estimable
45-64 / frail (AUROC 0.898)
45-64 / non_frail (AUROC 0.971) — not estimable
45-64 / pre_frail (AUROC 0.808)
65-79 / frail (AUROC 0.889)
65-79 / non_frail (AUROC 0.592) — not estimable
65-79 / pre_frail (AUROC 0.870)
80+ / frail (AUROC 0.796)
80+ / pre_frail (AUROC 0.740) — not estimable

Calibration

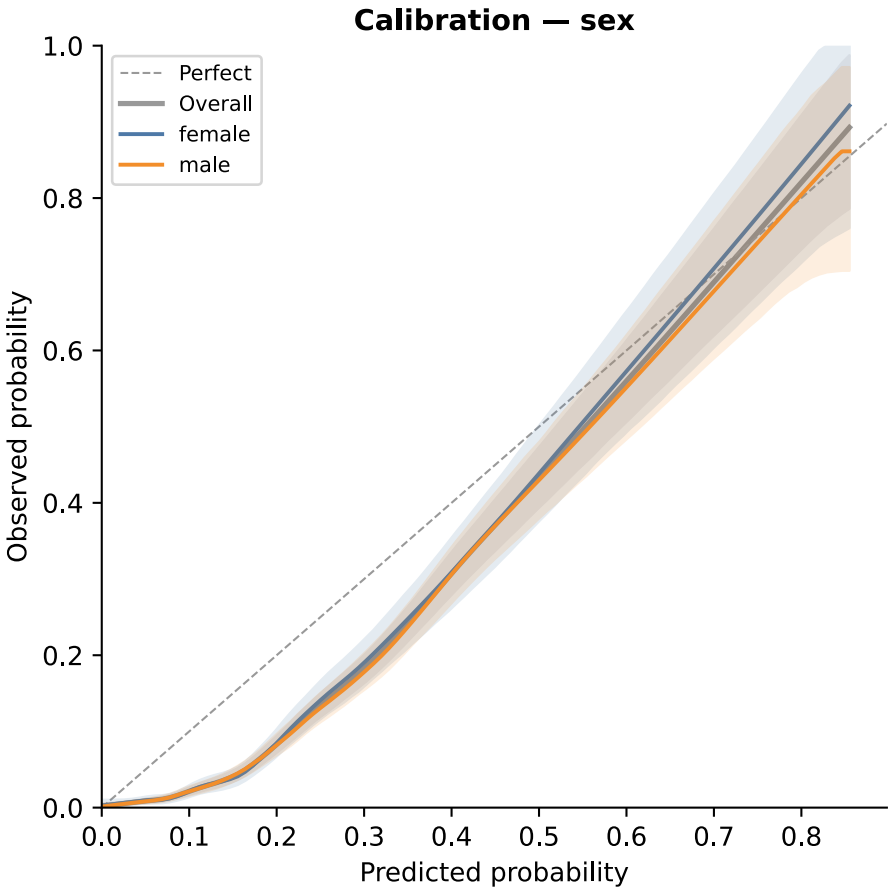
Attribute	Subgroup	n	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
sex	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
sex	female	3689	-0.963 (-1.112–-0.821)	1.561 (1.370–1.762)	0.043 (0.040–0.047)	0.0614	0.0600
sex	male	7030	-1.053 (-1.171–-0.940)	1.568 (1.423–1.730)	0.040 (0.037–0.043)	0.0629	0.0614
age_group	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
age_group	0-17	367	-0.962	1.542	0.055	0.0733	0.0713
age_group	18-44	941	-1.207 (-1.590–-0.906)	1.404	0.036 (0.029–0.044)	0.0647	0.0636
age_group	45-64	3715	-1.162 (-1.343–-1.007)	1.594 (1.386–1.845)	0.036 (0.033–0.040)	0.0646	0.0631
age_group	65-79	4684	-0.889 (-1.020–-0.768)	1.678 (1.502–1.876)	0.042 (0.039–0.046)	0.0568	0.0560
age_group	80+	1012	-1.069 (-1.376–-0.810)	1.164	0.053 (0.046–0.062)	0.0745	0.0748
frailty	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
frailty	frail	6139	-0.855 (-0.959–-0.757)	1.594 (1.463–1.741)	0.052 (0.049–0.055)	0.0637	0.0620
frailty	non_frail	1022	-2.475	1.546	0.011	0.0550	0.0550
frailty	pre_frail	3558	-1.342 (-1.554–-1.158)	1.282	0.031 (0.028–0.034)	0.0623	0.0617
clinic	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
clinic	ORL	233	-1.150	1.302	0.068	0.0984	0.0982
clinic	gynecology	452	-0.692	1.414	0.044	0.0423	0.0450
clinic	neuro	1117	-1.306 (-1.769–-0.963)	1.858	0.022 (0.017–0.028)	0.0522	0.0536
clinic	ortho	1690	-1.073 (-1.331–-0.866)	1.325	0.045 (0.039–0.051)	0.0654	0.0649
clinic	pediatric	65	-1.218	0.710	0.054	0.1069	0.0887
clinic	thoracic	723	-1.542	1.982	0.020	0.0560	0.0552
clinic	urology	349	-0.776	1.286	0.048	0.0546	0.0510
clinic	vascular	4852	-1.147 (-1.297–-1.008)	1.657 (1.443–1.895)	0.036 (0.033–0.039)	0.0639	0.0624
clinic	visceral	1238	-0.655 (-0.825–-0.479)	1.572 (1.339–1.849)	0.078 (0.071–0.086)	0.0713	0.0669
surgical_group	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
surgical_group	chopcluster_1	903	-1.323	1.834	0.023	0.0546	0.0551
surgical_group	chopcluster_10	346	-0.493 (-0.954–-0.138)	1.360	0.064 (0.047–0.082)	0.0599	0.0516
surgical_group	chopcluster_11	303	-1.111	1.123	0.030	0.0533	0.0502
surgical_group	chopcluster_12	1623	-1.174 (-1.447–-0.941)	1.339	0.040 (0.034–0.046)	0.0658	0.0650
surgical_group	chopcluster_13	213	-0.112 (-0.451–0.206)	1.377	0.127 (0.103–0.156)	0.0579	0.0323
surgical_group	chopcluster_3	807	-1.351	1.937	0.026	0.0587	0.0581
surgical_group	chopcluster_4	3764	-1.266 (-1.468–-1.093)	1.583 (1.306–1.889)	0.032 (0.029–0.035)	0.0644	0.0630

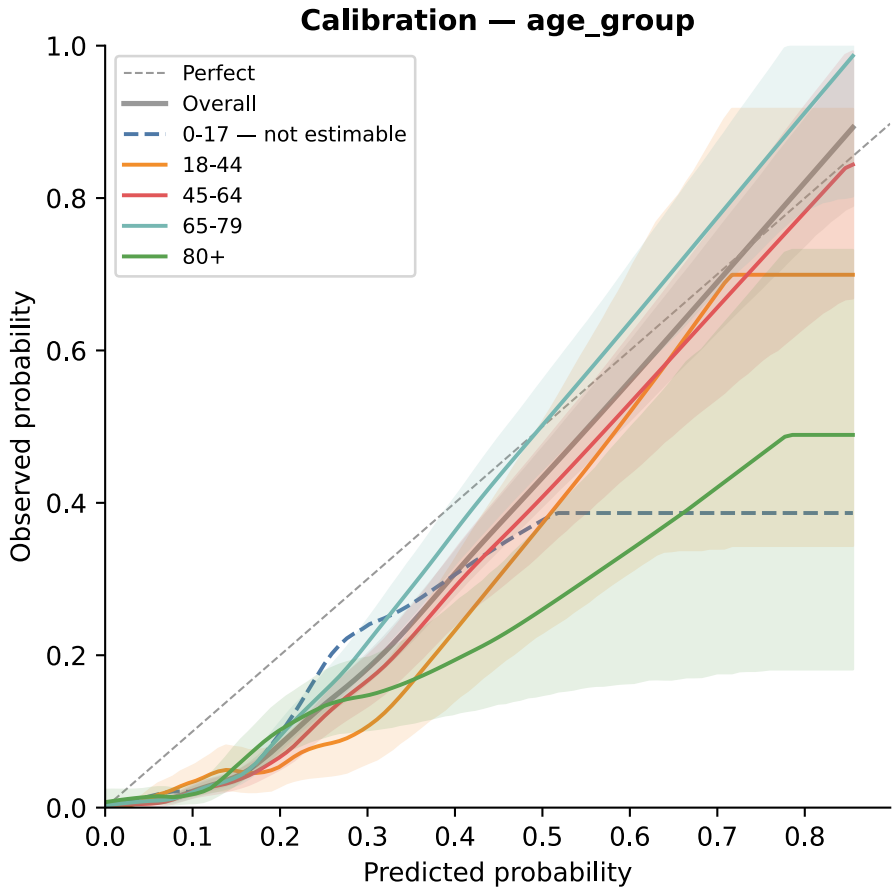
Attribute	Subgroup	n	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
surgical_group	chopcluster_5	769	-1.201 (-1.605–-0.877)	1.727	0.036 (0.029–0.044)	0.0684	0.0670
surgical_group	chopcluster_6	530	-0.565 (-0.838–-0.316)	1.621	0.078 (0.066–0.091)	0.0665	0.0646
surgical_group	chopcluster_7	298	-0.788 (-1.178–-0.462)	1.752	0.090 (0.074–0.105)	0.0932	0.0928
surgical_group	chopcluster_8	306	-0.903	1.085	0.068	0.0823	0.0739
surgical_group	none	857	-1.006 (-1.367–-0.717)	1.582	0.045 (0.038–0.053)	0.0644	0.0632
emergency	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
emergency	elective	9250	-1.156 (-1.270–-1.046)	1.488 (1.351–1.630)	0.034 (0.031–0.036)	0.0588	0.0577
emergency	emergency	1469	-0.705 (-0.860–-0.561)	1.727 (1.493–2.006)	0.089 (0.082–0.097)	0.0936	0.0878
sex × age_group	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
sex × age_group	*female / 0-17*	*175*	*-0.761*	*2.236*	*0.061*	*0.0771*	*0.0644*
sex × age_group	*female / 18-44*	*408*	*-1.395*	*1.530*	*0.034*	*0.0731*	*0.0720*
sex × age_group	female / 45-64	1179	-1.058 (-1.378–-0.785)	1.355	0.039 (0.032–0.046)	0.0606	0.0591
sex × age_group	female / 65-79	1470	-0.750 (-0.973–-0.546)	1.805	0.046 (0.040–0.053)	0.0519	0.0556
sex × age_group	*female / 80+*	*457*	*-1.310*	*1.179*	*0.048*	*0.0834*	*0.0839*
sex × age_group	*male / 0-17*	*192*	*-1.217*	*0.785*	*0.049*	*0.0859*	*0.0802*
sex × age_group	*male / 18-44*	*533*	*-1.072*	*1.334*	*0.038*	*0.0582*	*0.0578*
sex × age_group	male / 45-64	2536	-1.213 (-1.419–-1.023)	1.752	0.035 (0.031–0.039)	0.0666	0.0654
sex × age_group	male / 65-79	3214	-0.962 (-1.125–-0.809)	1.612 (1.403–1.863)	0.041 (0.037–0.045)	0.0590	0.0572
sex × age_group	*male / 80+*	*555*	*-0.904*	*1.168*	*0.058*	*0.0671*	*0.0676*
sex × frailty	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
sex × frailty	female / frail	1666	-0.819 (-1.006–-0.654)	1.683 (1.436–1.985)	0.058 (0.052–0.065)	0.0683	0.0672
sex × frailty	*female / non_frail*	*522*	*-3.691*	*4.400*	*0.009*	*0.0637*	*0.0638*
sex × frailty	female / pre_frail	1501	-0.983 (-1.257–-0.737)	1.226	0.039 (0.033–0.046)	0.0531	0.0525
sex × frailty	male / frail	4473	-0.872 (-0.999–-0.762)	1.561 (1.397–1.737)	0.050 (0.046–0.053)	0.0620	0.0601
sex × frailty	*male / non_frail*	*500*	*-1.834*	*1.374*	*0.014*	*0.0460*	*0.0459*
sex × frailty	*male / pre_frail*	*2057*	*-1.730*	*1.331*	*0.025*	*0.0691*	*0.0686*
age_group × frailty	Overall	10719	-1.020 (-1.116–-0.938)	1.566 (1.447–1.688)	0.041 (0.039–0.043)	0.0624	0.0609
age_group × frailty	*0-17 / frail*	*211*	*-0.689*	*1.442*	*0.079*	*0.0705*	*0.0664*
age_group × frailty	*0-17 / non_frail*	*45*	*—*	*—*	*0.007*	*0.0721*	*0.0721*
age_group × frailty	*0-17 / pre_frail*	*111*	*-1.857*	*0.242*	*0.028*	*0.0927*	*0.0844*
age_group × frailty	*18-44 / frail*	*300*	*-0.990*	*1.603*	*0.060*	*0.0827*	*0.0811*
age_group × frailty	*18-44 / non_frail*	*331*	*-2.473*	*1.062*	*0.013*	*0.0561*	*0.0562*
age_group × frailty	*18-44 / pre_frail*	*310*	*-1.078*	*0.965*	*0.039*	*0.0564*	*0.0567*

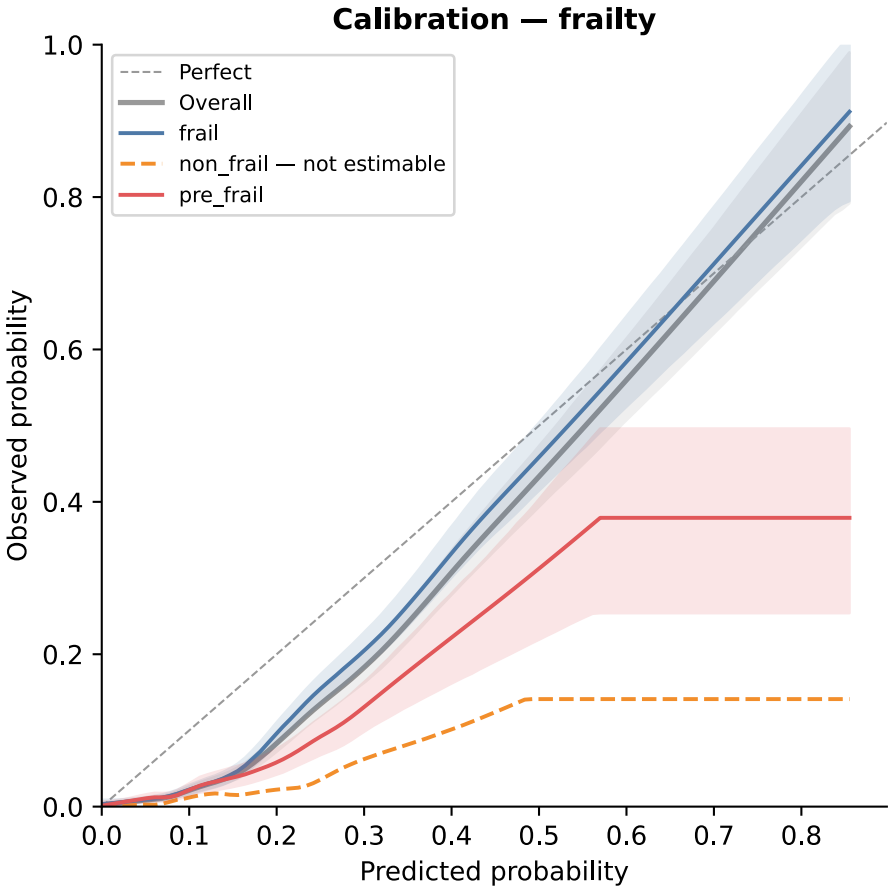
Attribute	Subgroup	n	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
age_group × frailty	45-64 / frail	1967	-1.000 (-1.198–-0.827)	1.633 (1.403–1.921)	0.046 (0.041–0.051)	0.0687	0.0667
age_group × frailty	*45-64 / non_frail*	*410*	*-2.215*	*2.787*	*0.011*	*0.0512*	*0.0510*
age_group × frailty	*45-64 / pre_frail*	*1338*	*-1.420*	*1.252*	*0.029*	*0.0625*	*0.0619*
age_group × frailty	65-79 / frail	3032	-0.758 (-0.907–-0.629)	1.727 (1.522–1.976)	0.050 (0.045–0.054)	0.0568	0.0575
age_group × frailty	*65-79 / non_frail*	*218*	*-2.693*	*0.196*	*0.011*	*0.0552*	*0.0553*
age_group × frailty	*65-79 / pre_frail*	*1434*	*-1.240*	*1.444*	*0.032*	*0.0593*	*0.0583*
age_group × frailty	*80+ / frail*	*629*	*-0.899*	*1.126*	*0.068*	*0.0753*	*0.0763*
age_group × frailty	*80+ / non_frail*	*18*	*—*	*—*	*0.012*	*0.0783*	*0.0783*
age_group × frailty	*80+ / pre_frail*	*365*	*-1.635*	*0.975*	*0.030*	*0.0739*	*0.0726*

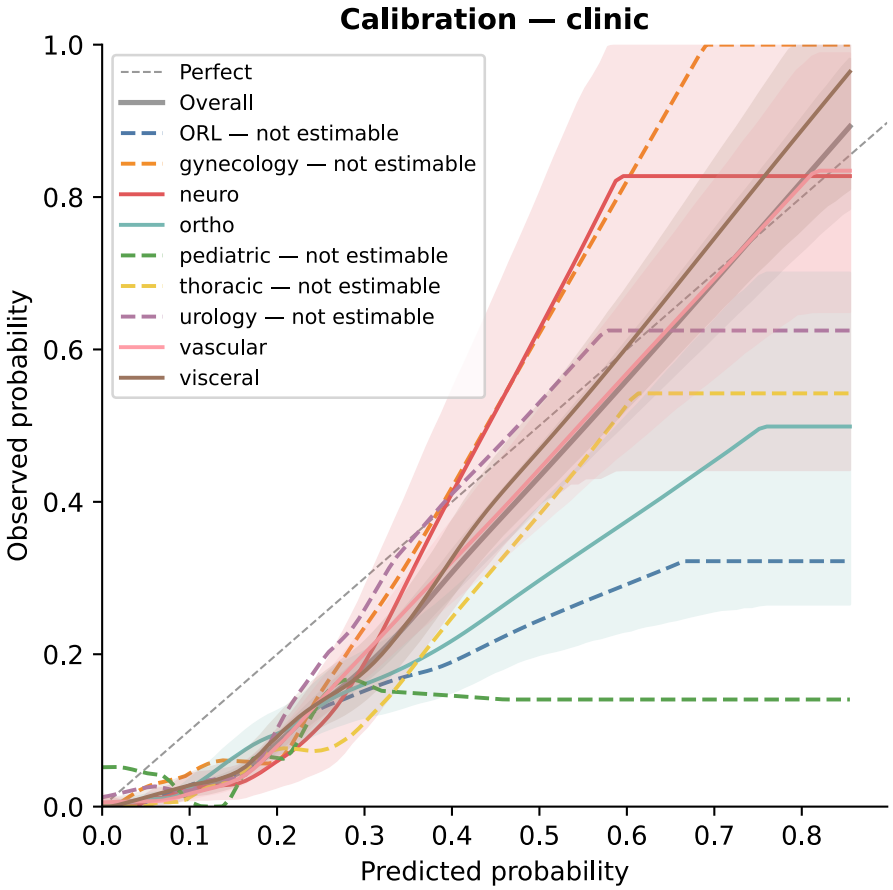
Calibration gap statistics

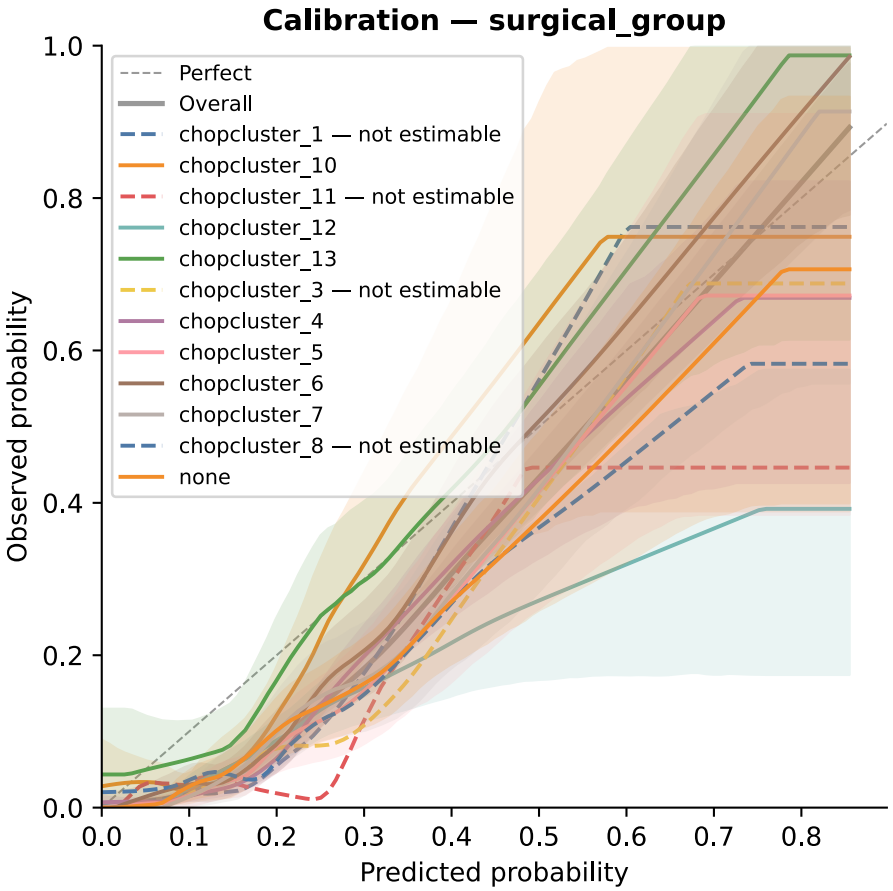
Attribute	Metric	Value (95% CI)
sex	max_intercept_gap	0.0905 (0.0048–0.2827)
sex	max_slope_gap	0.0066 (0.0038–0.2911)
age_group	max_intercept_gap	0.3186
age_group	max_slope_gap	0.5136
frailty	max_intercept_gap	1.6196
frailty	max_slope_gap	0.3126
clinic	max_intercept_gap	0.8870
clinic	max_slope_gap	1.2723
surgical_group	max_intercept_gap	1.2385
surgical_group	max_slope_gap	0.8526
emergency	max_intercept_gap	0.4515 (0.2610–0.6342)
emergency	max_slope_gap	0.2381 (0.0209–0.5473)
sex × age_group	max_intercept_gap	0.6449
sex × age_group	max_slope_gap	1.4507
sex × frailty	max_intercept_gap	2.8720
sex × frailty	max_slope_gap	3.1740
age_group × frailty	max_intercept_gap	2.0037
age_group × frailty	max_slope_gap	2.5912

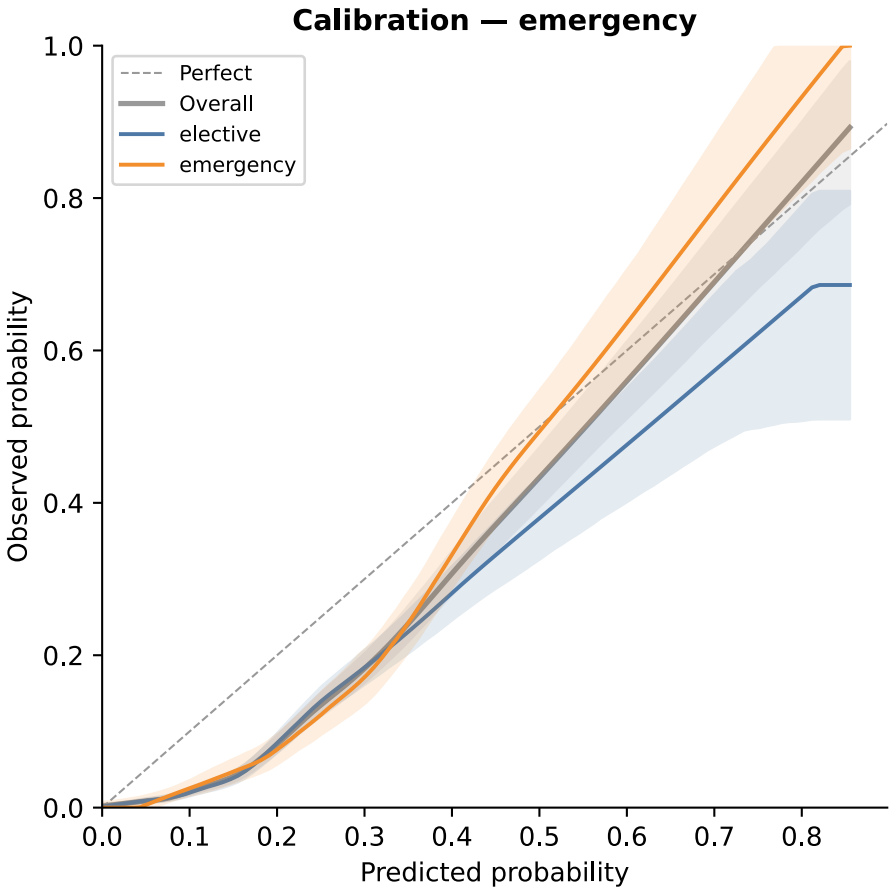

Calibration — sex
Perfect
Overall
female
male
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

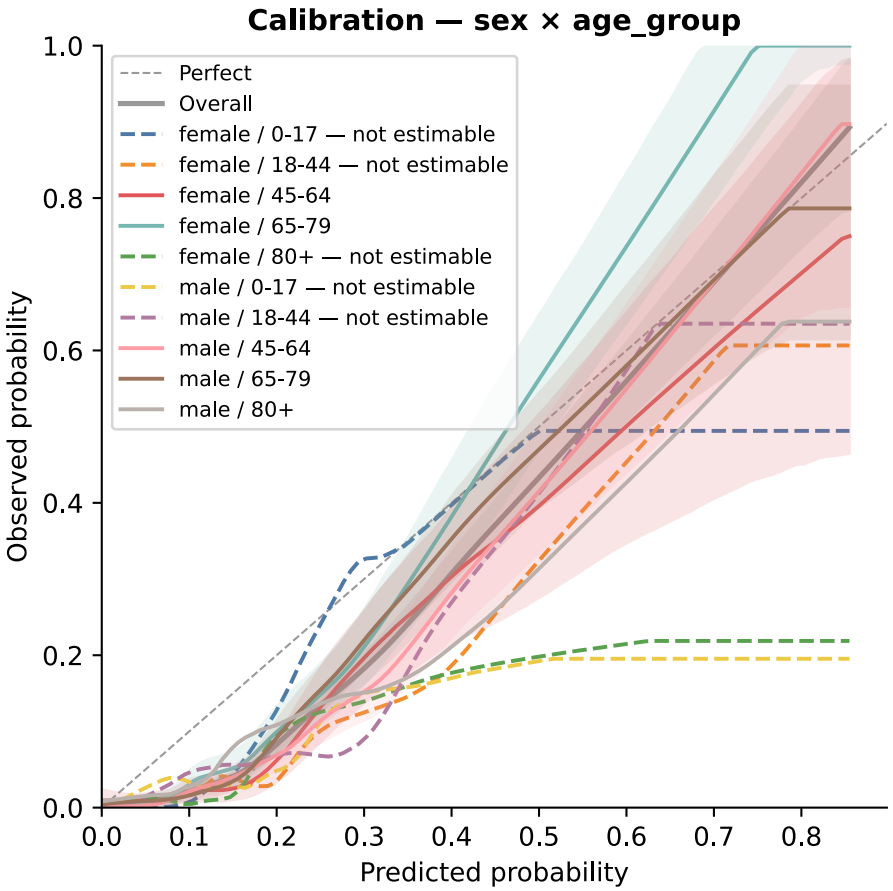

Calibration — age_group
Perfect
Overall
0-17 — not estimable
18-44
45-64
65-79
80+
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

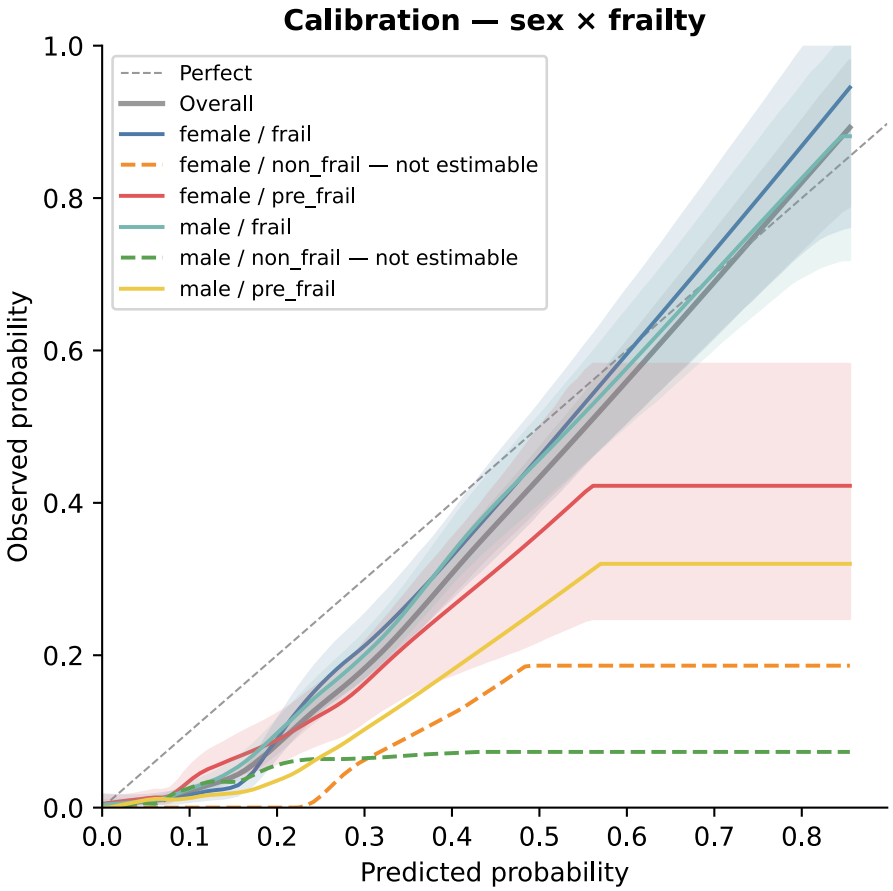

Calibration — frailty
Perfect
Overall
frail
non_frail — not estimable
pre_frail
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

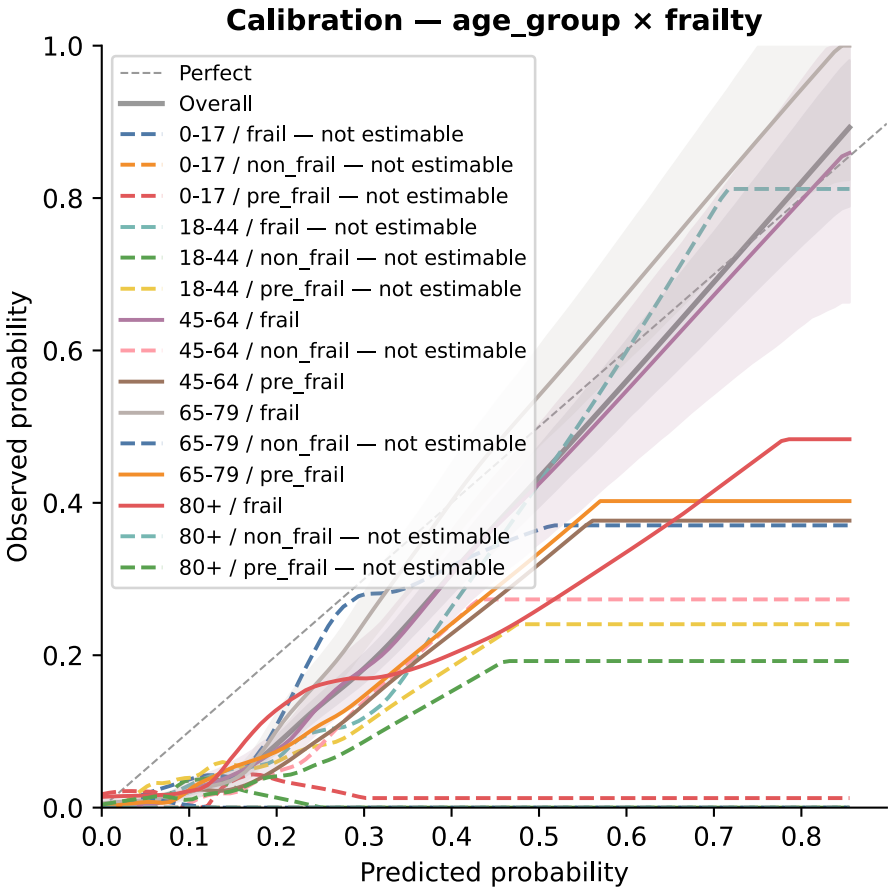

Calibration — clinic
Perfect
Overall
ORL — not estimable
gynecology — not estimable
neuro
ortho
pediatric — not estimable
thoracic — not estimable
urology — not estimable
vascular
visceral
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Calibration — surgical_group
Perfect
Overall
chopcluster_1 — not estimable
chopcluster_10
chopcluster_11 — not estimable
chopcluster_12
chopcluster_13
chopcluster_3 — not estimable
chopcluster_4
chopcluster_5
chopcluster_6
chopcluster_7
chopcluster_8 — not estimable
none
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7


Calibration — emergency
Perfect
Overall
elective
emergency
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7


Calibration — sex × age_group
Perfect
Overall
female / 0-17 — not estimable
female / 18-44 — not estimable
female / 45-64
female / 65-79
female / 80+ — not estimable
male / 0-17 — not estimable
male / 18-44 — not estimable
male / 45-64
male / 65-79
male / 80+
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7


Calibration — sex × frailty
Perfect
Overall
female / frail
female / non_frail — not estimable
female / pre_frail
male / frail
male / non_frail — not estimable
male / pre_frail
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7



Hosmer–Lemeshow goodness-of-fit

Read every statistic beside its *n* and its decile table. The Hosmer–Lemeshow statistic scales with subgroup size: a clinically irrelevant deviation becomes significant in a large subgroup, while a large deviation can go undetected in a small one. Values are therefore **not comparable across subgroups of different size**, and subgroups must not be ranked by p-value. This is why Kramer & Zimmerman (2007) require the statistic to be reported together with the number of patients, the per-decile observed-versus-expected table below, and the adjunct calibration measures in the table above — never on its own.

Attribute	Subgroup	n	Events	H–L statistic	df	p	Groups used	Decile slope	Decile intercept
sex	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035
sex	female	3689	190	177.76	10	0.0000	10	0.777	-0.037
sex	male	7030	314	352.91	10	0.0000	10	0.718	-0.033
age_group	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035

Attribute	Subgroup	n	Events	H–L statistic	df	p	Groups used	Decile slope	Decile intercept
age_group	0-17	367	21	22.54	10	0.0126	10	0.833	-0.050
age_group	18-44	941	34	50.47	10	0.0000	10	0.532	-0.018
age_group	45-64	3715	147	201.77	10	0.0000	10	0.675	-0.031
age_group	65-79	4684	249	220.84	10	0.0000	10	0.884	-0.044
age_group	80+	1012	53	60.51	10	0.0000	10	0.535	-0.016
frailty	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035
frailty	frail	6139	400	293.55	10	0.0000	10	0.860	-0.046
frailty	non_frail	1022	6	58.09	10	0.0000	10	0.150	-0.003
frailty	pre_frail	3558	98	187.73	10	0.0000	10	0.436	-0.011
clinic	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035
clinic	ORL	233	15	22.35	10	0.0134	10	0.537	-0.023
clinic	gynecology	452	24	13.57	10	0.1937	10	0.783	-0.021
clinic	neuro	1117	26	50.71	10	0.0000	10	0.561	-0.020
clinic	ortho	1690	76	86.86	10	0.0000	10	0.593	-0.021
clinic	pediatric	65	3	7.93	10	0.6359	10	0.335	-0.001
clinic	thoracic	723	14	37.02	10	0.0001	10	0.477	-0.016
clinic	urology	349	19	17.83	10	0.0578	10	0.856	-0.036
clinic	vascular	4852	184	263.12	10	0.0000	10	0.701	-0.034
clinic	visceral	1238	143	66.58	10	0.0000	10	0.988	-0.066
surgical_group	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035
surgical_group	chopcluster_1	903	22	44.33	10	0.0000	10	0.575	-0.022
surgical_group	chopcluster_10	346	28	15.64	10	0.1105	10	1.125	-0.056
surgical_group	chopcluster_11	303	9	13.28	10	0.2085	10	0.339	+0.003
surgical_group	chopcluster_12	1623	62	86.61	10	0.0000	10	0.563	-0.020
surgical_group	chopcluster_13	213	43	5.89	10	0.8241	10	1.242	-0.070
surgical_group	chopcluster_3	807	22	42.67	10	0.0000	10	0.604	-0.025
surgical_group	chopcluster_4	3764	118	209.96	10	0.0000	10	0.608	-0.027
surgical_group	chopcluster_5	769	30	46.17	10	0.0000	10	0.690	-0.034
surgical_group	chopcluster_6	530	67	25.39	10	0.0046	10	1.076	-0.074
surgical_group	chopcluster_7	298	36	23.18	10	0.0101	10	0.990	-0.092
surgical_group	chopcluster_8	306	23	21.05	10	0.0207	10	0.653	-0.023
surgical_group	none	857	44	43.12	10	0.0000	10	0.695	-0.030
emergency	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035

Attribute	Subgroup	n	Events	H–L statistic	df	p	Groups used	Decile slope	Decile intercept
emergency	elective	9250	316	441.97	10	0.0000	10	0.635	-0.025
emergency	emergency	1469	188	104.02	10	0.0000	10	1.123	-0.112
sex × age_group	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035
sex × age_group	*female / 0-17*	*175*	*13*	*13.85*	*10*	*0.1799*	*10*	*1.176*	*-0.091*
sex × age_group	*female / 18-44*	*408*	*13*	*27.47*	*10*	*0.0022*	*10*	*0.542*	*-0.025*
sex × age_group	female / 45-64	1179	52	60.90	10	0.0000	10	0.708	-0.030
sex × age_group	female / 65-79	1470	93	61.58	10	0.0000	10	0.946	-0.046
sex × age_group	*female / 80+*	*457*	*19*	*37.70*	*10*	*0.0000*	*10*	*0.466*	*-0.017*
sex × age_group	*male / 0-17*	*192*	*8*	*15.38*	*10*	*0.1188*	*10*	*0.388*	*-0.008*
sex × age_group	*male / 18-44*	*533*	*21*	*24.01*	*10*	*0.0076*	*10*	*0.521*	*-0.011*
sex × age_group	male / 45-64	2536	95	143.70	10	0.0000	10	0.648	-0.030
sex × age_group	male / 65-79	3214	156	159.99	10	0.0000	10	0.834	-0.041
sex × age_group	*male / 80+*	*555*	*34*	*27.49*	*10*	*0.0022*	*10*	*0.590*	*-0.014*
sex × frailty	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035
sex × frailty	female / frail	1666	129	95.15	10	0.0000	10	0.915	-0.056
sex × frailty	*female / non_frail*	*522*	*1*	*37.09*	*10*	*0.0001*	*10*	*0.077*	*-0.003*
sex × frailty	female / pre_frail	1501	60	57.99	10	0.0000	10	0.580	-0.014
sex × frailty	male / frail	4473	271	207.44	10	0.0000	10	0.837	-0.042
sex × frailty	*male / non_frail*	*500*	*5*	*21.78*	*10*	*0.0162*	*10*	*0.247*	*-0.004*
sex × frailty	*male / pre_frail*	*2057*	*38*	*135.81*	*10*	*0.0000*	*10*	*0.312*	*-0.009*
age_group × frailty	Overall	10719	504	531.10	10	0.0000	10	0.746	-0.035
age_group × frailty	*0-17 / frail*	*211*	*19*	*12.10*	*10*	*0.2786*	*10*	*0.942*	*-0.058*
age_group × frailty	*0-17 / non_frail*	*45*	*0*	—	—	—	*10*	—	—
age_group × frailty	*0-17 / pre_frail*	*111*	*2*	*11.79*	*10*	*0.2992*	*10*	*-0.020*	*+0.021*
age_group × frailty	*18-44 / frail*	*300*	*21*	*19.58*	*10*	*0.0335*	*10*	*0.721*	*-0.041*
age_group × frailty	*18-44 / non_frail*	*331*	*2*	*20.32*	*10*	*0.0264*	*10*	*0.078*	*+0.001*
age_group × frailty	*18-44 / pre_frail*	*310*	*11*	*14.85*	*10*	*0.1377*	*10*	*0.353*	*+0.004*
age_group × frailty	45-64 / frail	1967	111	109.75	10	0.0000	10	0.803	-0.044
age_group × frailty	*45-64 / non_frail*	*410*	*3*	*21.21*	*10*	*0.0197*	*10*	*0.315*	*-0.011*
age_group × frailty	*45-64 / pre_frail*	*1338*	*33*	*77.02*	*10*	*0.0000*	*10*	*0.419*	*-0.011*
age_group × frailty	65-79 / frail	3032	204	139.86	10	0.0000	10	0.985	-0.054
age_group × frailty	*65-79 / non_frail*	*218*	*1*	*13.51*	*10*	*0.1968*	*10*	*-0.015*	*+0.005*
age_group × frailty	*65-79 / pre_frail*	*1434*	*44*	*71.06*	*10*	*0.0000*	*10*	*0.529*	*-0.017*

Attribute	Subgroup	n	Events	H–L statistic	df	p	Groups used	Decile slope	Decile intercept
age_group × frailty	*80+ / frail*	*629*	*45*	*40.56*	*10*	*0.0000*	*10*	*0.592*	*-0.016*
age_group × frailty	*80+ / non_frail*	*18*	*0*	*—*	*—*	*—*	*10*	*—*	*—*
age_group × frailty	*80+ / pre_frail*	*365*	*8*	*26.77*	*10*	*0.0028*	*10*	*0.257*	*-0.002*

† Fewer risk groups than requested were usable, because the subgroup was small or its predicted probabilities were tied. Degrees of freedom follow the reduced count. A well-calibrated model gives a decile slope near 1 and a decile intercept near 0 (Kramer & Zimmerman 2007). 73 tests are reported here; no multiplicity adjustment has been applied. Adjust within an attribute if needed — do not pool the “Overall” rows with the subgroup rows.

Attribute	Metric	Value
sex	max_hl_gap	175.15
age_group	max_hl_gap	198.30
frailty	max_hl_gap	235.46
clinic	max_hl_gap	255.19
surgical_group	max_hl_gap	204.07
emergency	max_hl_gap	337.95
sex × age_group	max_hl_gap	146.14
sex × frailty	max_hl_gap	185.66
age_group × frailty	max_hl_gap	128.07

The gap is the largest pairwise difference in the statistic across subgroups. It inherits the sample-size dependence in full and should be read as a descriptive summary, not as evidence of differential calibration.

Observed versus expected events per decile

Deciles of predicted risk, of approximately equal size, with cut points snapped to tied predicted probabilities so that identical predictions always share a decile.

sex — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

sex — female

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	436	3	4.1	0.0069	0.0095	-1.1
2	278	1	6.3	0.0036	0.0227	-5.3
3	427	3	14.8	0.0070	0.0348	-11.8
4	389	3	19.3	0.0077	0.0496	-16.3
5	328	5	21.2	0.0152	0.0645	-16.2
6	343	4	28.2	0.0117	0.0823	-24.2
7	366	5	39.1	0.0137	0.1068	-34.1
8	393	19	57.9	0.0483	0.1473	-38.9
9	367	31	80.5	0.0845	0.2195	-49.5
10	362	116	145.1	0.3204	0.4008	-29.1

sex — male

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	791	1	7.2	0.0013	0.0091	-6.2
2	555	3	12.6	0.0054	0.0226	-9.6
3	895	7	31.2	0.0078	0.0348	-24.2
4	471	5	22.3	0.0106	0.0473	-17.3
5	878	6	54.7	0.0068	0.0623	-48.7
6	681	9	56.1	0.0132	0.0824	-47.1
7	684	15	73.2	0.0219	0.1070	-58.2
8	682	25	97.5	0.0367	0.1430	-72.5
9	679	47	137.5	0.0692	0.2025	-90.5
10	714	196	264.0	0.2745	0.3697	-68.0

age_group — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

age_group — 0-17

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	37	0	0.9	0.0000	0.0253	-0.9
2	36	1	1.7	0.0278	0.0476	-0.7
3	38	0	2.5	0.0000	0.0649	-2.5
4	32	2	2.6	0.0625	0.0803	-0.6
5	45	0	4.4	0.0000	0.0980	-4.4
6	33	1	3.9	0.0303	0.1185	-2.9
7	39	1	5.5	0.0256	0.1419	-4.5
8	35	2	6.2	0.0571	0.1773	-4.2
9	38	4	8.4	0.1053	0.2204	-4.4
10	34	10	11.8	0.2941	0.3469	-1.8

age_group — 18-44

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	82	0	0.5	0.0000	0.0063	-0.5
2	105	0	2.1	0.0000	0.0198	-2.1
3	85	0	2.8	0.0000	0.0331	-2.8
4	115	1	5.1	0.0087	0.0444	-4.1
5	86	0	4.9	0.0000	0.0574	-4.9
6	100	4	7.3	0.0400	0.0730	-3.3
7	86	2	8.1	0.0233	0.0945	-6.1
8	98	4	12.6	0.0408	0.1285	-8.6
9	92	5	18.0	0.0543	0.1957	-13.0
10	92	18	33.4	0.1957	0.3627	-15.4

age_group — 45-64

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	348	1	2.2	0.0029	0.0064	-1.2
2	473	2	9.6	0.0042	0.0202	-7.6
3	321	2	10.4	0.0062	0.0324	-8.4
4	404	2	18.1	0.0050	0.0448	-16.1
5	336	0	20.2	0.0000	0.0601	-20.2
6	372	4	28.5	0.0108	0.0767	-24.5
7	362	6	36.8	0.0166	0.1016	-30.8
8	370	16	50.9	0.0432	0.1376	-34.9
9	363	17	72.2	0.0468	0.1990	-55.2
10	366	97	137.9	0.2650	0.3768	-40.9

age_group — 65-79

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	524	2	4.9	0.0038	0.0093	-2.9
2	380	2	8.6	0.0053	0.0226	-6.6
3	413	3	13.4	0.0073	0.0324	-10.4
4	498	5	22.2	0.0100	0.0446	-17.2
5	565	8	35.1	0.0142	0.0621	-27.1
6	458	1	37.9	0.0022	0.0828	-36.9
7	449	10	47.9	0.0223	0.1067	-37.9
8	443	18	63.3	0.0406	0.1430	-45.3
9	493	34	101.9	0.0690	0.2066	-67.9
10	461	166	179.7	0.3601	0.3898	-13.7

age_group — 80+

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	98	0	1.3	0.0000	0.0132	-1.3
2	105	2	3.2	0.0190	0.0300	-1.2
3	109	1	4.9	0.0092	0.0453	-3.9
4	99	3	5.9	0.0303	0.0596	-2.9
5	106	2	8.1	0.0189	0.0768	-6.1
6	100	0	9.9	0.0000	0.0992	-9.9
7	93	3	11.8	0.0323	0.1274	-8.8
8	102	8	18.1	0.0784	0.1774	-10.1
9	97	14	24.1	0.1443	0.2481	-10.1
10	103	20	41.0	0.1942	0.3983	-21.0

frailty — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

frailty — frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	604	2	7.8	0.0033	0.0129	-5.8
2	688	7	20.8	0.0102	0.0302	-13.8
3	608	3	27.3	0.0049	0.0449	-24.3
4	577	5	34.6	0.0087	0.0600	-29.6
5	646	9	49.9	0.0139	0.0773	-40.9
6	595	12	59.2	0.0202	0.0995	-47.2
7	586	21	75.1	0.0358	0.1281	-54.1
8	597	28	102.0	0.0469	0.1708	-74.0
9	636	94	154.7	0.1478	0.2432	-60.7
10	602	219	259.7	0.3638	0.4314	-40.7

frailty — non_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	94	0	0.3	0.0000	0.0030	-0.3
2	143	0	1.8	0.0000	0.0124	-1.8
3	61	0	1.2	0.0000	0.0200	-1.2
4	106	0	2.9	0.0000	0.0273	-2.9
5	110	0	4.1	0.0000	0.0373	-4.1
6	96	1	4.5	0.0104	0.0469	-3.5
7	113	0	6.7	0.0000	0.0596	-6.7
8	97	0	7.7	0.0000	0.0796	-7.7
9	98	2	11.1	0.0204	0.1133	-9.1
10	104	3	21.9	0.0288	0.2108	-18.9

frailty — pre_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	382	1	2.3	0.0026	0.0062	-1.3
2	333	1	5.8	0.0030	0.0173	-4.8
3	354	2	9.7	0.0056	0.0275	-7.7
4	293	4	11.0	0.0137	0.0375	-7.0
5	367	4	18.2	0.0109	0.0496	-14.2
6	410	5	27.6	0.0122	0.0672	-22.6
7	372	5	33.1	0.0134	0.0891	-28.1
8	330	10	39.3	0.0303	0.1192	-29.3
9	368	16	63.2	0.0435	0.1717	-47.2
10	349	50	109.3	0.1433	0.3133	-59.3

clinic — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

clinic — ORL

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	24	0	0.2	0.0000	0.0100	-0.2
2	26	0	0.8	0.0000	0.0325	-0.8
3	23	0	1.4	0.0000	0.0630	-1.4
4	20	0	1.8	0.0000	0.0885	-1.8
5	23	2	2.7	0.0870	0.1170	-0.7
6	24	1	3.4	0.0417	0.1425	-2.4
7	23	0	4.2	0.0000	0.1830	-4.2
8	24	2	5.7	0.0833	0.2369	-3.7
9	23	6	7.3	0.2609	0.3193	-1.3
10	23	4	10.3	0.1739	0.4463	-6.3

clinic — gynecology

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	46	0	0.5	0.0000	0.0100	-0.5
2	46	0	1.0	0.0000	0.0227	-1.0
3	45	0	1.5	0.0000	0.0331	-1.5
4	39	2	1.7	0.0513	0.0429	+0.3
5	52	2	2.9	0.0385	0.0555	-0.9
6	41	1	3.0	0.0244	0.0741	-2.0
7	48	1	4.5	0.0208	0.0936	-3.5
8	45	3	5.4	0.0667	0.1208	-2.4
9	45	3	7.9	0.0667	0.1762	-4.9
10	45	12	14.6	0.2667	0.3251	-2.6

clinic — neuro

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	128	1	0.8	0.0078	0.0059	+0.2
2	116	0	2.0	0.0000	0.0172	-2.0
3	63	0	1.6	0.0000	0.0250	-1.6
4	165	0	5.7	0.0000	0.0347	-5.7
5	91	0	4.3	0.0000	0.0474	-4.3
6	94	0	5.6	0.0000	0.0593	-5.6
7	135	4	10.4	0.0296	0.0771	-6.4
8	102	1	10.2	0.0098	0.1001	-9.2
9	115	4	15.9	0.0348	0.1385	-11.9
10	108	16	27.8	0.1481	0.2574	-11.8

clinic — ortho

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	158	0	0.9	0.0000	0.0058	-0.9
2	199	1	4.0	0.0050	0.0201	-3.0
3	169	1	5.9	0.0059	0.0349	-4.9
4	167	2	8.4	0.0120	0.0501	-6.4
5	161	2	10.4	0.0124	0.0646	-8.4
6	146	1	12.0	0.0068	0.0822	-11.0
7	183	2	19.8	0.0109	0.1084	-17.8
8	172	12	26.1	0.0698	0.1520	-14.1
9	164	20	35.3	0.1220	0.2151	-15.3
10	171	35	63.7	0.2047	0.3726	-28.7

clinic — pediatric

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	7	0	0.2	0.0000	0.0229	-0.2
2	7	1	0.4	0.1429	0.0507	+0.6
3	7	0	0.5	0.0000	0.0729	-0.5
4	8	0	0.7	0.0000	0.0900	-0.7
5	5	0	0.5	0.0000	0.1070	-0.5
6	6	0	0.7	0.0000	0.1233	-0.7
7	7	0	1.0	0.0000	0.1443	-1.0
8	7	1	1.4	0.1429	0.1936	-0.4
9	5	0	1.2	0.0000	0.2500	-1.2
10	6	1	2.0	0.1667	0.3375	-1.0

clinic — thoracic

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	57	0	0.2	0.0000	0.0028	-0.2
2	101	0	1.3	0.0000	0.0125	-1.3
3	43	0	0.9	0.0000	0.0200	-0.9
4	85	0	2.3	0.0000	0.0276	-2.3
5	80	0	3.0	0.0000	0.0378	-3.0
6	66	0	3.3	0.0000	0.0493	-3.3
7	77	0	5.1	0.0000	0.0660	-5.1
8	75	1	7.1	0.0133	0.0951	-6.1
9	68	3	9.8	0.0441	0.1439	-6.8
10	71	10	21.6	0.1408	0.3042	-11.6

clinic — urology

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	37	1	0.4	0.0270	0.0114	+0.6
2	23	0	0.6	0.0000	0.0274	-0.6
3	43	1	1.7	0.0233	0.0391	-0.7
4	37	0	2.1	0.0000	0.0554	-2.1
5	35	2	2.4	0.0571	0.0693	-0.4
6	36	1	3.1	0.0278	0.0869	-2.1
7	35	0	3.9	0.0000	0.1126	-3.9
8	34	2	4.8	0.0588	0.1415	-2.8
9	35	1	6.6	0.0286	0.1894	-5.6
10	34	11	11.1	0.3235	0.3265	-0.1

clinic — vascular

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	456	2	4.4	0.0044	0.0096	-2.4
2	602	4	15.3	0.0066	0.0254	-11.3
3	423	5	15.8	0.0118	0.0374	-10.8
4	536	3	26.7	0.0056	0.0498	-23.7
5	479	3	31.1	0.0063	0.0649	-28.1
6	371	2	29.7	0.0054	0.0801	-27.7
7	563	9	57.1	0.0160	0.1015	-48.1
8	455	14	61.6	0.0308	0.1354	-47.6
9	483	22	91.2	0.0455	0.1888	-69.2
10	484	120	161.3	0.2479	0.3332	-41.3

clinic — visceral

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	116	0	1.4	0.0000	0.0117	-1.4
2	125	1	4.0	0.0080	0.0324	-3.0
3	146	2	8.0	0.0137	0.0550	-6.0
4	120	1	9.5	0.0083	0.0793	-8.5
5	109	4	11.6	0.0367	0.1068	-7.6
6	129	6	18.8	0.0465	0.1460	-12.8
7	125	7	25.3	0.0560	0.2024	-18.3
8	126	25	34.2	0.1984	0.2718	-9.2
9	120	25	44.6	0.2083	0.3714	-19.6
10	122	72	69.4	0.5902	0.5685	+2.6

surgical_group — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

surgical_group — chopcluster_1

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	107	1	0.6	0.0093	0.0059	+0.4
2	92	0	1.6	0.0000	0.0171	-1.6
3	92	0	2.5	0.0000	0.0274	-2.5
4	86	0	3.2	0.0000	0.0373	-3.2
5	76	0	3.6	0.0000	0.0474	-3.6
6	77	0	4.6	0.0000	0.0593	-4.6
7	102	3	7.8	0.0294	0.0766	-4.8
8	88	1	8.8	0.0114	0.1006	-7.8
9	94	2	13.3	0.0213	0.1417	-11.3
10	89	15	25.2	0.1685	0.2833	-10.2

surgical_group — chopcluster_10

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	34	1	0.5	0.0294	0.0151	+0.5
2	32	1	1.1	0.0312	0.0345	-0.1
3	36	1	1.8	0.0278	0.0504	-0.8
4	42	2	2.8	0.0476	0.0673	-0.8
5	32	1	2.7	0.0312	0.0859	-1.7
6	32	0	3.4	0.0000	0.1052	-3.4
7	38	1	5.0	0.0263	0.1317	-4.0
8	34	3	5.7	0.0882	0.1663	-2.7
9	33	3	7.1	0.0909	0.2162	-4.1
10	33	15	11.9	0.4545	0.3594	+3.1

surgical_group — chopcluster_11

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	29	0	0.2	0.0000	0.0078	-0.2
2	30	0	0.5	0.0000	0.0180	-0.5
3	33	0	0.9	0.0000	0.0271	-0.9
4	30	0	1.1	0.0000	0.0370	-1.1
5	34	2	1.7	0.0588	0.0487	+0.3
6	22	1	1.4	0.0455	0.0623	-0.4
7	37	1	2.9	0.0270	0.0793	-1.9
8	32	1	3.2	0.0312	0.0997	-2.2
9	26	1	3.5	0.0385	0.1362	-2.5
10	30	3	8.6	0.1000	0.2860	-5.6

surgical_group — chopcluster_12

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	159	0	0.9	0.0000	0.0059	-0.9
2	197	1	3.9	0.0051	0.0201	-2.9
3	108	0	3.5	0.0000	0.0322	-3.5
4	166	2	7.4	0.0120	0.0446	-5.4
5	168	0	10.1	0.0000	0.0600	-10.1
6	169	2	13.1	0.0118	0.0774	-11.1
7	172	3	17.5	0.0174	0.1016	-14.5
8	166	10	23.5	0.0602	0.1414	-13.5
9	153	12	30.6	0.0784	0.2003	-18.6
10	165	32	58.2	0.1939	0.3530	-26.2

surgical_group — chopcluster_13

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	24	1	1.2	0.0417	0.0500	-0.2
2	20	2	1.6	0.1000	0.0808	+0.4
3	22	1	2.3	0.0455	0.1066	-1.3
4	21	1	2.8	0.0476	0.1312	-1.8
5	22	2	3.5	0.0909	0.1607	-1.5
6	19	2	3.8	0.1053	0.1995	-1.8
7	22	6	5.2	0.2727	0.2343	+0.8
8	21	7	6.4	0.3333	0.3048	+0.6
9	21	7	8.2	0.3333	0.3888	-1.2
10	21	14	11.5	0.6667	0.5464	+2.5

surgical_group — chopcluster_3

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	58	0	0.2	0.0000	0.0030	-0.2
2	109	0	1.4	0.0000	0.0126	-1.4
3	91	0	2.0	0.0000	0.0224	-2.0
4	81	0	2.6	0.0000	0.0322	-2.6
5	78	0	3.3	0.0000	0.0421	-3.3
6	78	0	4.3	0.0000	0.0557	-4.3
7	73	1	5.4	0.0137	0.0743	-4.4
8	76	0	8.0	0.0000	0.1057	-8.0
9	84	4	14.1	0.0476	0.1684	-10.1
10	79	17	28.0	0.2152	0.3541	-11.0

surgical_group — chopcluster_4

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	338	2	3.4	0.0059	0.0100	-1.4
2	481	4	12.3	0.0083	0.0255	-8.3
3	351	4	13.1	0.0114	0.0373	-9.1
4	302	1	14.3	0.0033	0.0475	-13.3
5	409	3	24.5	0.0073	0.0600	-21.5
6	426	4	32.8	0.0094	0.0771	-28.8
7	302	2	29.2	0.0066	0.0967	-27.2
8	396	10	49.0	0.0253	0.1237	-39.0
9	380	12	65.8	0.0316	0.1731	-53.8
10	379	76	116.0	0.2005	0.3061	-40.0

surgical_group — chopcluster_5

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	67	0	0.4	0.0000	0.0062	-0.4
2	96	0	1.9	0.0000	0.0201	-1.9
3	65	1	2.1	0.0154	0.0324	-1.1
4	72	0	3.2	0.0000	0.0448	-3.2
5	80	0	5.0	0.0000	0.0628	-5.0
6	77	1	6.3	0.0130	0.0822	-5.3
7	80	1	8.6	0.0125	0.1078	-7.6
8	82	3	12.0	0.0366	0.1465	-9.0
9	76	4	15.8	0.0526	0.2076	-11.8
10	74	20	27.2	0.2703	0.3673	-7.2

surgical_group — chopcluster_6

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	55	0	1.2	0.0000	0.0215	-1.2
2	50	1	2.0	0.0200	0.0406	-1.0
3	52	0	3.0	0.0000	0.0573	-3.0
4	59	1	4.7	0.0169	0.0800	-3.7
5	49	3	5.4	0.0612	0.1108	-2.4
6	50	2	7.3	0.0400	0.1467	-5.3
7	58	3	11.5	0.0517	0.1982	-8.5
8	50	8	12.9	0.1600	0.2570	-4.9
9	54	16	19.5	0.2963	0.3618	-3.5
10	53	33	31.3	0.6226	0.5904	+1.7

surgical_group — chopcluster_7

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	32	0	1.1	0.0000	0.0344	-1.1
2	30	0	2.0	0.0000	0.0657	-2.0
3	31	0	2.9	0.0000	0.0929	-2.9
4	27	0	3.2	0.0000	0.1204	-3.2
5	31	1	4.8	0.0323	0.1548	-3.8
6	30	4	6.1	0.1333	0.2025	-2.1
7	28	4	7.0	0.1429	0.2507	-3.0
8	31	6	9.4	0.1935	0.3029	-3.4
9	29	5	10.9	0.1724	0.3755	-5.9
10	29	16	16.4	0.5517	0.5657	-0.4

surgical_group — chopcluster_8

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	30	0	0.4	0.0000	0.0127	-0.4
2	34	1	1.1	0.0294	0.0334	-0.1
3	30	2	1.5	0.0667	0.0498	+0.5
4	28	0	1.8	0.0000	0.0646	-1.8
5	37	0	3.2	0.0000	0.0857	-3.2
6	26	2	3.0	0.0769	0.1156	-1.0
7	33	2	5.2	0.0606	0.1562	-3.2
8	29	0	6.3	0.0000	0.2164	-6.3
9	29	6	8.8	0.2069	0.3021	-2.8
10	30	10	14.9	0.3333	0.4953	-4.9

surgical_group — none

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	98	0	0.6	0.0000	0.0061	-0.6
2	71	0	1.2	0.0000	0.0175	-1.2
3	75	0	2.1	0.0000	0.0277	-2.1
4	105	0	4.3	0.0000	0.0408	-4.3
5	88	1	5.5	0.0114	0.0622	-4.5
6	72	0	6.1	0.0000	0.0849	-6.1
7	96	5	11.2	0.0521	0.1168	-6.2
8	85	4	13.8	0.0471	0.1625	-9.8
9	83	10	19.3	0.1205	0.2326	-9.3
10	84	24	35.1	0.2857	0.4176	-11.1

emergency — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

emergency — elective

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	819	2	5.3	0.0024	0.0064	-3.3
2	1219	6	24.5	0.0049	0.0201	-18.5
3	890	5	28.9	0.0056	0.0325	-23.9
4	772	7	32.8	0.0091	0.0424	-25.8
5	976	9	53.7	0.0092	0.0550	-44.7
6	813	10	56.6	0.0123	0.0696	-46.6
7	993	13	88.4	0.0131	0.0890	-75.4
8	884	21	105.2	0.0238	0.1190	-84.2
9	952	44	163.2	0.0462	0.1714	-119.2
10	932	199	301.3	0.2135	0.3233	-102.3

emergency — emergency

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	148	1	6.5	0.0068	0.0438	-5.5
2	132	1	10.0	0.0076	0.0759	-9.0
3	148	1	14.8	0.0068	0.0998	-13.8
4	167	11	21.0	0.0659	0.1259	-10.0
5	153	7	24.1	0.0458	0.1577	-17.1
6	140	7	26.6	0.0500	0.1901	-19.6
7	137	10	30.5	0.0730	0.2226	-20.5
8	156	25	43.7	0.1603	0.2802	-18.7
9	144	37	53.9	0.2569	0.3744	-16.9
10	144	88	81.9	0.6111	0.5689	+6.1

sex × age_group — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

sex × age_group — female / 0-17

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	20	0	0.5	0.0000	0.0273	-0.5
2	15	0	0.7	0.0000	0.0487	-0.7
3	20	0	1.4	0.0000	0.0700	-1.4
4	18	0	1.6	0.0000	0.0894	-1.6
5	16	0	1.7	0.0000	0.1047	-1.7
6	18	0	2.4	0.0000	0.1317	-2.4
7	17	2	2.8	0.1176	0.1638	-0.8
8	17	0	3.3	0.0000	0.1935	-3.3
9	17	5	4.0	0.2941	0.2374	+1.0
10	17	6	6.1	0.3529	0.3600	-0.1

sex × age_group — female / 18-44

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	41	0	0.4	0.0000	0.0091	-0.4
2	35	0	0.9	0.0000	0.0254	-0.9
3	43	0	1.6	0.0000	0.0374	-1.6
4	51	0	2.6	0.0000	0.0501	-2.6
5	37	0	2.3	0.0000	0.0615	-2.3
6	40	2	3.1	0.0500	0.0773	-1.1
7	38	0	3.7	0.0000	0.0968	-3.7
8	44	1	5.6	0.0227	0.1275	-4.6
9	40	2	8.1	0.0500	0.2019	-6.1
10	39	8	14.6	0.2051	0.3755	-6.6

sex × age_group — female / 45-64

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	122	1	0.8	0.0082	0.0064	+0.2
2	111	1	1.9	0.0090	0.0173	-0.9
3	107	1	2.9	0.0093	0.0273	-1.9
4	138	2	5.5	0.0145	0.0400	-3.5
5	101	0	5.5	0.0000	0.0550	-5.5
6	131	2	9.5	0.0153	0.0726	-7.5
7	116	1	11.3	0.0086	0.0970	-10.3
8	122	5	16.3	0.0410	0.1338	-11.3
9	113	3	22.5	0.0265	0.1987	-19.5
10	118	36	46.7	0.3051	0.3954	-10.7

sex × age_group — female / 65-79

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	170	1	1.6	0.0059	0.0094	-0.6
2	115	0	2.6	0.0000	0.0226	-2.6
3	176	0	6.1	0.0000	0.0344	-6.1
4	149	1	7.4	0.0067	0.0497	-6.4
5	132	4	8.6	0.0303	0.0648	-4.6
6	130	0	10.8	0.0000	0.0832	-10.8
7	159	4	17.3	0.0252	0.1089	-13.3
8	145	11	22.2	0.0759	0.1531	-11.2
9	147	14	32.8	0.0952	0.2235	-18.8
10	147	58	59.8	0.3946	0.4070	-1.8

sex × age_group — female / 80+

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	48	0	0.6	0.0000	0.0128	-0.6
2	49	1	1.4	0.0204	0.0294	-0.4
3	32	0	1.4	0.0000	0.0428	-1.4
4	58	2	3.2	0.0345	0.0544	-1.2
5	40	0	2.8	0.0000	0.0704	-2.8
6	51	0	4.8	0.0000	0.0949	-4.8
7	43	0	5.3	0.0000	0.1222	-5.3
8	46	1	7.7	0.0217	0.1671	-6.7
9	43	8	11.0	0.1860	0.2560	-3.0
10	47	7	18.9	0.1489	0.4030	-11.9

sex × age_group — male / 0-17

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	21	0	0.6	0.0000	0.0262	-0.6
2	17	1	0.8	0.0588	0.0485	+0.2
3	18	0	1.1	0.0000	0.0622	-1.1
4	26	2	2.0	0.0769	0.0771	-0.0
5	12	0	1.1	0.0000	0.0950	-1.1
6	26	1	2.9	0.0385	0.1123	-1.9
7	16	0	2.1	0.0000	0.1313	-2.1
8	18	1	2.8	0.0556	0.1581	-1.8
9	20	0	4.0	0.0000	0.2010	-4.0
10	18	3	5.8	0.1667	0.3239	-2.8

sex × age_group — male / 18-44

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	53	0	0.3	0.0000	0.0060	-0.3
2	49	0	0.8	0.0000	0.0172	-0.8
3	72	0	2.2	0.0000	0.0307	-2.2
4	46	0	2.0	0.0000	0.0427	-2.0
5	42	1	2.2	0.0238	0.0529	-1.2
6	55	2	3.7	0.0364	0.0675	-1.7
7	57	2	5.1	0.0351	0.0900	-3.1
8	54	3	7.0	0.0556	0.1294	-4.0
9	51	3	9.7	0.0588	0.1902	-6.7
10	54	10	19.0	0.1852	0.3510	-9.0

sex × age_group — male / 45-64

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	226	0	1.4	0.0000	0.0064	-1.4
2	304	0	6.2	0.0000	0.0203	-6.2
3	224	2	7.3	0.0089	0.0324	-5.3
4	280	0	12.6	0.0000	0.0449	-12.6
5	237	0	14.3	0.0000	0.0602	-14.3
6	247	2	18.9	0.0081	0.0765	-16.9
7	260	5	26.3	0.0192	0.1013	-21.3
8	260	11	36.0	0.0423	0.1383	-25.0
9	241	13	47.6	0.0539	0.1974	-34.6
10	257	62	93.5	0.2412	0.3636	-31.5

sex × age_group — male / 65-79

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	354	1	3.3	0.0028	0.0093	-2.3
2	265	2	6.0	0.0075	0.0225	-4.0
3	283	3	9.2	0.0106	0.0324	-6.2
4	349	4	15.5	0.0115	0.0443	-11.5
5	387	4	24.0	0.0103	0.0621	-20.0
6	328	1	27.1	0.0030	0.0827	-26.1
7	268	5	28.1	0.0187	0.1048	-23.1
8	334	10	46.3	0.0299	0.1387	-36.3
9	324	19	64.6	0.0586	0.1994	-45.6
10	322	107	121.6	0.3323	0.3777	-14.6

sex × age_group — male / 80+

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	50	0	0.7	0.0000	0.0135	-0.7
2	56	1	1.7	0.0179	0.0305	-0.7
3	69	1	3.3	0.0145	0.0471	-2.3
4	59	1	3.8	0.0169	0.0649	-2.8
5	44	2	3.5	0.0455	0.0803	-1.5
6	54	0	5.3	0.0000	0.0988	-5.3
7	60	3	7.9	0.0500	0.1323	-4.9
8	51	7	9.4	0.1373	0.1845	-2.4
9	56	7	13.6	0.1250	0.2421	-6.6
10	56	12	22.0	0.2143	0.3931	-10.0

sex × frailty — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

sex × frailty — female / frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	165	1	2.7	0.0061	0.0162	-1.7
2	162	1	5.6	0.0062	0.0348	-4.6
3	170	1	8.5	0.0059	0.0500	-7.5
4	180	5	11.9	0.0278	0.0663	-6.9
5	146	0	12.7	0.0000	0.0868	-12.7
6	194	3	21.6	0.0155	0.1116	-18.6
7	142	5	20.7	0.0352	0.1461	-15.7
8	175	6	33.9	0.0343	0.1935	-27.9
9	167	40	47.0	0.2395	0.2812	-7.0
10	165	67	78.2	0.4061	0.4736	-11.2

sex × frailty — female / non_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	38	0	0.1	0.0000	0.0028	-0.1
2	69	0	0.9	0.0000	0.0125	-0.9
3	58	0	1.3	0.0000	0.0224	-1.3
4	52	0	1.7	0.0000	0.0329	-1.7
5	27	0	1.1	0.0000	0.0400	-1.1
6	73	0	3.6	0.0000	0.0488	-3.6
7	48	0	3.0	0.0000	0.0633	-3.0
8	55	0	4.7	0.0000	0.0862	-4.7
9	50	0	6.0	0.0000	0.1193	-6.0
10	52	1	11.9	0.0192	0.2287	-10.9

sex × frailty — female / pre_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	157	1	1.0	0.0064	0.0063	+0.0
2	144	1	2.5	0.0069	0.0170	-1.5
3	157	1	4.3	0.0064	0.0274	-3.3
4	167	3	6.7	0.0180	0.0399	-3.7
5	138	1	7.6	0.0072	0.0550	-6.6
6	129	2	9.1	0.0155	0.0703	-7.1
7	156	2	13.8	0.0128	0.0886	-11.8
8	153	9	18.6	0.0588	0.1218	-9.6
9	154	12	27.9	0.0779	0.1812	-15.9
10	146	28	48.2	0.1918	0.3299	-20.2

sex × frailty — male / frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	480	1	6.1	0.0021	0.0128	-5.1
2	350	4	9.6	0.0114	0.0276	-5.6
3	492	5	19.5	0.0102	0.0396	-14.5
4	409	3	22.5	0.0073	0.0550	-19.5
5	504	3	36.1	0.0060	0.0717	-33.1
6	463	8	43.2	0.0173	0.0933	-35.2
7	423	16	51.3	0.0378	0.1214	-35.3
8	451	23	72.8	0.0510	0.1615	-49.8
9	457	56	104.7	0.1225	0.2290	-48.7
10	444	152	182.2	0.3423	0.4105	-30.2

sex × frailty — male / non_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	56	0	0.2	0.0000	0.0032	-0.2
2	40	0	0.4	0.0000	0.0100	-0.4
3	65	0	1.1	0.0000	0.0174	-1.1
4	29	0	0.7	0.0000	0.0250	-0.7
5	57	0	1.9	0.0000	0.0326	-1.9
6	45	0	1.9	0.0000	0.0424	-1.9
7	57	1	3.1	0.0175	0.0544	-2.1
8	49	0	3.5	0.0000	0.0712	-3.5
9	50	2	5.2	0.0400	0.1045	-3.2
10	52	2	10.0	0.0385	0.1920	-8.0

sex × frailty — male / pre_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	225	0	1.4	0.0000	0.0061	-1.4
2	189	0	3.3	0.0000	0.0175	-3.3
3	197	1	5.4	0.0051	0.0277	-4.4
4	181	2	6.8	0.0110	0.0376	-4.8
5	221	2	11.0	0.0090	0.0496	-9.0
6	234	3	15.7	0.0128	0.0670	-12.7
7	178	2	15.5	0.0112	0.0872	-13.5
8	232	3	26.9	0.0129	0.1159	-23.9
9	196	4	32.9	0.0204	0.1678	-28.9
10	204	21	61.2	0.1029	0.3000	-40.2

age_group × frailty — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	1227	4	11.4	0.0033	0.0093	-7.4
2	833	4	18.9	0.0048	0.0227	-14.9
3	1322	10	46.0	0.0076	0.0348	-36.0
4	741	6	35.1	0.0081	0.0473	-29.1
5	1325	13	82.4	0.0098	0.0622	-69.4
6	1024	13	84.4	0.0127	0.0824	-71.4
7	1050	20	112.3	0.0190	0.1069	-92.3
8	1010	42	144.2	0.0416	0.1428	-102.2
9	1120	78	231.0	0.0696	0.2063	-153.0
10	1067	314	407.3	0.2943	0.3818	-93.3

age_group × frailty — 0-17 / frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	23	0	0.9	0.0000	0.0370	-0.9
2	20	0	1.3	0.0000	0.0638	-1.3
3	18	2	1.5	0.1111	0.0831	+0.5
4	23	0	2.3	0.0000	0.1020	-2.3
5	26	1	3.3	0.0385	0.1256	-2.3
6	17	1	2.5	0.0588	0.1488	-1.5
7	22	1	4.0	0.0455	0.1798	-3.0
8	18	2	3.8	0.1111	0.2106	-1.8
9	23	5	5.8	0.2174	0.2502	-0.8
10	21	7	8.2	0.3333	0.3893	-1.2

age_group × frailty — 0-17 / non_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	4	0	0.1	0.0000	0.0163	-0.1
2	5	0	0.1	0.0000	0.0270	-0.1
3	6	0	0.2	0.0000	0.0375	-0.2
4	6	0	0.3	0.0000	0.0525	-0.3
5	4	0	0.3	0.0000	0.0675	-0.3
6	3	0	0.2	0.0000	0.0767	-0.2
7	4	0	0.3	0.0000	0.0862	-0.3
8	6	0	0.6	0.0000	0.1017	-0.6
9	3	0	0.4	0.0000	0.1217	-0.4
10	4	0	0.7	0.0000	0.1713	-0.7

age_group × frailty — 0-17 / pre_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	14	0	0.3	0.0000	0.0232	-0.3
2	9	1	0.4	0.1111	0.0450	+0.6
3	11	0	0.6	0.0000	0.0582	-0.6
4	9	0	0.6	0.0000	0.0678	-0.6
5	13	0	1.0	0.0000	0.0800	-1.0
6	9	0	0.8	0.0000	0.0928	-0.8
7	12	0	1.3	0.0000	0.1121	-1.3
8	12	0	1.6	0.0000	0.1321	-1.6
9	12	1	2.0	0.0833	0.1696	-1.0
10	10	0	2.4	0.0000	0.2415	-2.4

age_group × frailty — 18-44 / frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	33	0	0.7	0.0000	0.0198	-0.7
2	24	0	0.9	0.0000	0.0377	-0.9
3	30	0	1.5	0.0000	0.0498	-1.5
4	32	1	2.1	0.0312	0.0669	-1.1
5	33	1	3.0	0.0303	0.0898	-2.0
6	25	0	2.9	0.0000	0.1172	-2.9
7	33	2	5.0	0.0606	0.1523	-3.0
8	30	2	6.4	0.0667	0.2133	-4.4
9	30	5	9.2	0.1667	0.3078	-4.2
10	30	10	14.1	0.3333	0.4685	-4.1

age_group × frailty — 18-44 / non_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	25	0	0.1	0.0000	0.0034	-0.1
2	41	0	0.5	0.0000	0.0122	-0.5
3	34	0	0.8	0.0000	0.0225	-0.8
4	40	0	1.3	0.0000	0.0335	-1.3
5	18	0	0.7	0.0000	0.0400	-0.7
6	34	0	1.6	0.0000	0.0465	-1.6
7	37	0	2.1	0.0000	0.0572	-2.1
8	37	0	2.7	0.0000	0.0738	-2.7
9	33	2	3.6	0.0606	0.1102	-1.6
10	32	0	7.1	0.0000	0.2220	-7.1

age_group × frailty — 18-44 / pre_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	28	0	0.2	0.0000	0.0059	-0.2
2	32	0	0.6	0.0000	0.0189	-0.6
3	28	0	0.9	0.0000	0.0325	-0.9
4	44	1	2.1	0.0227	0.0480	-1.1
5	26	1	1.6	0.0385	0.0631	-0.6
6	28	2	2.2	0.0714	0.0770	-0.2
7	33	0	3.1	0.0000	0.0944	-3.1
8	30	2	3.6	0.0667	0.1212	-1.6
9	31	2	5.3	0.0645	0.1697	-3.3
10	30	3	8.9	0.1000	0.2958	-5.9

age_group × frailty — 45-64 / frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	174	1	1.7	0.0057	0.0100	-0.7
2	202	1	5.2	0.0050	0.0255	-4.2
3	211	0	8.3	0.0000	0.0394	-8.3
4	185	0	10.1	0.0000	0.0547	-10.1
5	232	2	16.6	0.0086	0.0715	-14.6
6	182	1	17.1	0.0055	0.0938	-16.1
7	190	8	23.1	0.0421	0.1218	-15.1
8	206	11	33.7	0.0534	0.1636	-22.7
9	194	19	46.2	0.0979	0.2379	-27.2
10	191	68	84.2	0.3560	0.4410	-16.2

age_group × frailty — 45-64 / non_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	45	0	0.1	0.0000	0.0029	-0.1
2	36	0	0.4	0.0000	0.0100	-0.4
3	34	0	0.5	0.0000	0.0150	-0.5
4	48	0	1.1	0.0000	0.0225	-1.1
5	41	0	1.3	0.0000	0.0323	-1.3
6	41	0	1.7	0.0000	0.0426	-1.7
7	37	0	2.1	0.0000	0.0557	-2.1
8	44	0	3.2	0.0000	0.0720	-3.2
9	44	0	4.9	0.0000	0.1115	-4.9
10	40	3	8.7	0.0750	0.2178	-5.7

age_group × frailty — 45-64 / pre_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	152	0	0.9	0.0000	0.0058	-0.9
2	128	1	2.2	0.0078	0.0175	-1.2
3	144	1	3.9	0.0069	0.0273	-2.9
4	105	3	3.9	0.0286	0.0374	-0.9
5	148	0	7.3	0.0000	0.0493	-7.3
6	139	0	9.4	0.0000	0.0678	-9.4
7	111	3	9.7	0.0270	0.0869	-6.7
8	144	5	16.6	0.0347	0.1150	-11.6
9	139	1	23.8	0.0072	0.1715	-22.8
10	128	19	38.9	0.1484	0.3038	-19.9

age_group × frailty — 65-79 / frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	322	1	4.2	0.0031	0.0131	-3.2
2	243	3	6.7	0.0123	0.0277	-3.7
3	343	4	13.6	0.0117	0.0396	-9.6
4	283	3	15.6	0.0106	0.0552	-12.6
5	318	3	22.9	0.0094	0.0720	-19.9
6	289	3	26.4	0.0104	0.0912	-23.4
7	323	9	38.3	0.0279	0.1185	-29.3
8	305	14	49.3	0.0459	0.1618	-35.3
9	296	35	67.6	0.1182	0.2285	-32.6
10	310	129	128.5	0.4161	0.4145	+0.5

age_group × frailty — 65-79 / non_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	23	0	0.1	0.0000	0.0028	-0.1
2	27	0	0.3	0.0000	0.0126	-0.3
3	16	0	0.3	0.0000	0.0200	-0.3
4	24	0	0.7	0.0000	0.0277	-0.7
5	19	0	0.7	0.0000	0.0374	-0.7
6	27	1	1.3	0.0370	0.0476	-0.3
7	18	0	1.1	0.0000	0.0594	-1.1
8	21	0	1.8	0.0000	0.0855	-1.8
9	23	0	2.7	0.0000	0.1191	-2.7
10	20	0	4.0	0.0000	0.2022	-4.0

age_group × frailty — 65-79 / pre_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	173	1	1.1	0.0058	0.0064	-0.1
2	83	0	1.2	0.0000	0.0150	-1.2
3	208	0	5.2	0.0000	0.0252	-5.2
4	130	1	4.9	0.0077	0.0378	-3.9
5	134	1	6.7	0.0075	0.0499	-5.7
6	126	2	8.1	0.0159	0.0646	-6.1
7	152	0	12.8	0.0000	0.0841	-12.8
8	142	4	16.6	0.0282	0.1170	-12.6
9	146	11	25.2	0.0753	0.1727	-14.2
10	140	24	46.8	0.1714	0.3340	-22.8

age_group × frailty — 80+ / frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	65	0	1.3	0.0000	0.0204	-1.3
2	60	2	2.4	0.0333	0.0405	-0.4
3	72	2	4.0	0.0278	0.0551	-2.0
4	50	0	3.5	0.0000	0.0698	-3.5
5	71	1	6.4	0.0141	0.0902	-5.4
6	57	0	6.5	0.0000	0.1146	-6.5
7	70	4	10.8	0.0571	0.1538	-6.8
8	61	10	12.9	0.1639	0.2109	-2.9
9	61	14	17.4	0.2295	0.2848	-3.4
10	62	12	27.2	0.1935	0.4394	-15.2

age_group × frailty — 80+ / non_frail

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	3	0	0.0	0.0000	0.0083	-0.0
2	1	0	0.0	0.0000	0.0150	-0.0
3	2	0	0.0	0.0000	0.0200	-0.0
4	2	0	0.1	0.0000	0.0250	-0.1
5	3	0	0.2	0.0000	0.0633	-0.2
6	1	0	0.1	0.0000	0.0750	-0.1
7	2	0	0.2	0.0000	0.1050	-0.2
8	2	0	0.3	0.0000	0.1550	-0.3
9	1	0	0.2	0.0000	0.2350	-0.2
10	1	0	0.3	0.0000	0.2600	-0.3

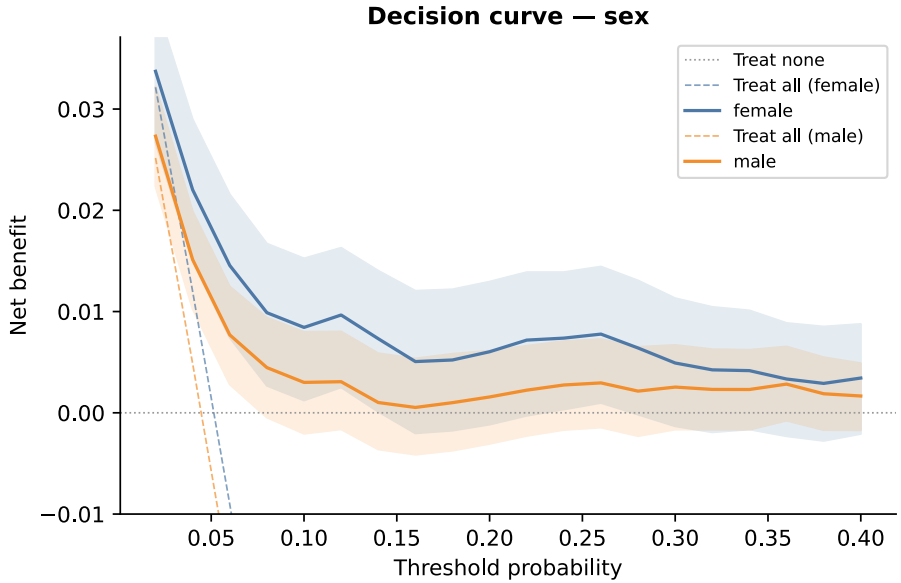
age_group × frailty — 80+ / pre_frail

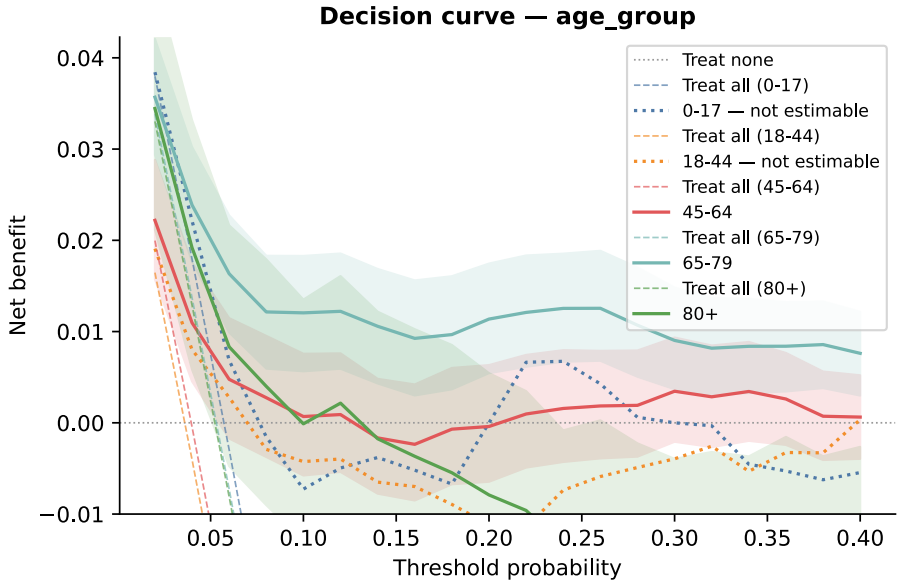
Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	42	0	0.4	0.0000	0.0100	-0.4
2	34	1	0.8	0.0294	0.0229	+0.2
3	39	0	1.3	0.0000	0.0323	-1.3
4	38	0	1.7	0.0000	0.0455	-1.7
5	28	1	1.6	0.0357	0.0579	-0.6
6	35	0	2.5	0.0000	0.0711	-2.5
7	41	1	3.8	0.0244	0.0924	-2.8
8	36	0	4.5	0.0000	0.1257	-4.5
9	37	2	7.1	0.0541	0.1909	-5.1
10	35	3	10.9	0.0857	0.3113	-7.9

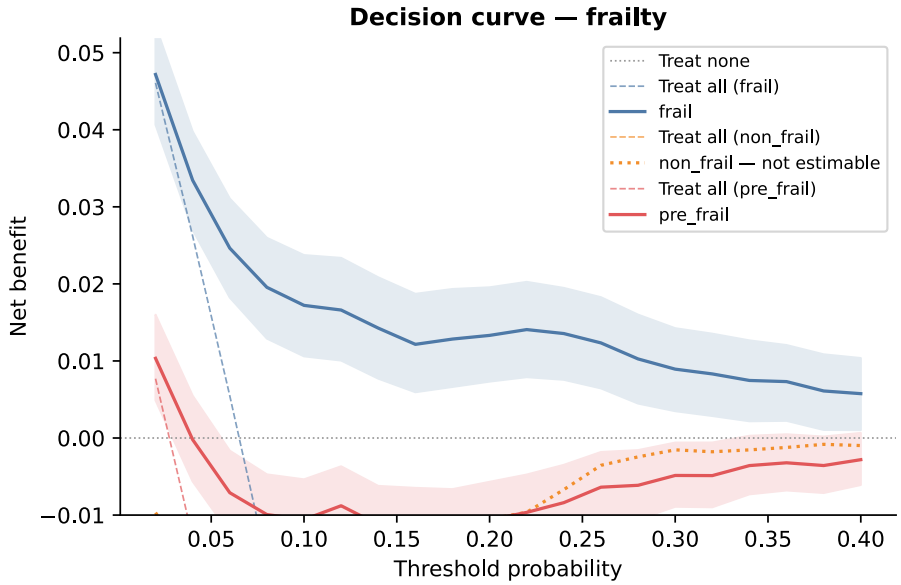
Clinical utility (decision curves)

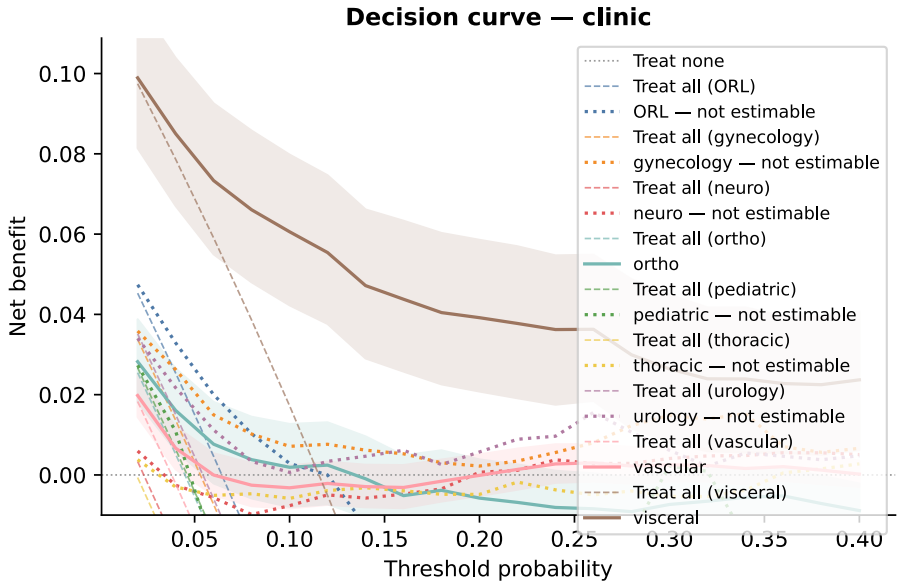
Clinical utility gap statistics

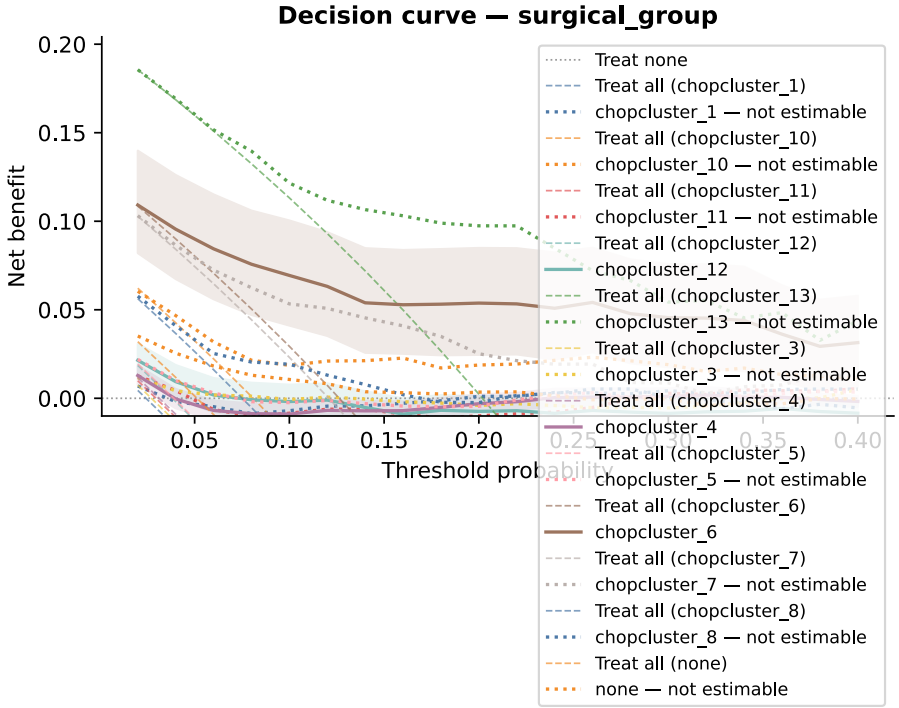
Attribute	Metric	Value (95% CI)
sex	max_net_benefit_gap	0.0069 (0.0030–0.0164)
age_group	max_net_benefit_gap	0.0260
frailty	max_net_benefit_gap	0.0569
clinic	max_net_benefit_gap	0.0951
surgical_group	max_net_benefit_gap	0.1787
emergency	max_net_benefit_gap	0.0936 (0.0761–0.1116)
sex × age_group	max_net_benefit_gap	0.0523
sex × frailty	max_net_benefit_gap	0.0733
age_group × frailty	max_net_benefit_gap	0.0911

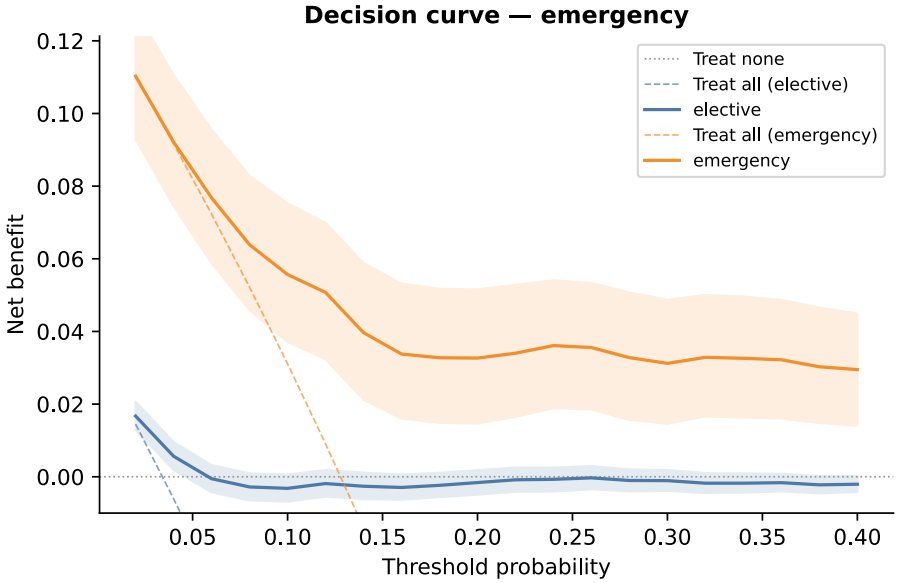


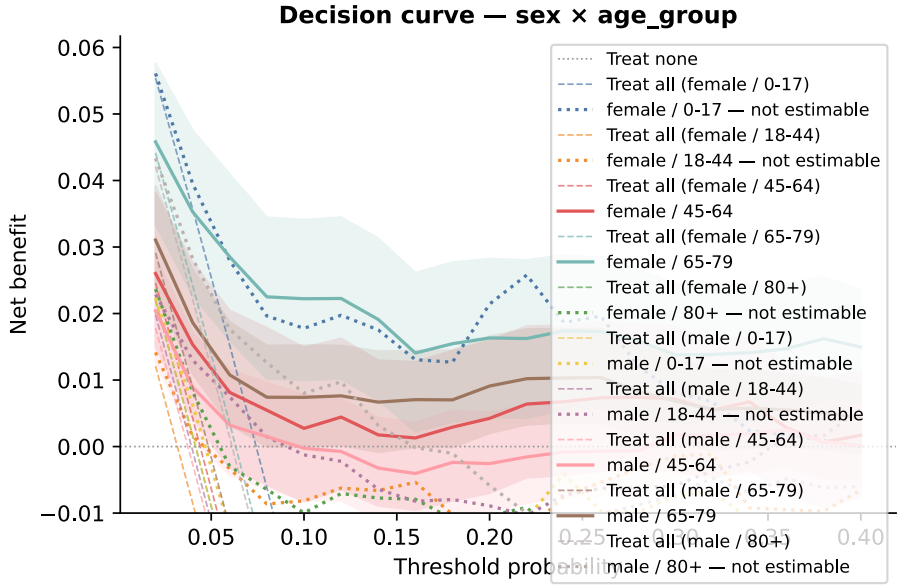


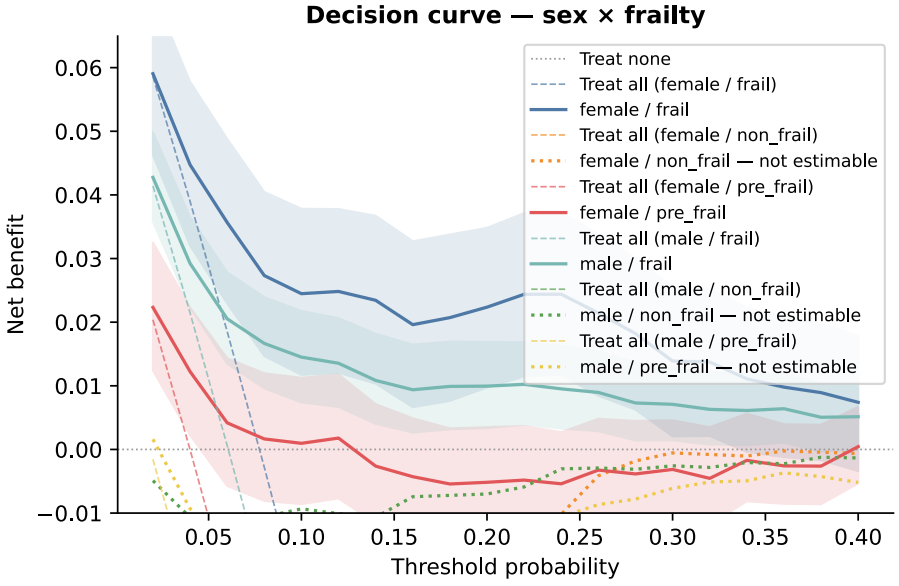

Decision curve — frailty
Net benefit
Threshold probability
Treat none
Treat all (frail)
frail
Treat all (non_frail)
non_frail — not estimable
Treat all (pre_frail)
pre_frail

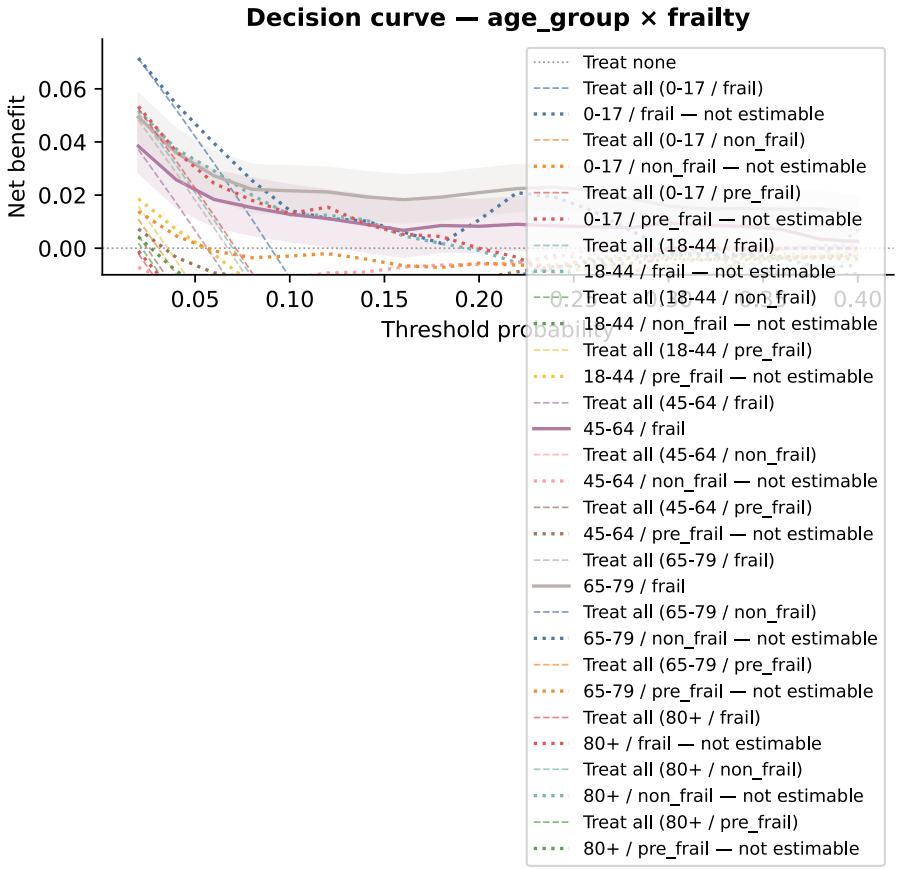

Decision curve — clinic
Net benefit
Threshold probability
Treat none
Treat all (ORL)
ORL — not estimable
Treat all (gynecology)
gynecology — not estimable
Treat all (neuro)
neuro — not estimable
Treat all (ortho)
ortho
Treat all (pediatric)
pediatric — not estimable
Treat all (thoracic)
thoracic — not estimable
Treat all (urology)
urology — not estimable
Treat all (vascular)
vascular
Treat all (visceral)
visceral


Decision curve — surgical_group
Net benefit
Threshold probability
0.20
0.15
0.10
0.05
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Treat none
Treat all (chopcluster_1)
chopcluster_1 — not estimable
Treat all (chopcluster_10)
chopcluster_10 — not estimable
Treat all (chopcluster_11)
chopcluster_11 — not estimable
Treat all (chopcluster_12)
chopcluster_12
Treat all (chopcluster_13)
chopcluster_13 — not estimable
Treat all (chopcluster_3)
chopcluster_3 — not estimable
Treat all (chopcluster_4)
chopcluster_4
Treat all (chopcluster_5)
chopcluster_5 — not estimable
Treat all (chopcluster_6)
chopcluster_6
Treat all (chopcluster_7)
chopcluster_7 — not estimable
Treat all (chopcluster_8)
chopcluster_8 — not estimable
Treat all (none)
none — not estimable


Decision curve — emergency
Net benefit
Threshold probability
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Treat none
Treat all (elective)
elective
Treat all (emergency)
emergency


Decision curve — sex × age_group
Net benefit
Threshold probability
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Treat none
Treat all (female / 0-17)
female / 0-17 — not estimable
Treat all (female / 18-44)
female / 18-44 — not estimable
Treat all (female / 45-64)
female / 45-64
Treat all (female / 65-79)
female / 65-79
Treat all (female / 80+)
female / 80+ — not estimable
Treat all (male / 0-17)
male / 0-17 — not estimable
Treat all (male / 18-44)
male / 18-44 — not estimable
Treat all (male / 45-64)
male / 45-64
Treat all (male / 65-79)
male / 65-79
Treat all (male / 80+)
male / 80+ — not estimable


Decision curve — sex × frailty
Net benefit
Threshold probability
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Treat none
Treat all (female / frail)
female / frail
Treat all (female / non_frail)
female / non_frail — not estimable
Treat all (female / pre_frail)
female / pre_frail
Treat all (male / frail)
male / frail
Treat all (male / non_frail)
male / non_frail — not estimable
Treat all (male / pre_frail)
male / pre_frail — not estimable


Decision curve — age_group × frailty
Net benefit
0.06
0.04
0.02
0.00
0.05
0.10
0.15
0.20
0.40
Threshold probability
Treat none
Treat all (0-17 / frail)
0-17 / frail — not estimable
Treat all (0-17 / non_frail)
0-17 / non_frail — not estimable
Treat all (0-17 / pre_frail)
0-17 / pre_frail — not estimable
Treat all (18-44 / frail)
18-44 / frail — not estimable
Treat all (18-44 / non_frail)
18-44 / non_frail — not estimable
Treat all (18-44 / pre_frail)
18-44 / pre_frail — not estimable
Treat all (45-64 / frail)
45-64 / frail
Treat all (45-64 / non_frail)
45-64 / non_frail — not estimable
Treat all (45-64 / pre_frail)
45-64 / pre_frail — not estimable
Treat all (65-79 / frail)
65-79 / frail
Treat all (65-79 / non_frail)
65-79 / non_frail — not estimable
Treat all (65-79 / pre_frail)
65-79 / pre_frail — not estimable
Treat all (80+ / frail)
80+ / frail — not estimable
Treat all (80+ / non_frail)
80+ / non_frail — not estimable
Treat all (80+ / pre_frail)
80+ / pre_frail — not estimable

Interaction cell diagnostics

Attribute	Subgroup	n	Events	Shrunk	Marginal rate	Cell rate	Shrinkage factor
sex × age_group	*female / 0-17*	*175*	*13*	*Yes*	*0.0515*	*0.0743*	*0.5020*
sex × age_group	*female / 18-44*	*408*	*13*	*Yes*	*0.0515*	*0.0319*	*0.3018*
sex × age_group	female / 45-64	1179	52	No	0.0515	0.0441	0.1301
sex × age_group	female / 65-79	1470	93	No	0.0515	0.0633	0.1071
sex × age_group	*female / 80+*	*457*	*19*	*Yes*	*0.0515*	*0.0416*	*0.2785*
sex × age_group	*male / 0-17*	*192*	*8*	*Yes*	*0.0447*	*0.0417*	*0.4788*
sex × age_group	*male / 18-44*	*533*	*21*	*Yes*	*0.0447*	*0.0394*	*0.2487*
sex × age_group	male / 45-64	2536	95	No	0.0447	0.0375	0.0650
sex × age_group	male / 65-79	3214	156	No	0.0447	0.0485	0.0520
sex × age_group	*male / 80+*	*555*	*34*	*Yes*	*0.0447*	*0.0613*	*0.2412*
sex × frailty	female / frail	1666	129	No	0.0515	0.0774	0.1119
sex × frailty	*female / non_frail*	*522*	*1*	*Yes*	*0.0515*	*0.0019*	*0.2869*
sex × frailty	female / pre_frail	1501	60	No	0.0515	0.0400	0.1227
sex × frailty	male / frail	4473	271	No	0.0447	0.0606	0.0448
sex × frailty	*male / non_frail*	*500*	*5*	*Yes*	*0.0447*	*0.0100*	*0.2958*
sex × frailty	*male / pre_frail*	*2057*	*38*	*Yes*	*0.0447*	*0.0185*	*0.0926*
age_group × frailty	*0-17 / frail*	*211*	*19*	*Yes*	*0.0572*	*0.0900*	*0.3891*
age_group × frailty	*0-17 / non_frail*	*45*	*0*	*Yes*	*0.0572*	*0.0000*	*0.7492*
age_group × frailty	*0-17 / pre_frail*	*111*	*2*	*Yes*	*0.0572*	*0.0180*	*0.5477*
age_group × frailty	*18-44 / frail*	*300*	*21*	*Yes*	*0.0361*	*0.0700*	*0.3094*
age_group × frailty	*18-44 / non_frail*	*331*	*2*	*Yes*	*0.0361*	*0.0060*	*0.2888*
age_group × frailty	*18-44 / pre_frail*	*310*	*11*	*Yes*	*0.0361*	*0.0355*	*0.3024*
age_group × frailty	45-64 / frail	1967	111	No	0.0396	0.0564	0.0640
age_group × frailty	*45-64 / non_frail*	*410*	*3*	*Yes*	*0.0396*	*0.0073*	*0.2469*
age_group × frailty	*45-64 / pre_frail*	*1338*	*33*	*Yes*	*0.0396*	*0.0247*	*0.0913*
age_group × frailty	65-79 / frail	3032	204	No	0.0532	0.0673	0.0424
age_group × frailty	*65-79 / non_frail*	*218*	*1*	*Yes*	*0.0532*	*0.0046*	*0.3814*
age_group × frailty	*65-79 / pre_frail*	*1434*	*44*	*Yes*	*0.0532*	*0.0307*	*0.0857*
age_group × frailty	*80+ / frail*	*629*	*45*	*Yes*	*0.0524*	*0.0715*	*0.1761*
age_group × frailty	*80+ / non_frail*	*18*	*0*	*Yes*	*0.0524*	*0.0000*	*0.8819*
age_group × frailty	*80+ / pre_frail*	*365*	*8*	*Yes*	*0.0524*	*0.0219*	*0.2691*

Appendix: TRIPOD+AI item coverage

Fairness-relevant items from the TRIPOD+AI checklist (Collins et al., 2024, *BMJ*). Items addressed by this audit are linked to the relevant report section. Items marked "User responsibility" require information that the audit tool cannot produce automatically.

Item	Description	Status	Detail
3c	Describe any known health inequalities between sociodemographic groups	*User responsibility*	User must provide context in manuscript introduction
5a	Describe data sources, rationale, and representativeness	*User responsibility*	User must describe data representativeness in manuscript methods
7	Describe data pre-processing, including whether similar across sociodemographic groups	*User responsibility*	User must verify preprocessing does not vary by group
8a	Outcome assessment consistency across sociodemographic groups	*User responsibility*	User must verify outcome labels are measured consistently
8b	Assessor demographics for outcome assessment	*User responsibility*	User must report assessor demographics
9c	Assessor demographics for predictor measurement	*User responsibility*	User must report assessor demographics
12e	All measures used to evaluate model performance	**Addressed**	Discrimination, Calibration, and Clinical utility sections
12f	Describe any model updating (eg recalibration) for sociodemographic groups or settings	Not included	No mitigation requested
14	Fairness approaches and rationale	**Addressed**	Audit summary and gap statistics throughout
20b	Report demographic characteristics and group differences	**Addressed**	Discrimination section (subgroup sizes, prevalences)
20c	Compare predictor distributions including demographics	*User responsibility*	User must report predictor distributions per group in manuscript
23a	Report performance estimates with CIs, including for key sociodemographic subgroups	**Addressed**	Discrimination, Calibration, and Clinical utility sections
25	Overall interpretation including fairness issues	*User responsibility*	User must interpret results in manuscript discussion
26	Limitations including non-representative sample, bias effects	*User responsibility*	User must discuss limitations in manuscript

Source: 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. doi: 10.1136/bmj-2023-078378. This checklist covers fairness-relevant items only; see the full statement for the complete checklist.

Appendix: How to read this report

Discrimination

Discrimination measures how well the model separates positive from negative cases.

AUROC (Area Under the ROC Curve)
Probability that the model ranks a random positive case higher than a random negative case. Range [0, 1]; 0.5 = no discrimination; 1.0 = perfect separation.

AUPRC (Area Under the Precision-Recall Curve)
Summarises precision across all recall levels. More informative than AUROC for low-prevalence outcomes. Higher is better; baseline equals the prevalence.

Sensitivity (True Positive Rate)
Proportion of true positive cases correctly identified at the chosen threshold. = TP / (TP + FN).

Specificity (True Negative Rate)
Proportion of true negative cases correctly identified at the chosen threshold. = TN / (TN + FP).

PPV (Positive Predictive Value)
Proportion of positive predictions that are truly positive. = TP / (TP + FP).

NPV (Negative Predictive Value)
Proportion of negative predictions that are truly negative. = TN / (TN + FN).

Calibration

Calibration measures whether predicted probabilities match observed event rates.

Calibration intercept
Intercept from logistic recalibration with slope fixed to 1 (Steyerberg 2009). Perfect calibration = 0. Positive values indicate the model underestimates risk on average; negative values indicate overestimation.

Calibration slope
Slope from logistic recalibration with free intercept and slope. Perfect calibration = 1. Values < 1 indicate overconfident predictions (predicted probabilities too extreme); values > 1 indicate underconfident predictions.

Brier score
Mean squared difference between predicted probabilities and binary outcomes. Range [0, 1]; lower is better.

ECE (Expected Calibration Error)
Weighted average of absolute calibration error across equal-frequency bins. Lower is better.

ICI (Integrated Calibration Index)
Mean absolute difference between LOESS-smoothed observed probabilities and predicted probabilities (Austin & Steyerberg 2019). Lower is better; more robust than ECE as it avoids binning artefacts.

Hosmer–Lemeshow statistic
Chi-square-distributed goodness-of-fit statistic comparing observed with expected events across deciles of predicted risk. A large value indicates lack of fit. **It scales with sample size**, so it is not comparable between subgroups of different size and must be read alongside *n*, the per-decile table, and the calibration measures above (Kramer & Zimmerman 2007).

Hosmer–Lemeshow degrees of freedom
Equal to the number of risk groups used. The classical convention subtracts 2, which is the penalty for estimating an intercept and slope on the data under test; this package evaluates predictions supplied by a model fitted elsewhere, so no degrees of freedom are spent and none are subtracted.

Decile slope and intercept
Least-squares regression of the observed on the expected event rate across risk groups, recommended by Kramer & Zimmerman (2007) as an adjunct to the statistic. Perfect calibration gives a slope of 1 and an intercept of 0.

Clinical utility (decision curves)

Decision curve analysis quantifies the clinical value of using the model at a given decision threshold.

Net benefit
TP/n − FP/n × $p_t/(1 - p_t)$, where p_t is the decision threshold. Higher is better. A model is clinically useful when its net benefit exceeds both 0 (treat none) and the treat-all reference.

Standardised net benefit
Net benefit divided by prevalence (Naderalvojoud 2025). Normalised to a [0, 1] scale for comparison across cohorts with different prevalences.

Treat-all reference
Net benefit if all patients receive treatment. = prevalence − (1 − prevalence) × $p_t/(1 - p_t)$. Becomes strongly negative at high thresholds.

Treat-none reference
Net benefit = 0. The baseline of no intervention.

Gap statistics

Gap statistics quantify the largest disparity between subgroups for each metric. All gaps are non-negative; 0 = perfect parity.

Equal opportunity difference
max − min of sensitivity across subgroups. Measures fairness in true positive identification.

Predictive equality
max − min of false positive rate (1 − specificity) across subgroups. Measures fairness in false alarm rates.

Equalized odds gap
Maximum of equal opportunity difference and predictive equality.

Statistical parity
max − min of positive prediction rate across subgroups. Measures whether the model predicts positive at equal rates.

Max intercept / slope gap
Maximum pairwise difference in calibration intercept or slope across subgroups, with bootstrap confidence intervals.

Max net benefit gap
Maximum across all thresholds of the maximum pairwise net-benefit difference between subgroups, with bootstrap confidence intervals.

Confidence intervals and shrinkage

Bootstrap confidence intervals
95% percentile intervals from subgroup-stratified patient-level resamples. Patients are resampled with replacement within each subgroup, preserving the subgroup's size while letting its event count vary.

Empirical Bayes shrinkage
Interaction cells with fewer events than the configured threshold (shown in italics) use Beta-Binomial shrinkage toward the marginal subgroup rate. This reduces instability in small cells. Confidence intervals are suppressed for shrunk estimates.

Generated by **isitfair**. Three-axis evaluation: discrimination, calibration, and clinical utility reported jointly per TRIPOD+AI recommendations.

Bootstrap confidence intervals (95%) from 2000 subgroup-stratified patient-level resamples. Interaction cells with fewer than the configured event threshold use empirical Bayes shrinkage (italicised rows).

ICU mortality prediction (SUPPORT2) -- subgroup deployment-readiness audit

Model	Logistic regression
Cohort	SUPPORT2 (UCI, n=9,105); held-out test partition n=1,821, 472 deaths
Outcome	In-hospital mortality
Threshold	0.3
Uncertainty	200 patient-level bootstrap resamples; event floors applied
Role	Software reproducibility example, not external validation of the perioperative infection model. Not pooled with the Bern analysis.

Audit summary

N: 1821	**Events:** 472	**Prevalence:** 0.259	**Threshold:** 0.300
Bootstrap resamples: 200	**Resampling:** subgroup-stratified patient-level	**Attributes:** 3	

Discrimination

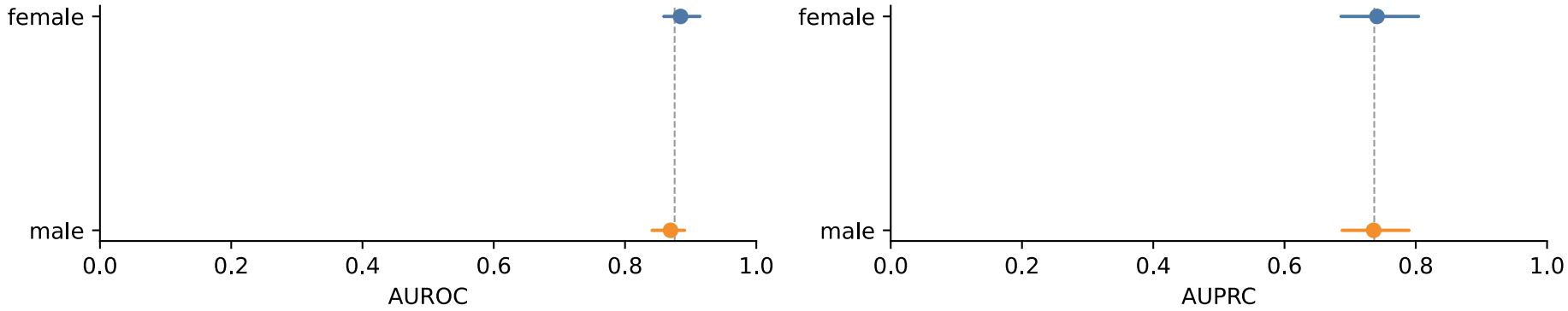
Attribute	Subgroup	n	Events	Prevalence	AUROC (95% CI)	AUPRC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
sex	Overall	1821	472	0.259	0.876 (0.861–0.896)	0.737 (0.697–0.780)	0.746 (0.711–0.782)	0.856 (0.838–0.872)
sex	female	787	199	0.253	0.885 (0.859–0.914)	0.741 (0.686–0.804)	0.784 (0.726–0.849)	0.835 (0.806–0.861)
sex	male	1034	273	0.264	0.869 (0.841–0.891)	0.736 (0.688–0.790)	0.718 (0.668–0.767)	0.873 (0.846–0.900)
age_group	Overall	1821	472	0.259	0.876 (0.861–0.896)	0.737 (0.697–0.780)	0.746 (0.711–0.782)	0.856 (0.838–0.872)
age_group	18-44	257	65	0.253	0.911 (0.862–0.950)	0.800 (0.671–0.884)	0.754 (0.619–0.867)	0.875 (0.834–0.918)
age_group	45-64	684	171	0.250	0.867 (0.832–0.900)	0.748 (0.694–0.812)	0.696 (0.635–0.759)	0.879 (0.854–0.910)
age_group	65-79	677	172	0.254	0.877 (0.841–0.906)	0.746 (0.679–0.796)	0.802 (0.740–0.854)	0.834 (0.800–0.861)
age_group	80+	203	64	0.315	0.852 (0.798–0.906)	0.681 (0.562–0.814)	0.719 (0.594–0.829)	0.827 (0.752–0.895)
race_ethnicity	Overall	1821	472	0.259	0.876 (0.861–0.896)	0.737 (0.697–0.780)	0.746 (0.711–0.782)	0.856 (0.838–0.872)
race_ethnicity	asian	13	5	0.385	0.850	0.796	0.400	1.000
race_ethnicity	black	266	82	0.308	0.879 (0.832–0.922)	0.786 (0.713–0.858)	0.732 (0.625–0.838)	0.837 (0.783–0.888)
race_ethnicity	hispanic	57	10	0.175	0.909	0.813	0.900	0.830
race_ethnicity	other	32	9	0.281	0.971	0.919	0.778	0.957
race_ethnicity	white	1453	366	0.252	0.873 (0.852–0.892)	0.722 (0.677–0.767)	0.749 (0.702–0.787)	0.857 (0.839–0.875)

Gap statistics

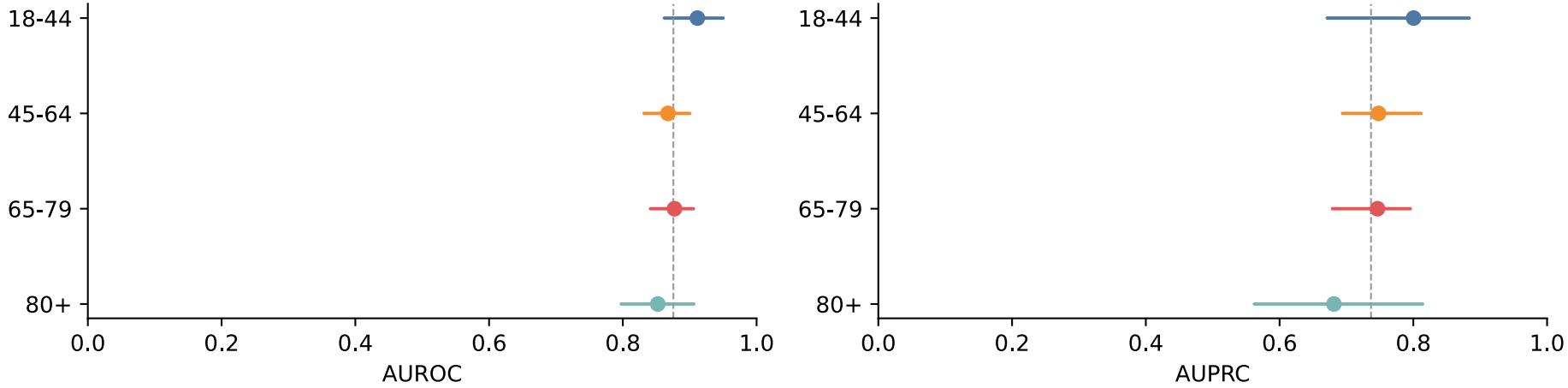
Effect-size column shows Cohen's *h* on rate-metric gaps for two-group attributes; multi-group attributes and ranking metrics (AUROC / AUPRC) show "—".

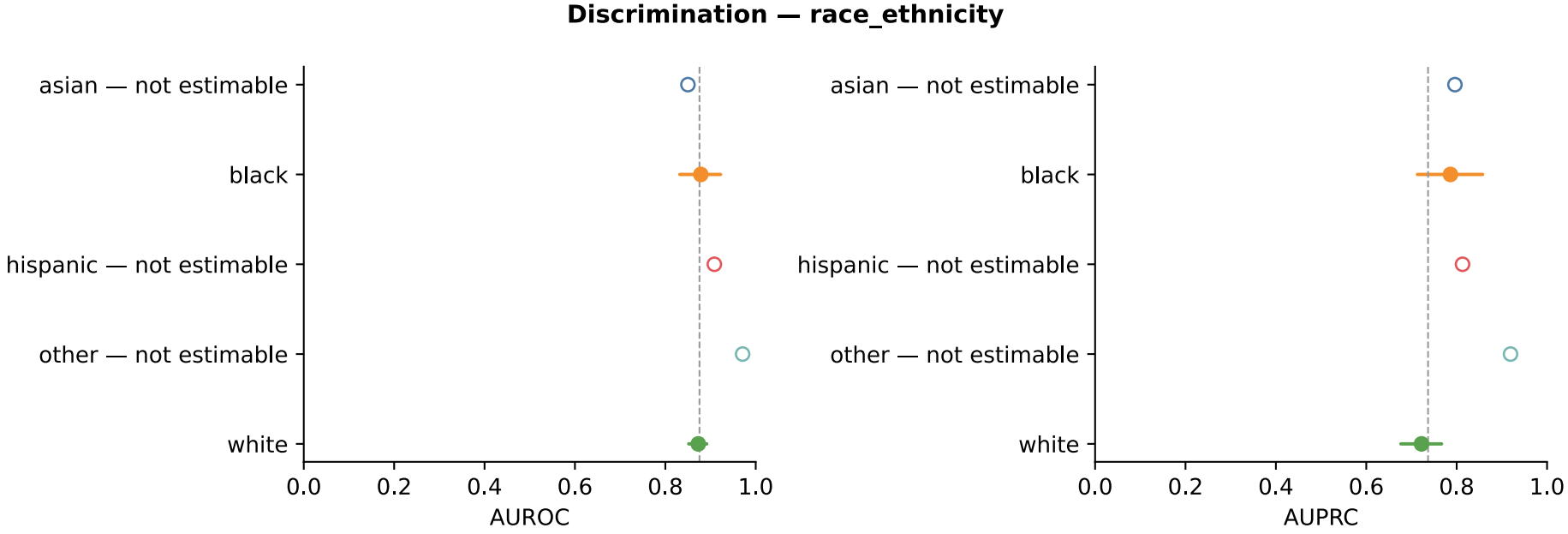
Attribute	Metric	Value	Effect size
sex	equal_opportunity_difference	0.0660	h=+0.15 (negligible)
sex	predictive_equality	0.0375	h=-0.11 (negligible)
sex	equalized_odds_gap	0.0660	—
sex	statistical_parity	0.0381	—
age_group	equal_opportunity_difference	0.1064	—
age_group	predictive_equality	0.0518	—
age_group	equalized_odds_gap	0.1064	—
age_group	statistical_parity	0.0802	—
race_ethnicity	equal_opportunity_difference	0.5000	—
race_ethnicity	predictive_equality	0.1702	—
race_ethnicity	equalized_odds_gap	0.5000	—
race_ethnicity	statistical_parity	0.1845	—

Discrimination — sex



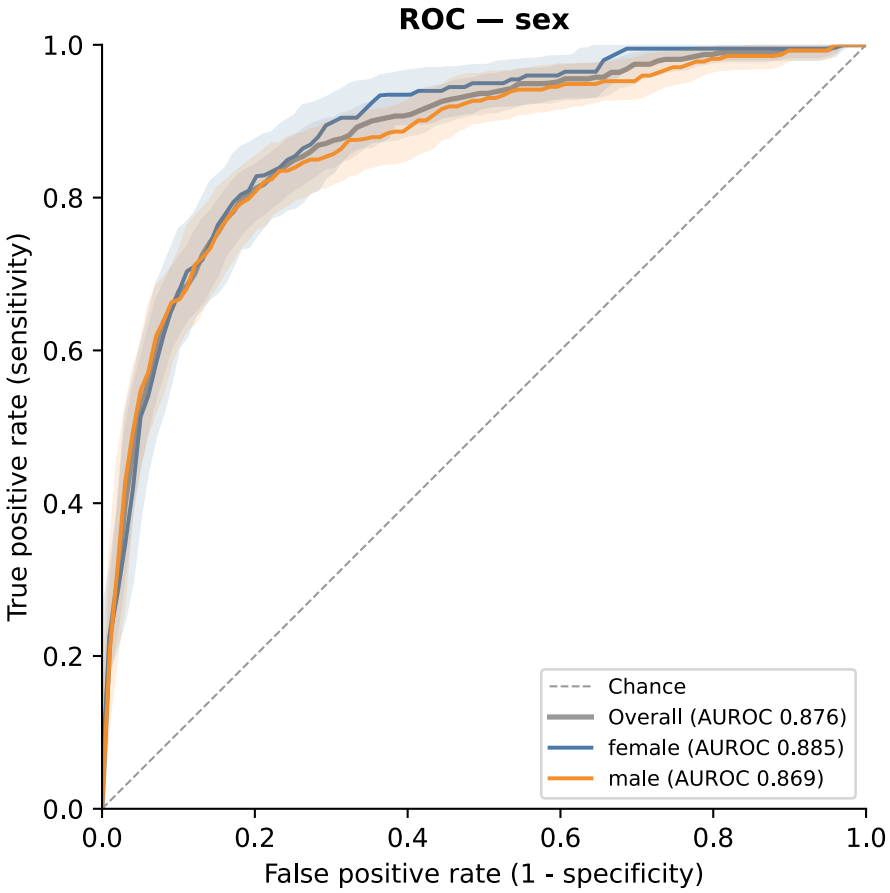
Discrimination — age_group

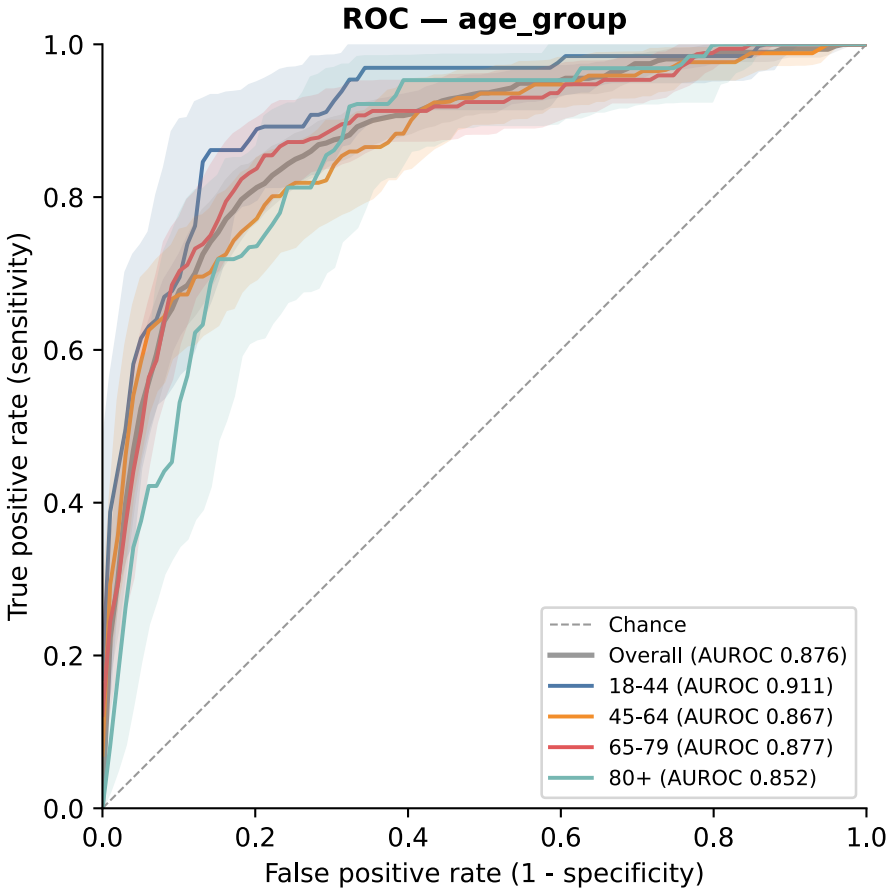


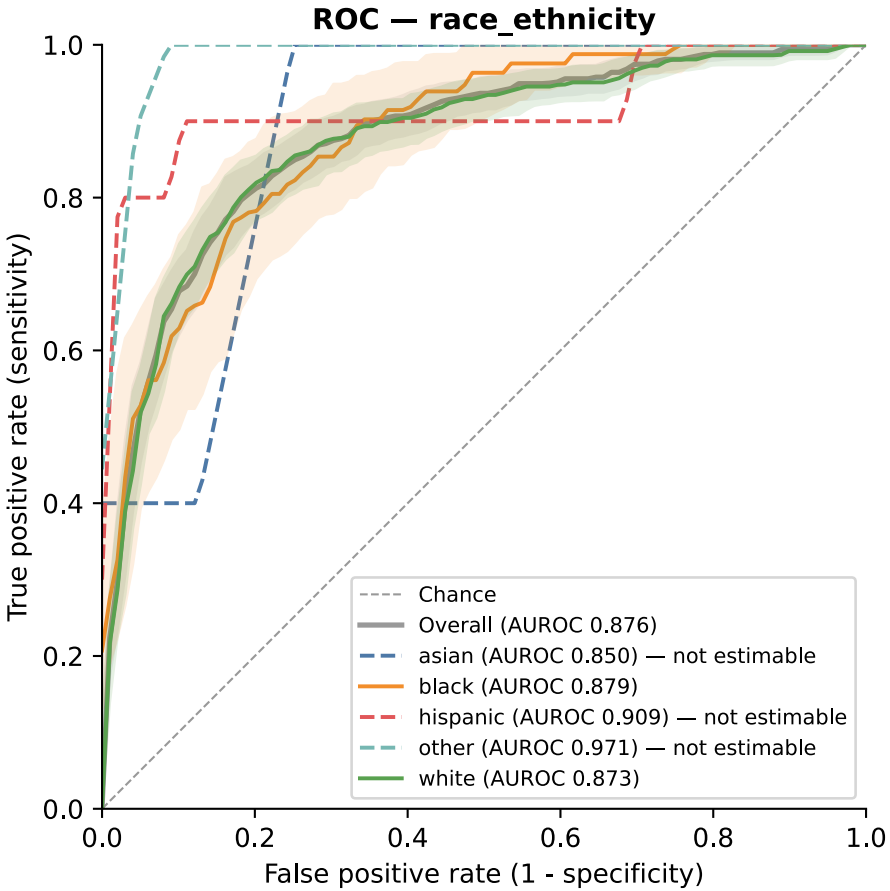


ROC curves

Per-subgroup ROC curves with pointwise 95% bootstrap bands. AUROC values in the legend are taken from the table above, not integrated from the drawn curve. Curves that cross indicate that no single subgroup dominates at every operating point — the ranking of subgroups depends on the threshold chosen.


ROC — sex
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
female (AUROC 0.885)
male (AUROC 0.869)


ROC — age_group
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
18-44 (AUROC 0.911)
45-64 (AUROC 0.867)
65-79 (AUROC 0.877)
80+ (AUROC 0.852)

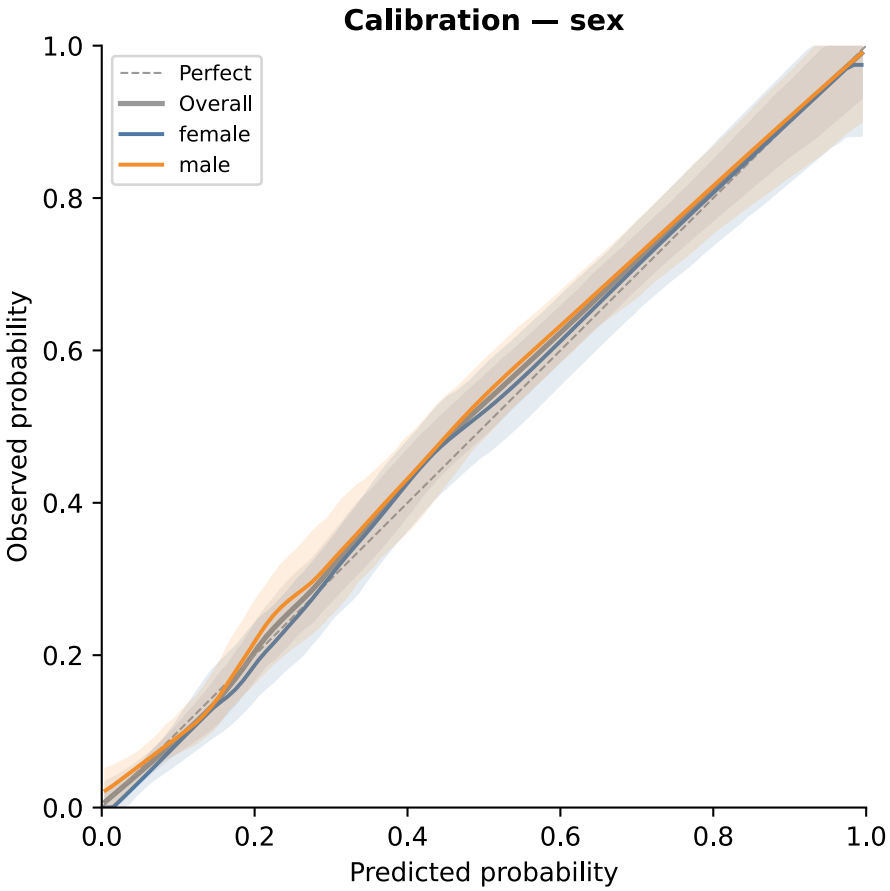

ROC — race_ethnicity
True positive rate (sensitivity)
False positive rate (1 - specificity)
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Overall (AUROC 0.876)
asian (AUROC 0.850) — not estimable
black (AUROC 0.879)
hispanic (AUROC 0.909) — not estimable
other (AUROC 0.971) — not estimable
white (AUROC 0.873)

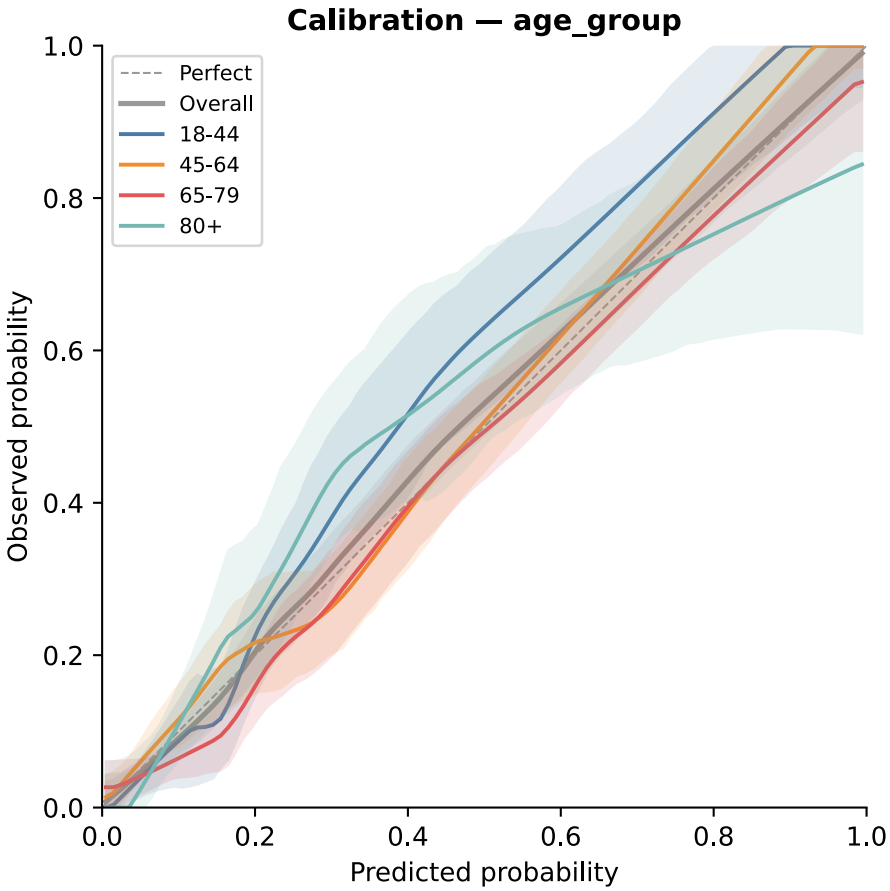
Calibration

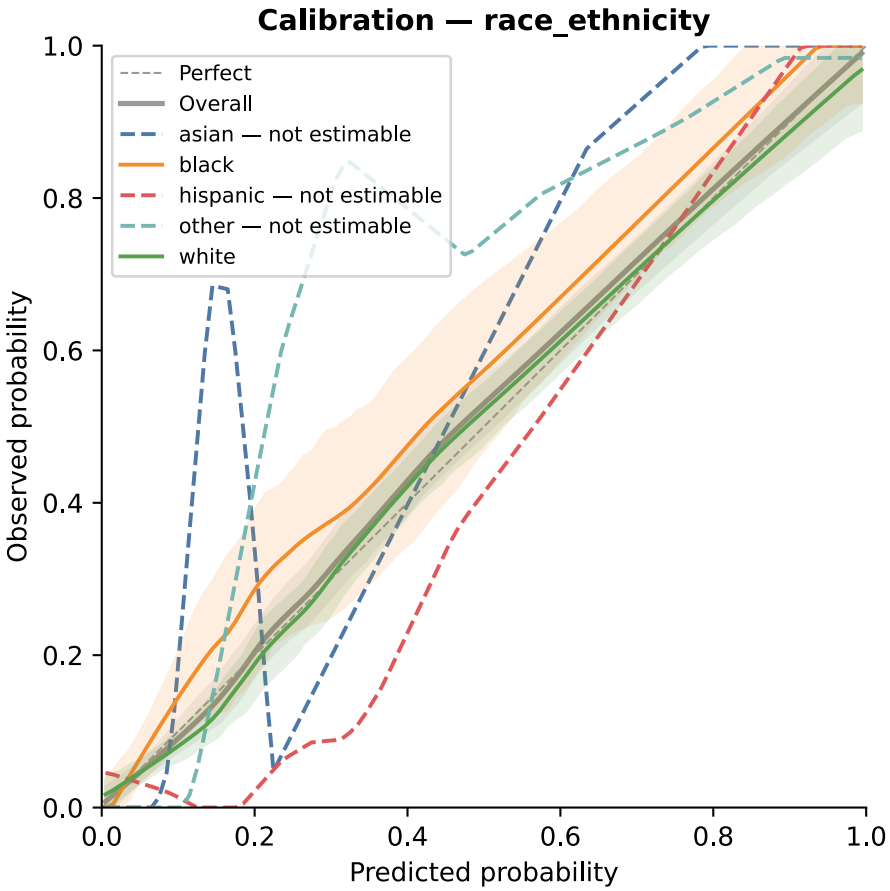
Attribute	Subgroup	n	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
sex	Overall	1821	0.035 (-0.080–0.149)	1.041 (0.957–1.146)	0.115 (0.105–0.123)	0.0178	0.0099
sex	female	787	-0.039 (-0.228–0.146)	1.108 (0.969–1.315)	0.113 (0.098–0.124)	0.0322	0.0136
sex	male	1034	0.093 (-0.078–0.277)	1.000 (0.862–1.128)	0.116 (0.105–0.129)	0.0240	0.0135
age_group	Overall	1821	0.035 (-0.080–0.149)	1.041 (0.957–1.146)	0.115 (0.105–0.123)	0.0178	0.0099
age_group	18-44	257	0.269 (-0.111–0.564)	1.342	0.103 (0.077–0.125)	0.0537	0.0453
age_group	45-64	684	0.117 (-0.131–0.325)	1.006 (0.862–1.170)	0.112 (0.095–0.127)	0.0321	0.0186
age_group	65-79	677	-0.192 (-0.427–-0.008)	1.064 (0.906–1.231)	0.112 (0.098–0.127)	0.0405	0.0235
age_group	80+	203	0.221 (-0.160–0.551)	0.918	0.149 (0.118–0.181)	0.0818	0.0568
race_ethnicity	Overall	1821	0.035 (-0.080–0.149)	1.041 (0.957–1.146)	0.115 (0.105–0.123)	0.0178	0.0099
race_ethnicity	asian	13	1.107	2.336	0.184	0.3010	0.2098
race_ethnicity	black	266	0.379 (-0.024–0.682)	1.059	0.130 (0.103–0.155)	0.0533	0.0459
race_ethnicity	hispanic	57	-0.615	1.306	0.068	0.1092	0.0771
race_ethnicity	other	32	0.510	2.100	0.100	0.1481	0.1361
race_ethnicity	white	1453	-0.033 (-0.198–0.103)	1.027 (0.919–1.128)	0.114 (0.105–0.125)	0.0225	0.0115

Calibration gap statistics

Attribute	Metric	Value (95% CI)
sex	max_intercept_gap	0.1324 (0.0083–0.4089)
sex	max_slope_gap	0.1076 (0.0039–0.3160)
age_group	max_intercept_gap	0.4614 (0.2440–0.8955)
age_group	max_slope_gap	0.4240
race_ethnicity	max_intercept_gap	1.7220
race_ethnicity	max_slope_gap	1.3091


Calibration — sex
Perfect
Overall
female
male
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0


Calibration — age_group
Perfect
Overall
18-44
45-64
65-79
80+
Observed probability
Predicted probability
0.0
0.2
0.4
0.6
0.8
1.0



Hosmer–Lemeshow goodness-of-fit

Read every statistic beside its *n* and its decile table. The Hosmer–Lemeshow statistic scales with subgroup size: a clinically irrelevant deviation becomes significant in a large subgroup, while a large deviation can go undetected in a small one. Values are therefore **not comparable across subgroups of different size**, and subgroups must not be ranked by p-value. This is why Kramer & Zimmerman (2007) require the statistic to be reported together with the number of patients, the per-decile observed-versus-expected table below, and the adjunct calibration measures in the table above — never on its own.

Attribute	Subgroup	n	Events	H–L statistic	df	p	Groups used	Decile slope	Decile intercept
sex	Overall	1821	472	6.04	10	0.8120	10	1.032	-0.004
sex	female	787	199	12.14	10	0.2756	10	1.027	-0.012
sex	male	1034	273	6.78	10	0.7456	10	1.024	+0.005
age_group	Overall	1821	472	6.04	10	0.8120	10	1.032	-0.004

Attribute	Subgroup	n	Events	H–L statistic	df	p	Groups used	Decile slope	Decile intercept
age_group	18-44	257	65	8.63	10	0.5672	10	1.229	-0.020
age_group	45-64	684	171	8.29	10	0.6001	10	1.033	+0.005
age_group	65-79	677	172	18.61	10	0.0456	10	0.986	-0.019
age_group	80+	203	64	12.55	10	0.2497	10	0.987	+0.035
race_ethnicity	Overall	1821	472	6.04	10	0.8120	10	1.032	-0.004
race_ethnicity	asian	13	5	17.53	10	0.0635	10	1.038	+0.226
race_ethnicity	black	266	82	9.65	10	0.4715	10	1.063	+0.030
race_ethnicity	hispanic	57	10	7.75	10	0.6536	10	1.023	-0.063
race_ethnicity	other	32	9	10.01	10	0.4396	10	1.392	-0.024
race_ethnicity	white	1453	366	8.73	10	0.5580	10	1.013	-0.007

† Fewer risk groups than requested were usable, because the subgroup was small or its predicted probabilities were tied. Degrees of freedom follow the reduced count. A well-calibrated model gives a decile slope near 1 and a decile intercept near 0 (Kramer & Zimmerman 2007). 14 tests are reported here; no multiplicity adjustment has been applied. Adjust within an attribute if needed — do not pool the “Overall” rows with the subgroup rows.

Attribute	Metric	Value
sex	max_hl_gap	5.36
age_group	max_hl_gap	10.31
race_ethnicity	max_hl_gap	9.78

The gap is the largest pairwise difference in the statistic across subgroups. It inherits the sample-size dependence in full and should be read as a descriptive summary, not as evidence of differential calibration.

Observed versus expected events per decile

Deciles of predicted risk, of approximately equal size, with cut points snapped to tied predicted probabilities so that identical predictions always share a decile.

sex — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	183	5	3.5	0.0273	0.0193	+1.5
2	182	5	6.7	0.0275	0.0367	-1.7
3	182	11	9.9	0.0604	0.0543	+1.1
4	182	10	13.8	0.0549	0.0760	-3.8
5	182	16	19.1	0.0879	0.1047	-3.1
6	182	26	28.0	0.1429	0.1540	-2.0
7	182	47	44.5	0.2582	0.2447	+2.5
8	182	78	72.0	0.4286	0.3955	+6.0
9	182	122	113.1	0.6703	0.6216	+8.9
10	182	152	153.8	0.8352	0.8451	-1.8

sex — female

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	79	1	1.5	0.0127	0.0187	-0.5
2	79	0	2.8	0.0000	0.0356	-2.8
3	79	6	4.2	0.0759	0.0534	+1.8
4	79	4	6.1	0.0506	0.0776	-2.1
5	79	4	8.6	0.0506	0.1085	-4.6
6	79	15	13.2	0.1899	0.1675	+1.8
7	79	21	21.4	0.2658	0.2713	-0.4
8	78	32	32.0	0.4103	0.4102	+0.0
9	78	55	47.7	0.7051	0.6118	+7.3
10	78	61	65.1	0.7821	0.8349	-4.1

sex — male

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	104	4	2.1	0.0385	0.0199	+1.9
2	104	5	3.9	0.0481	0.0375	+1.1
3	104	5	5.8	0.0481	0.0553	-0.8
4	104	6	7.8	0.0577	0.0754	-1.8
5	103	13	10.6	0.1262	0.1030	+2.4
6	103	12	15.3	0.1165	0.1488	-3.3
7	103	27	23.5	0.2621	0.2282	+3.5
8	103	43	39.7	0.4175	0.3856	+3.3
9	103	71	65.2	0.6893	0.6328	+5.8
10	103	87	87.9	0.8447	0.8533	-0.9

age_group — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	183	5	3.5	0.0273	0.0193	+1.5
2	182	5	6.7	0.0275	0.0367	-1.7
3	182	11	9.9	0.0604	0.0543	+1.1
4	182	10	13.8	0.0549	0.0760	-3.8
5	182	16	19.1	0.0879	0.1047	-3.1
6	182	26	28.0	0.1429	0.1540	-2.0
7	182	47	44.5	0.2582	0.2447	+2.5
8	182	78	72.0	0.4286	0.3955	+6.0
9	182	122	113.1	0.6703	0.6216	+8.9
10	182	152	153.8	0.8352	0.8451	-1.8

age_group — 18-44

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	26	0	0.3	0.0000	0.0118	-0.3
2	26	1	0.6	0.0385	0.0212	+0.4
3	26	0	0.9	0.0000	0.0354	-0.9
4	26	1	1.4	0.0385	0.0522	-0.4
5	26	1	2.3	0.0385	0.0896	-1.3
6	26	4	3.8	0.1538	0.1449	+0.2
7	26	7	6.3	0.2692	0.2413	+0.7
8	25	11	8.5	0.4400	0.3405	+2.5
9	25	17	12.9	0.6800	0.5142	+4.1
10	25	23	20.1	0.9200	0.8058	+2.9

age_group — 45-64

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	69	2	1.2	0.0290	0.0172	+0.8
2	69	3	2.1	0.0435	0.0305	+0.9
3	69	4	3.3	0.0580	0.0476	+0.7
4	69	3	4.6	0.0435	0.0667	-1.6
5	68	10	6.2	0.1471	0.0911	+3.8
6	68	9	9.0	0.1324	0.1321	+0.0
7	68	16	13.7	0.2353	0.2009	+2.3
8	68	19	23.9	0.2794	0.3517	-4.9
9	68	47	41.9	0.6912	0.6159	+5.1
10	68	58	56.2	0.8529	0.8272	+1.8

age_group — 65-79

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	68	0	1.9	0.0000	0.0279	-1.9
2	68	7	3.2	0.1029	0.0470	+3.8
3	68	2	4.5	0.0294	0.0657	-2.5
4	68	4	5.9	0.0588	0.0867	-1.9
5	68	2	7.8	0.0294	0.1151	-5.8
6	68	7	11.7	0.1029	0.1723	-4.7
7	68	21	18.9	0.3088	0.2786	+2.1
8	67	31	30.9	0.4627	0.4614	+0.1
9	67	44	44.5	0.6567	0.6645	-0.5
10	67	54	58.2	0.8060	0.8682	-4.2

age_group — 80+

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	21	0	0.7	0.0000	0.0346	-0.7
2	21	2	1.1	0.0952	0.0502	+0.9
3	21	1	1.6	0.0476	0.0757	-0.6
4	20	0	2.1	0.0000	0.1050	-2.1
5	20	5	3.0	0.2500	0.1519	+2.0
6	20	7	4.6	0.3500	0.2321	+2.4
7	20	8	6.2	0.4000	0.3076	+1.8
8	20	13	8.7	0.6500	0.4362	+4.3
9	20	12	12.5	0.6000	0.6265	-0.5
10	20	16	17.3	0.8000	0.8638	-1.3

race_ethnicity — Overall

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	183	5	3.5	0.0273	0.0193	+1.5
2	182	5	6.7	0.0275	0.0367	-1.7
3	182	11	9.9	0.0604	0.0543	+1.1
4	182	10	13.8	0.0549	0.0760	-3.8
5	182	16	19.1	0.0879	0.1047	-3.1
6	182	26	28.0	0.1429	0.1540	-2.0
7	182	47	44.5	0.2582	0.2447	+2.5
8	182	78	72.0	0.4286	0.3955	+6.0
9	182	122	113.1	0.6703	0.6216	+8.9
10	182	152	153.8	0.8352	0.8451	-1.8

race_ethnicity — asian

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	2	0	0.1	0.0000	0.0517	-0.1
2	2	0	0.2	0.0000	0.0814	-0.2
3	2	0	0.2	0.0000	0.1012	-0.2
4	1	1	0.1	1.0000	0.1424	+0.9
5	1	1	0.2	1.0000	0.1652	+0.8
6	1	1	0.2	1.0000	0.1765	+0.8
7	1	0	0.2	0.0000	0.2006	-0.2
8	1	0	0.2	0.0000	0.2236	-0.2
9	1	1	0.6	1.0000	0.6347	+0.4
10	1	1	0.9	1.0000	0.8612	+0.1

race_ethnicity — black

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	27	0	0.4	0.0000	0.0154	-0.4
2	27	1	0.9	0.0370	0.0328	+0.1
3	27	1	1.3	0.0370	0.0491	-0.3
4	27	3	1.9	0.1111	0.0719	+1.1
5	27	6	2.9	0.2222	0.1078	+3.1
6	27	6	4.7	0.2222	0.1732	+1.3
7	26	11	7.6	0.4231	0.2923	+3.4
8	26	11	11.1	0.4231	0.4279	-0.1
9	26	20	16.5	0.7692	0.6358	+3.5
10	26	23	22.1	0.8846	0.8514	+0.9

race_ethnicity — hispanic

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	6	0	0.1	0.0000	0.0134	-0.1
2	6	0	0.2	0.0000	0.0293	-0.2
3	6	1	0.3	0.1667	0.0447	+0.7
4	6	0	0.4	0.0000	0.0702	-0.4
5	6	0	0.5	0.0000	0.0842	-0.5
6	6	0	0.8	0.0000	0.1303	-0.8
7	6	0	1.6	0.0000	0.2661	-1.6
8	5	1	1.8	0.2000	0.3686	-0.8
9	5	4	3.3	0.8000	0.6552	+0.7
10	5	4	4.4	0.8000	0.8725	-0.4

race_ethnicity — other

Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	4	0	0.1	0.0000	0.0225	-0.1
2	4	0	0.2	0.0000	0.0425	-0.2
3	3	0	0.2	0.0000	0.0610	-0.2
4	3	0	0.2	0.0000	0.0820	-0.2
5	3	0	0.3	0.0000	0.1005	-0.3
6	3	0	0.4	0.0000	0.1326	-0.4
7	3	1	0.6	0.3333	0.2035	+0.4
8	3	3	0.9	1.0000	0.2900	+2.1
9	3	2	1.8	0.6667	0.5907	+0.2
10	3	3	2.4	1.0000	0.7991	+0.6

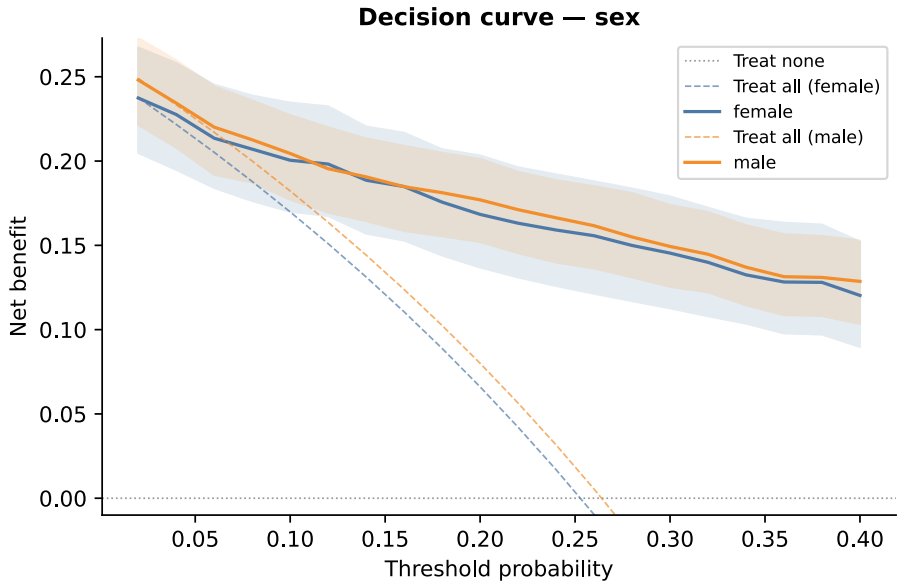
race_ethnicity — white

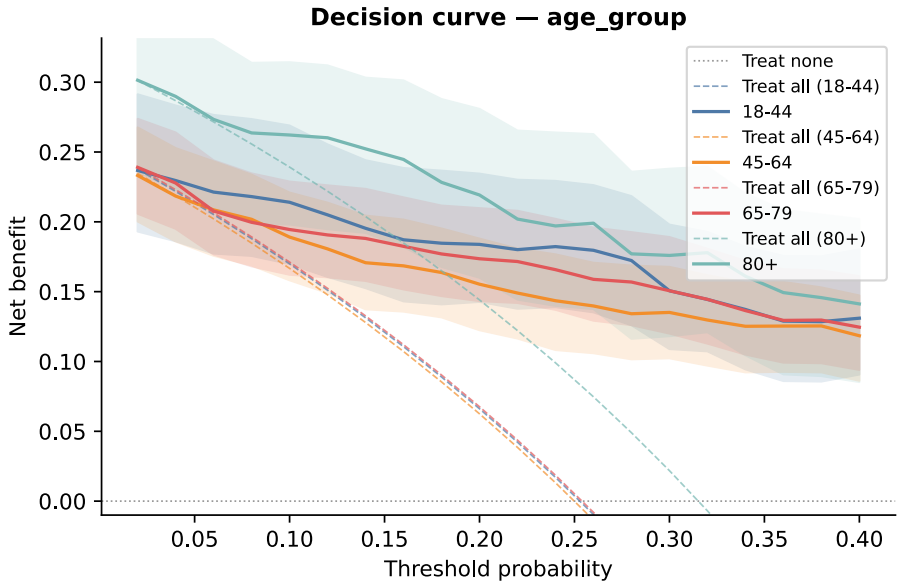
Decile	n	Observed	Expected	Observed rate	Expected rate	Difference
1	146	5	3.0	0.0342	0.0204	+2.0
2	146	3	5.5	0.0205	0.0377	-2.5
3	146	10	8.1	0.0685	0.0557	+1.9
4	145	8	11.2	0.0552	0.0775	-3.2
5	145	11	15.4	0.0759	0.1065	-4.4
6	145	17	22.4	0.1172	0.1548	-5.4
7	145	37	35.4	0.2552	0.2441	+1.6
8	145	62	57.7	0.4276	0.3980	+4.3
9	145	94	90.3	0.6483	0.6228	+3.7
10	145	119	122.5	0.8207	0.8451	-3.5

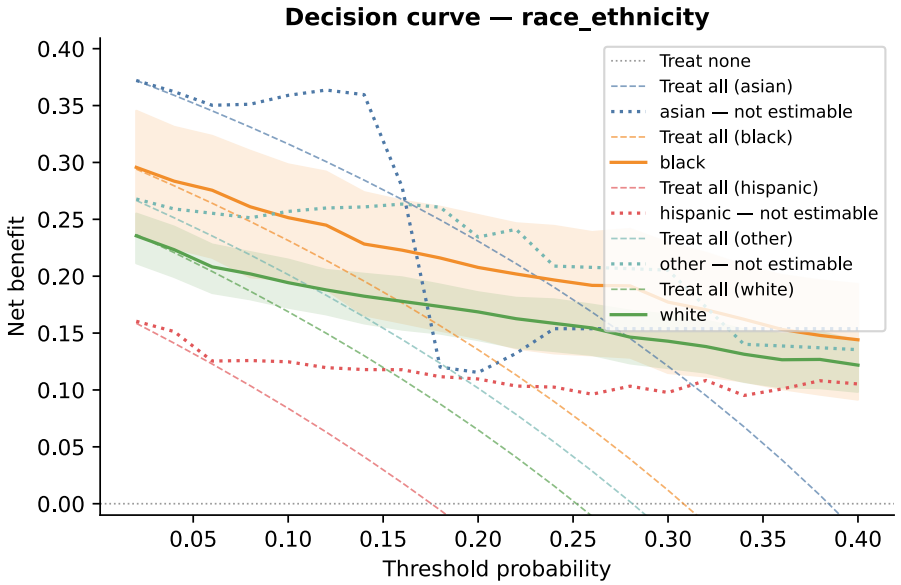
Clinical utility (decision curves)

Clinical utility gap statistics

Attribute	Metric	Value (95% CI)
sex	max_net_benefit_gap	0.0107
age_group	max_net_benefit_gap	0.0815 (0.0518–0.1678)
race_ethnicity	max_net_benefit_gap	0.2440






Decision curve — race_ethnicity
Net benefit
Threshold probability
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Treat none
Treat all (asian)
asian — not estimable
Treat all (black)
black
Treat all (hispanic)
hispanic — not estimable
Treat all (other)
other — not estimable
Treat all (white)
white

Mitigation comparison

sex

Calibration

Recalibration targets calibration, so this is where its effect belongs. Compare the intercept against 0 and the slope against 1 within each subgroup, and the gap statistics below across subgroups.

Before

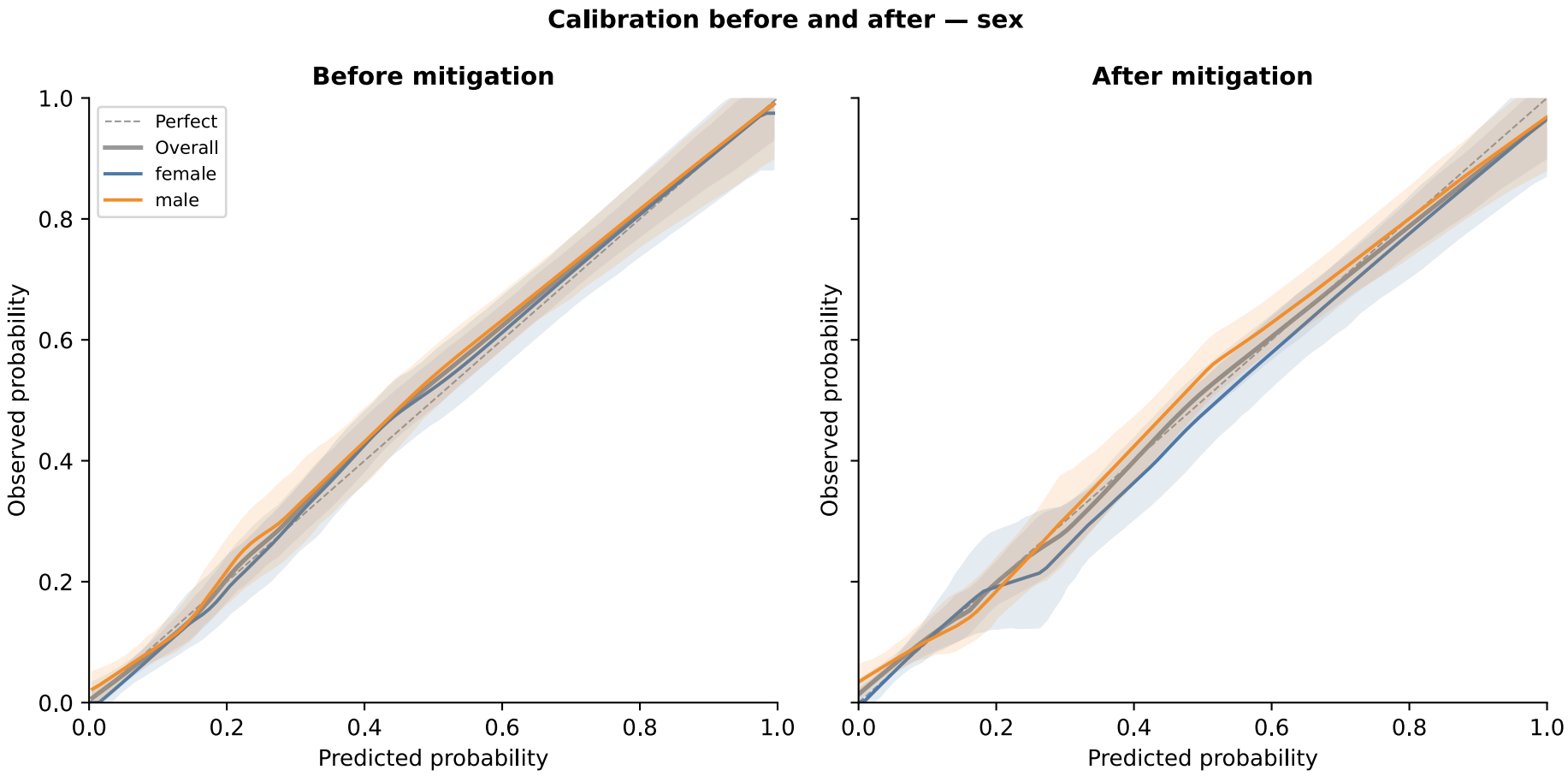
Subgroup	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
female	-0.039 (-0.228–0.146)	1.108 (0.969–1.315)	0.113 (0.098–0.124)	0.0322	0.0136
male	0.093 (-0.078–0.277)	1.000 (0.862–1.128)	0.116 (0.105–0.129)	0.0240	0.0135

After

Subgroup	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
female	-0.075 (-0.277–0.121)	0.766 (0.498–1.072)	0.116 (0.101–0.130)	0.0276	0.0114
male	0.146 (-0.043–0.360)	0.819 (0.637–0.998)	0.118 (0.107–0.131)	0.0343	0.0194

Calibration gaps

Metric	Before (95% CI)	After (95% CI)
max_intercept_gap	0.1324 (0.0083–0.4089)	0.2205 (0.0286–0.5335)
max_slope_gap	0.1076 (0.0039–0.3160)	0.0536 (0.0061–0.3926)



Discrimination (sanity check)

Per-subgroup recalibration is a monotone transform within each subgroup, so within-subgroup AUROC is invariant except for ties that isotonic regression introduces, and any sensitivity change is a threshold artefact. These columns barely moving is the expected result, not a null finding — a large shift here means the mitigation did something other than recalibrate.

Before

Subgroup	AUROC	Sensitivity
female	0.885 (0.859–0.914)	0.784 (0.726–0.849)
male	0.869 (0.841–0.891)	0.718 (0.668–0.767)

After

Subgroup	AUROC	Sensitivity
female	0.879 (0.850–0.907)	0.824 (0.771–0.882)
male	0.862 (0.834–0.885)	0.725 (0.669–0.776)

age_group

Calibration

Recalibration targets calibration, so this is where its effect belongs. Compare the intercept against 0 and the slope against 1 within each subgroup, and the gap statistics below across subgroups.

Before

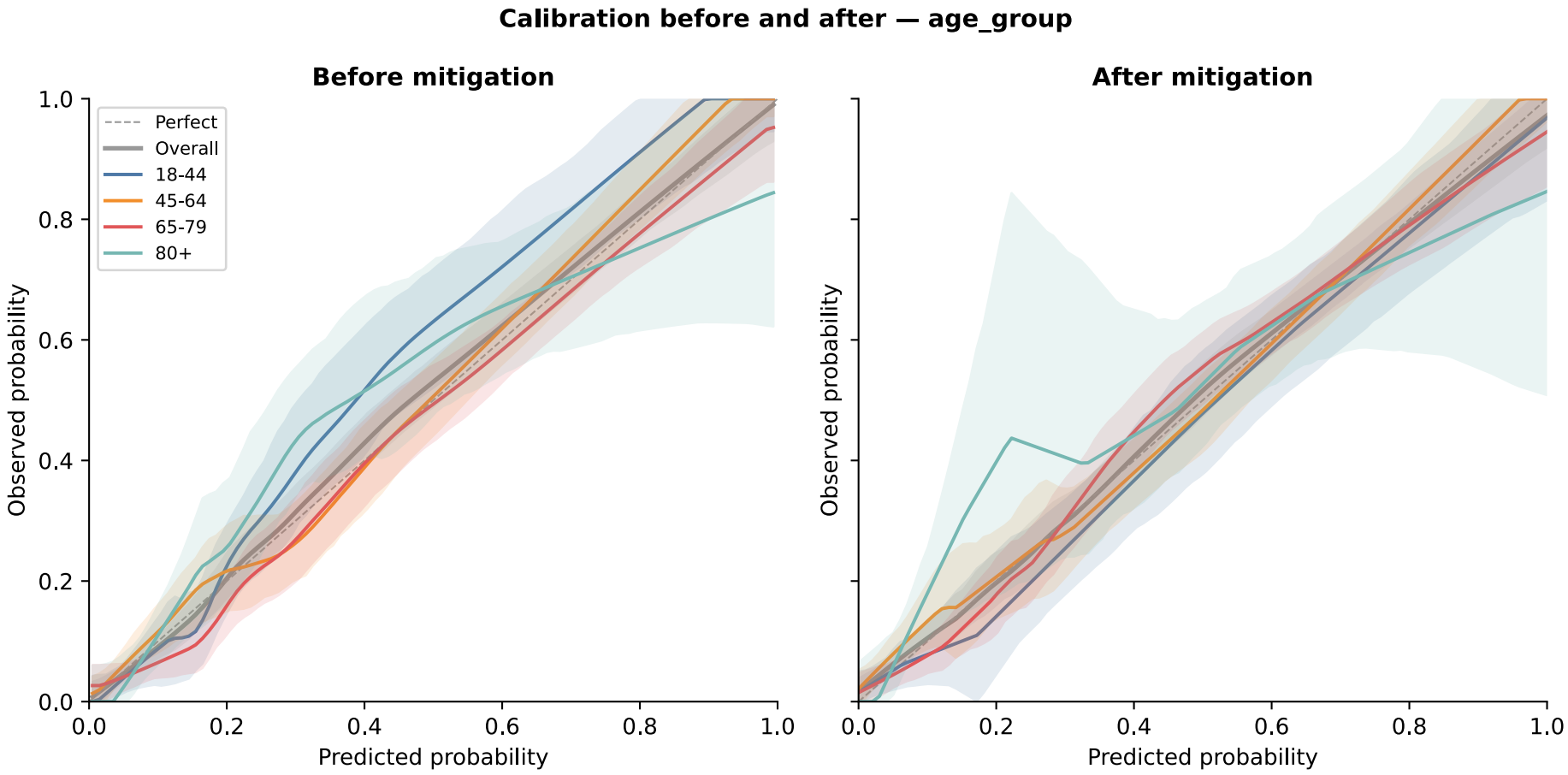
Subgroup	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
18-44	0.269 (-0.111–0.564)	1.342	0.103 (0.077–0.125)	0.0537	0.0453
45-64	0.117 (-0.131–0.325)	1.006 (0.862–1.170)	0.112 (0.095–0.127)	0.0321	0.0186
65-79	-0.192 (-0.427—0.008)	1.064 (0.906–1.231)	0.112 (0.098–0.127)	0.0405	0.0235
80+	0.221 (-0.160–0.551)	0.918	0.149 (0.118–0.181)	0.0818	0.0568

After

Subgroup	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
18-44	-0.094 (-0.570–0.273)	0.641	0.098 (0.071–0.120)	0.0215	0.0183
45-64	0.147 (-0.120–0.370)	0.865 (0.732–1.024)	0.115 (0.098–0.130)	0.0390	0.0217
65-79	-0.007 (-0.236–0.177)	1.031 (0.855–1.223)	0.112 (0.098–0.130)	0.0435	0.0202
80+	0.252 (-0.148–0.608)	0.227	0.155 (0.125–0.189)	0.0747	0.0498

Calibration gaps

Metric	Before (95% CI)	After (95% CI)
max_intercept_gap	0.4614 (0.2440–0.8955)	0.3459 (0.1336–1.0749)
max_slope_gap	0.4240	0.8041



Discrimination (sanity check)

Per-subgroup recalibration is a monotone transform within each subgroup, so within-subgroup AUROC is invariant except for ties that isotonic regression introduces, and any sensitivity change is a threshold artefact. These columns barely moving is the expected result, not a null finding — a large shift here means the mitigation did something other than recalibrate.

Before

Subgroup	AUROC	Sensitivity
18-44	0.911 (0.862–0.950)	0.754 (0.619–0.867)
45-64	0.867 (0.832–0.900)	0.696 (0.635–0.759)

After

Subgroup	AUROC	Sensitivity
18-44	0.904 (0.852–0.948)	0.862 (0.759–0.938)
45-64	0.865 (0.828–0.897)	0.696 (0.635–0.759)

Subgroup	AUROC	Sensitivity
65-79	0.877 (0.841–0.906)	0.802 (0.740–0.854)
80+	0.852 (0.798–0.906)	0.719 (0.594–0.829)

Subgroup	AUROC	Sensitivity
65-79	0.875 (0.841–0.905)	0.797 (0.727–0.848)
80+	0.844 (0.790–0.897)	0.750 (0.645–0.855)

race_ethnicity

Calibration

Recalibration targets calibration, so this is where its effect belongs. Compare the intercept against 0 and the slope against 1 within each subgroup, and the gap statistics below across subgroups.

Before

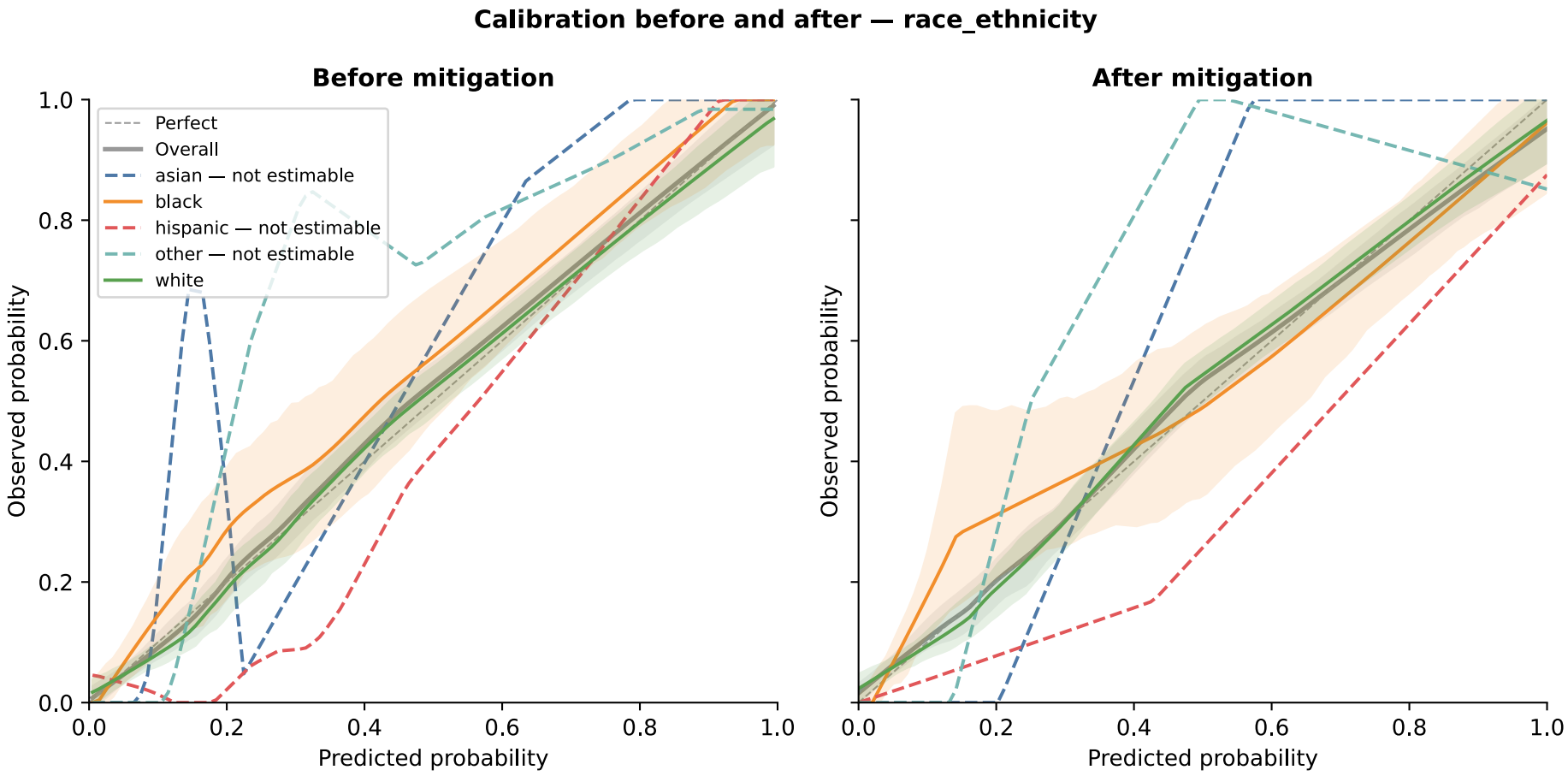
Subgroup	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
asian	1.107	2.336	0.184	0.3010	0.2098
black	0.379 (-0.024–0.682)	1.059	0.130 (0.103–0.155)	0.0533	0.0459
hispanic	-0.615	1.306	0.068	0.1092	0.0771
other	0.510	2.100	0.100	0.1481	0.1361
white	-0.033 (-0.198–0.103)	1.027 (0.919–1.128)	0.114 (0.105–0.125)	0.0225	0.0115

After

Subgroup	Intercept (95% CI)	Slope (95% CI)	Brier (95% CI)	ECE	ICI
asian	1.368	1.252	0.227	0.3263	0.1515
black	0.169 (-0.343–0.540)	0.802	0.131 (0.099–0.157)	0.0636	0.0469
hispanic	-1.678	0.184	0.080	0.1137	0.1005
other	0.583	0.133	0.087	0.0980	0.0953
white	0.090 (-0.080–0.226)	0.698 (0.499–0.926)	0.116 (0.107–0.127)	0.0163	0.0160

Calibration gaps

Metric	Before (95% CI)	After (95% CI)
max_intercept_gap	1.7220	3.0458
max_slope_gap	1.3091	1.1189



Discrimination (sanity check)

Per-subgroup recalibration is a monotone transform within each subgroup, so within-subgroup AUROC is invariant except for ties that isotonic regression introduces, and any sensitivity change is a threshold artefact. These columns barely moving is the expected result, not a null finding — a large shift here means the mitigation did something other than recalibrate.

Before

Subgroup	AUROC	Sensitivity
asian	0.850	0.400
black	0.879 (0.832–0.922)	0.732 (0.625–0.838)

After

Subgroup	AUROC	Sensitivity
asian	0.662	0.400
black	0.872 (0.824–0.916)	0.756 (0.649–0.851)

Subgroup	AUROC	Sensitivity
hispanic	0.909	0.900
other	0.971	0.778
white	0.873 (0.852–0.892)	0.749 (0.702–0.787)

Subgroup	AUROC	Sensitivity
hispanic	0.920	0.900
other	0.959	0.778
white	0.870 (0.850–0.889)	0.754 (0.710–0.791)

Appendix: TRIPOD+AI item coverage

Fairness-relevant items from the TRIPOD+AI checklist (Collins et al., 2024, *BMJ*). Items addressed by this audit are linked to the relevant report section. Items marked "User responsibility" require information that the audit tool cannot produce automatically.

Item	Description	Status	Detail
3c	Describe any known health inequalities between sociodemographic groups	*User responsibility*	User must provide context in manuscript introduction
5a	Describe data sources, rationale, and representativeness	*User responsibility*	User must describe data representativeness in manuscript methods
7	Describe data pre-processing, including whether similar across sociodemographic groups	*User responsibility*	User must verify preprocessing does not vary by group
8a	Outcome assessment consistency across sociodemographic groups	*User responsibility*	User must verify outcome labels are measured consistently
8b	Assessor demographics for outcome assessment	*User responsibility*	User must report assessor demographics
9c	Assessor demographics for predictor measurement	*User responsibility*	User must report assessor demographics
12e	All measures used to evaluate model performance	**Addressed**	Discrimination, Calibration, and Clinical utility sections
12f	Describe any model updating (eg recalibration) for sociodemographic groups or settings	**Addressed**	Mitigation section
14	Fairness approaches and rationale	**Addressed**	Audit summary and gap statistics throughout
20b	Report demographic characteristics and group differences	**Addressed**	Discrimination section (subgroup sizes, prevalences)
20c	Compare predictor distributions including demographics	*User responsibility*	User must report predictor distributions per group in manuscript
23a	Report performance estimates with CIs, including for key sociodemographic subgroups	**Addressed**	Discrimination, Calibration, and Clinical utility sections
25	Overall interpretation including fairness issues	*User responsibility*	User must interpret results in manuscript discussion
26	Limitations including non-representative sample, bias effects	*User responsibility*	User must discuss limitations in manuscript

Source: 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. doi: 10.1136/bmj-2023-078378. This checklist covers fairness-relevant items only; see the full statement for the complete checklist.

Appendix: How to read this report

Discrimination

Discrimination measures how well the model separates positive from negative cases.

AUROC (Area Under the ROC Curve)
Probability that the model ranks a random positive case higher than a random negative case. Range [0, 1]; 0.5 = no discrimination; 1.0 = perfect separation.

AUPRC (Area Under the Precision-Recall Curve)
Summarises precision across all recall levels. More informative than AUROC for low-prevalence outcomes. Higher is better; baseline equals the prevalence.

Sensitivity (True Positive Rate)
Proportion of true positive cases correctly identified at the chosen threshold. = TP / (TP + FN).

Specificity (True Negative Rate)
Proportion of true negative cases correctly identified at the chosen threshold. = TN / (TN + FP).

PPV (Positive Predictive Value)
Proportion of positive predictions that are truly positive. = TP / (TP + FP).

NPV (Negative Predictive Value)
Proportion of negative predictions that are truly negative. = TN / (TN + FN).

Calibration

Calibration measures whether predicted probabilities match observed event rates.

Calibration intercept
Intercept from logistic recalibration with slope fixed to 1 (Steyerberg 2009). Perfect calibration = 0. Positive values indicate the model underestimates risk on average; negative values indicate overestimation.

Calibration slope
Slope from logistic recalibration with free intercept and slope. Perfect calibration = 1. Values < 1 indicate overconfident predictions (predicted probabilities too extreme); values > 1 indicate underconfident predictions.

Brier score
Mean squared difference between predicted probabilities and binary outcomes. Range [0, 1]; lower is better.

ECE (Expected Calibration Error)
Weighted average of absolute calibration error across equal-frequency bins. Lower is better.

ICI (Integrated Calibration Index)
Mean absolute difference between LOESS-smoothed observed probabilities and predicted probabilities (Austin & Steyerberg 2019). Lower is better; more robust than ECE as it avoids binning artefacts.

Hosmer–Lemeshow statistic
Chi-square-distributed goodness-of-fit statistic comparing observed with expected events across deciles of predicted risk. A large value indicates lack of fit. **It scales with sample size**, so it is not comparable between subgroups of different size and must be read alongside *n*, the per-decile table, and the calibration measures above (Kramer & Zimmerman 2007).

Hosmer–Lemeshow degrees of freedom
Equal to the number of risk groups used. The classical convention subtracts 2, which is the penalty for estimating an intercept and slope on the data under test; this package evaluates predictions supplied by a model fitted elsewhere, so no degrees of freedom are spent and none are subtracted.

Decile slope and intercept
Least-squares regression of the observed on the expected event rate across risk groups, recommended by Kramer & Zimmerman (2007) as an adjunct to the statistic. Perfect calibration gives a slope of 1 and an intercept of 0.

Clinical utility (decision curves)

Decision curve analysis quantifies the clinical value of using the model at a given decision threshold.

Net benefit
TP/n − FP/n × $p_t/(1 - p_t)$, where p_t is the decision threshold. Higher is better. A model is clinically useful when its net benefit exceeds both 0 (treat none) and the treat-all reference.

Standardised net benefit
Net benefit divided by prevalence (Naderalvojoud 2025). Normalised to a [0, 1] scale for comparison across cohorts with different prevalences.

Treat-all reference
Net benefit if all patients receive treatment. = prevalence − (1 − prevalence) × $p_t/(1 - p_t)$. Becomes strongly negative at high thresholds.

Treat-none reference
Net benefit = 0. The baseline of no intervention.

Gap statistics

Gap statistics quantify the largest disparity between subgroups for each metric. All gaps are non-negative; 0 = perfect parity.

Equal opportunity difference
max − min of sensitivity across subgroups. Measures fairness in true positive identification.

Predictive equality
max − min of false positive rate (1 − specificity) across subgroups. Measures fairness in false alarm rates.

Equalized odds gap
Maximum of equal opportunity difference and predictive equality.

Statistical parity
max − min of positive prediction rate across subgroups. Measures whether the model predicts positive at equal rates.

Max intercept / slope gap
Maximum pairwise difference in calibration intercept or slope across subgroups, with bootstrap confidence intervals.

Max net benefit gap
Maximum across all thresholds of the maximum pairwise net-benefit difference between subgroups, with bootstrap confidence intervals.

Confidence intervals and shrinkage

Bootstrap confidence intervals
95% percentile intervals from subgroup-stratified patient-level resamples. Patients are resampled with replacement within each subgroup, preserving the subgroup's size while letting its event count vary.

Empirical Bayes shrinkage
Interaction cells with fewer events than the configured threshold (shown in italics) use Beta-Binomial shrinkage toward the marginal subgroup rate. This reduces instability in small cells. Confidence intervals are suppressed for shrunk estimates.

Generated by **isitfair**. Three-axis evaluation: discrimination, calibration, and clinical utility reported jointly per TRIPOD+AI recommendations.

Bootstrap confidence intervals (95%) from 200 subgroup-stratified patient-level resamples.