

Subgroup AUROC-gap tests are not evidence of unfairness in clinical prediction models: case-mix differences make the equal-discrimination null false

Rohith Reddy Bellibatlu

Florida International University

Manpreet Singh

Boston University

Yash Jajoo

New York University

Abhishek Israni

Boston University

Rahul Joshi

rahulj@sitpune.edu.in

Symbiosis International University

Research Article

Keywords: subgroup analysis, case mix, discrimination, AUROC, algorithmic fairness, clinical prediction model, estimand, permutation test

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10718552/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Subgroup AUROC-gap tests are not evidence of
unfairness in clinical prediction models: case-mix
differences make the equal-discrimination null false

Rohith Reddy Bellibatlu, ^{1†}, Manpreet Singh, ^{2†}, Yash Jajoo,
^{3†}, Abhishek Israni⁴, Rahul Joshi, ^{5*}

¹Florida International University, Miami, FL, USA.

²Boston University, Boston, MA, USA.

³New York University, New York, NY, USA.

⁴School of Public Health, Boston University, Boston, MA, USA.

⁵Symbiosis Institute of Technology, Pune, Symbiosis International
(Deemed University), Pune, India.

*Corresponding author(s). E-mail(s): rahulj@sitpune.edu.in;

Contributing authors: rohithreddybc@gmail.com; manni@bu.edu;
yj1499@nyu.edu; abhi1097@bu.edu;

[†]These authors contributed equally to this work.

Abstract

Background. Fairness audits of clinical prediction models reduce subgroup AUROC to a cut-point, ratio-rule or permutation comparison, all testing one null: equal true AUROC in every subgroup. That null is false for a perfectly fair model whenever case mix differs, and we asked by how much.

Methods. Seven procedures ran on common simulated data at $\alpha = 0.05$: a fixed cut-point, the four-fifths rule, raw and studentized permutation tests, and three readings of a double-corrected variance estimator. Case-mix cells made one shared model *be* the data-generating probability, differing across subgroups only in predictor spread, swept across 16 values and measured in 21 demographic partitions of five clinical cohorts. A closed form turns each partition's own geometry, not the ratio, into the gap a fair model would show; we refitted every cohort 24 times and built a noise-only reference. A subgroup Cox calibration test ran on

the same sweep and on matched pairs varying only the mechanism behind a fixed gap.

Results. The raw permutation test’s false-flag rate for a perfectly fair model rises from 0.056 at $\rho = 1.0$ to 0.918 at 2.0 and 1.000 at 3.0. Studentizing tracks it to within 0.008; raising n from 2,000 to 10,000 at a fixed true gap drives five procedures to 1.000. Measured $\hat{\rho}$ has median 1.145; across 24 refits the median runs 1.132–1.223 and the maximum 1.999–6.388. The implied case-mix gap is no more stable: its median runs 0.013–0.030, its maximum 0.131–0.219 and the count reaching the 0.05 disparity convention 2–7 of 21, so we quote no point value; 9 of 21 exceed a noise-only reference. Where scores are true conditional probabilities, the calibration test held nominal (mean 0.0499) while separating mechanism at fixed true gaps (0.049 vs 0.078, 0.051 vs 0.293, 0.047 vs 0.502), which no AUROC-gap procedure did. We could not decompose observed gaps.

Conclusions. Subgroup AUROC-gap tests should not be used as fairness evidence for clinical prediction models. A gap that cannot be decomposed cannot be interpreted. Report subgroup AUROC descriptively beside the case-mix distribution, with subgroup calibration and net benefit: existing advice we measured rather than invented.

Keywords: subgroup analysis, case mix, discrimination, AUROC, algorithmic fairness, clinical prediction model, estimand, permutation test

1 Background

Disaggregated evaluation, reporting a model’s performance separately within demographic subgroups, is now standard in clinical prediction model reporting, written into reporting guidance [1, 2], into widely used auditing toolkits [3–5] and into regulatory-facing testing plans [6]. In practice the disaggregated numbers are almost always reduced to one summary before a judgement is made, and two reductions dominate: the maximum-minus-minimum range of subgroup AUROC compared against a fixed cut-point, usually 0.05; and the four-fifths ratio rule, imported from employment-discrimination regulation [7], which flags a model when the smallest subgroup value falls below 80% of the largest. Both are deterministic screens with no reference distribution and no stated error rate. A third family adds inference to the same summary, by permutation [8] or by correcting the variance of the subgroup estimates [9].

All of them share one thing, and it is the thing this paper is about. Whatever the statistic and whatever the calibration, the null hypothesis being tested is that true AUROC is the same in every subgroup, and that null is treated as synonymous with the model being fair. It is not.

What is already known

The statistical shortcomings of the individual procedures are well documented and none is our discovery. The four-fifths rule’s false-positive behaviour is roughly forty-seven years old: Greenberg [10] showed that the rule “in many situations can lead to large statistical errors of both kinds,” Boardman [11] computed its exact Type I and Type II error probabilities in the same volume, Morris and Lobsenz [12] supplied a significance test for the impact ratio, Collins and Morris [13] compared the Type I error and power of competing significance tests for adverse impact specifically in the small-sample regime, and Besse et al. [14] derived the asymptotic distribution of the disparate-impact indexes themselves so that confidence intervals, rather than a single point estimate, can accompany a disparate-impact judgement. Roth et al. [15] re-confirmed by simulation that it “often resulted in false-positive readings of adverse impact even when there were no underlying (population) standardized group differences.” Watkins and Chen [16] traced how a single clause of disparate-impact law was abstracted into a metric and codified into machine-learning fairness toolkits, which is how it reached the audits studied here.

Studentized permutation inference for the AUC is likewise established. Janssen [17] proved that studentized permutation statistics are asymptotically exact under extended, non-exchangeable nulls, the generalized Behrens–Fisher problem, with the nonparametric counterpart in Neubert and Brunner [18], the general theory including maxima of studentized absolute differences in Chung and Romano [19, 20], DiCiccio

and Romano [21], and, most directly, permutation-based inference for the AUC specifically in Pauly et al. [22]. Any result we report about studentization confirms that literature rather than adding to it, and inherits its open problems: Munko and Sattler [23] show that standard studentization does not deliver the correct *joint* limiting distribution in multiple-testing settings, a live caveat on any maximum-over-partitions construction including ours.

The estimation-side literature is equally clear. Lum et al. [9] showed that common group-wise disparity measures are biased estimators of what they represent, worst when categories have few members, and proposed a double-corrected variance estimator; the analogous bias in binned calibration-error estimators is established [24, 25]; subgroup-specific tests in randomised trials reject at between 7% and 66% under a true null [26]; and apparent subgroup disparities can be artefacts of training stochasticity, small subgroups, or the instability of AUROC across resamples [27–29]. Evaluation choices have been framed as “fairness hacking” [30] and their multiplicity addressed [31], while concurrent work continues to build tests of the same null [32, 33]. The chain linking a rank correlation to the Mann–Whitney statistic, and hence to AUROC, has four distinct links. Cureton [34] introduced the rank-biserial correlation coefficient and proved its equivalence to Kendall’s tau and Spearman’s rho, without reference to the Mann–Whitney statistic; Glass [35] proved his simplified formula algebraically equivalent to Cureton’s; Wendt [36] gave the simplified, U-based form that is the rank-biserial/U identity itself; and Willson [37] showed that the coefficient is a linear function of U in order to supply distribution theory and critical values, a contribution distinct from the identity itself. The corresponding AUC identity is older than the source usually given for it. Bamber [38] is routinely credited with $AUC = U/(mn)$, including in earlier versions of this manuscript, but Bamber’s own abstract presents that relation as established material under review and claims as new the systematic treatment: asymptotic normality, variance formulas, competing confidence-interval

constructions, and the maximum variance of the area estimator over convex ordinal
 dominance graphs. The identity itself predates him. Green [39] states the equivalence
 between the area under a yes–no ROC curve and forced-choice proportion correct, and
 Birnbaum and Klose [40] were already deriving bounds on the variance of the Mann–
 Whitney statistic in that role. We therefore cite Bamber for the systematic treatment
 and the variance methods, Green and Birnbaum–Klose for priority on the identity, and
 Hanley and McNeil [41] for the clinical statement, that the area equals the probability
 a randomly chosen diseased subject is ranked above a randomly chosen non-diseased
 one, and that this is what the Wilcoxon statistic estimates, which is the form in which
 the identity actually reached this literature. The same U statistic underlies Somers’
 D, which is itself an affine function of AUROC [42, 43].

The objection that has already been made in this journal’s field

Set against all of that is a separate literature that has been saying, for a decade, that
 subgroup discrimination is not a like-for-like quantity. Discrimination measures how
 well a model separates the cases it is shown, so a subgroup whose predictor values are
 more spread out is easier to separate, and a single model with one coefficient vector
 therefore *expects* unequal subgroup AUROC whenever covariate distributions differ.
 A cluster-specific c-index varies substantially across centres for one model [44]; model-
 based concordance was introduced to quantify how much of a c-statistic change is
 attributable to case mix rather than to the model [45]; across 224 external validations,
 c-statistic variation is mainly associated with case mix [46]; CHARGE-AF discrimi-
 nation falls from 0.72 to 0.57 across age strata within one model [47]; AUC varies by
 10–20% between subpopulations, higher for subgroups whose scores are more spread
 out [48]; correlation among predictors shifts AUC in both directions depending on the
 sign of the effects it correlates [49]; and discrimination and calibration respond differ-
 ently to case-mix shift depending on the causal direction of prediction [50]. Closest of

231 all, Nieboer et al. [51] found that a permutation test on the c-statistic “may provide
232 misleading results in the presence of case-mix differences.”
233

234 Pfohl et al. [52] state the general form of the objection for algorithmic fairness,
235
236 arguing through causal graphical models that equal performance across subgroups is
237 an unreliable fairness target even when the data are representative and the null is
238 correctly specified. That paper has been widely cited as a caution and rarely acted on.
239
240 Petersen et al. [53] make the parallel argument that apparent underperformance in a
241 subgroup can reflect an irreducibly harder prediction task rather than a remediable
242 model deficiency, and Jones et al. [54] give a causal taxonomy of the dataset-bias
243 mechanisms that produce subgroup performance gaps in medical imaging specifically.
244
245 This paper is the empirical demonstration of Pfohl et al. [52]’s central point in the case
246 that dominates practice, and it goes one step further: the failure requires no selection
247 bias, no unrepresentative data and no misspecification, only that subgroups differ in
248 case mix. Subgroups do differ in case mix, by more than sampling noise alone would
249 produce in 9 of the 21 clinical partitions measured below, and by amounts that span
250
251 a wide range, and it is the amount, not the fact, that decides how badly an audit is
252 misled.
253
254
255
256
257
258
259

260 **The impossibility results, and what they cost the alternative**

261
262 A separate and older literature bears directly on where this paper ends up, and it is
263 not comfortable for the direction we point in. Chouldechova [55] and Kleinberg et al.
264 [56] showed that calibration within groups, balance for the positive class and balance
265 for the negative class cannot hold simultaneously unless base rates are equal or the
266 predictor is perfect. Subgroup prevalence is essentially never equal in a clinical cohort,
267
268 so any audit that asks for calibration *and* equal error rates is asking for something
269 arithmetically unavailable. That is a reason the equal-AUROC target is doomed on
270
271 its own terms, and it is prior to anything we measure.
272
273
274
275
276

Pleiss et al. [57] sharpen it in the direction that costs us most. They show that calibration is compatible with only one of the equal-error-rate constraints, and, the part a paper recommending calibration has to confront, that satisfying the relaxed criterion is achieved by withholding information from an existing classifier’s predictions, so it is no better than randomising a fraction of them. Calibration is cheap to satisfy and weak as a fairness criterion: a model can be perfectly calibrated in every subgroup and still be useless, or harmful, in one of them. We take this as a constraint on what our Section 3.4 result can mean. Showing that a subgroup calibration test does not inherit the case-mix failure is not showing that calibration is a sufficient fairness criterion, and the impossibility results say it cannot be one. What survives is comparative and narrow, and we keep it that way throughout.

What this paper adds

We take one shared model, with identical coefficients, which *is* the exact data-generating probability, Bayes-optimal and perfectly calibrated in every subgroup, treating no subgroup differently in any respect, and let subgroups differ only in the distribution of the predictor. Every flag raised on such data is a false alarm about fairness; we measure how often seven procedures raise one.

Five things here are, to our knowledge, new. First, a head-to-head Type I error comparison of these seven procedures on common simulated nulls; the individual procedures have been analysed, but not against each other on the same data. Second, a sensitivity analysis for the subgroup-admissibility rule specifically, the minimum-size threshold below which a subgroup is dropped, which Simson et al. [30] anticipate as an instance of the general multiverse problem but do not study. Third, quantification of the case-mix failure across seven auditing procedures at once, which localises it in the estimand rather than in any one test; the two closest antecedents do something different, since Pfohl et al. [52] argue the point theoretically without quantifying it

323 across procedures, and Nieboer et al. [51] quantify it for a single permutation test in an
324 external-validation setting rather than a subgroup fairness audit. Fourth, an AUROC-
325 specific argument that a ratio scale is meaningless for a quantity whose uninformative
326 point is 0.5, which is why the four-fifths rule transfers particularly badly here. Fifth,
327 and this is the part that makes the rest usable, the case-mix magnitude is not assumed
328 but measured, in every demographic partition of our own clinical cohorts, and each
329 partition’s *own* measured predictor geometry is then evaluated in closed form, so the
330 output is the subgroup AUROC gap a perfectly fair model would produce on that
331 partition rather than a rate borrowed from a chosen simulation geometry. We claim
332 novelty for nothing else: not for studentization, not for small subgroups inflating noise,
333 not for the four-fifths rule’s false positives, and not for the case-mix objection itself.

341 One further result is a measurement of other people’s recommendations rather
342 than a contribution of ours, and we are explicit about the distinction. We put a
343 subgroup calibration test and model-based concordance [45, 58, 59] through the same
344 sweep and the same matched-pair positive control as the seven procedures, and report
345 their operating characteristics under case mix (Section 3.4). We did not devise either
346 quantity, we do not propose them as fairness tests, and the finding is comparative and
347 bounded to the data-generating process under which it was measured.

353 We also report a negative result. We attempted to decompose our cohorts’ observed
354 gaps into a case-mix component and a genuine residual, applying van Klaveren et al.
355 [45]’s existing estimator, and could not make the decomposition stable enough to
356 report (Section 3.2). We report the attempt and its failure because a gap that cannot
357 be decomposed is a gap that cannot be interpreted, which is the paper’s thesis rather
358 than an exception to it.

Disclosure: five things we got wrong	369
	370
We state these at the front, once, and do not return to them. The reader is entitled	371
to know that the authors have been wrong about this evidence base five times, and	372
by how much.	373
	374
<i>One. The raw range test was reported as well behaved.</i> Our first analysis concluded	375
that no clinical cohort produced a subgroup gap distinguishable from its own permu-	376
tation null, and framed that as a negative result about a class of published findings.	377
It rested entirely on the raw max–min range permutation test under one admissibility	378
rule, with no measurement of that procedure’s operating characteristics.	379
	380
<i>Two. The studentized test was recommended.</i> A later analysis of the same data	381
reported the studentized permutation test as uniquely well behaved, nominal under	382
both nulls and stable across rules, and recommended it. Two things were wrong. Sta-	383
bility was measured by counting flagged cohorts, and a count is invariant to which	384
cohorts are in the set; measured by identity, the studentized test is among the least	385
stable procedures in the comparison (Table 5). And the recommendation rested on	386
calibration under an equal-AUROC null, which Table 2 shows is not the null that mat-	387
ters. Neither analysis was released, no preprint, no repository, no conference version,	388
and this manuscript supersedes both.	389
	390
<i>Three. A case-mix magnitude was assumed rather than measured.</i> An earlier draft	391
of this paper headlined a false-flag rate of 0.918, obtained at a per-level linear-predictor	392
standard-deviation ratio of $\rho = 2.0$, and described that ratio as a moderate and	393
clinically ordinary amount of case-mix heterogeneity. That description had no citation	394
and no measurement behind it; it was an assumption wearing the clothes of a finding.	395
The quantity was directly measurable in the ten fitted models we already had, and	396
when measured its median is 1.145 across 21 clinical partitions, with exactly one	397
reaching 2.0 (Section 3.1). The correction cost the paper its single dramatic number	398
and returned a distribution with a predictable structure, which is a better result, but	399
	400
	401
	402
	403
	404
	405
	406
	407
	408
	409
	410
	411
	412
	413
	414

415 we did not know that when we went looking, and it is a smaller finding at the median
416 than the one previously claimed.

418 *Four. The attributable fraction was computed correctly and measured its own*
419 *machinery.* An earlier version reported that case mix explains a median **93%** of
420 observed subgroup AUROC gaps; a subsequent repair reported **29%**. **Both figures**
421 **are withdrawn, and so is the attributable fraction as a reported quantity.**

422 This error is of a different kind from the first three. The number was not assumed.
423 It was computed, from code that ran, on the right data. It was wrong because the
424 estimator was near-tautological: its numerator had been Cox-recalibrated within each
425 level, and because Cox recalibration is monotone it leaves the level's observed AUROC
426 exactly unchanged while forcing calibration, so model-based concordance of the recal-
427 ibrated predictions reproduces the observed AUROC almost exactly, at a median
428 absolute difference of 0.0030 and a correlation of 0.9956. An estimator can be imple-
429 mented correctly, tested and reproducible, and still be measuring itself; nothing about
430 the number's provenance would have revealed that. What did was asking what the
431 quantity would equal if case mix played no part at all. That much is a clean method-
432 ological finding and we keep it (Section 3.2). The 29% that replaced it is withdrawn
433 for a different reason again, and the 93% is not recoverable from the current result
434 files, because the estimator that produced it no longer exists: the fraction's denomi-
435 nator is an observed gap frequently near zero at these sample sizes, so the ratio is not
436 estimable here, and no repair of the estimator changes that. The first error produced
437 a wrong number. The second showed there was no number to produce. The residual
438 gap, which is a difference rather than a ratio and does not inherit that pathology, we
439 do report.

440 *Five. The implied case-mix gap was reported from one fit.* The version before this
441 one withdrew the maximum $\hat{\rho}$ as unstable across refits and kept its downstream prod-
442 uct, the implied case-mix gap, at full precision: median 0.022, maximum 0.173, and
443
444

5 of 21 partitions reaching the 0.05 convention. That was not defensible, because the implied gap is computed from the same fitted linear predictor. Propagated through the same 24 specifications, its median runs 0.013 to 0.030, its maximum 0.131 to 0.219, and the count reaching 0.05 runs 2 to 7. **All three point values are withdrawn.** Two further checks that should have been run before, and now are, cut in the same direction: the implied gap has a noise floor, because it is a maximum minus a minimum, and only 9 of the 21 partitions clear it; and the earlier text asserted that Eq. (2) assumes a Gaussian linear predictor while also asserting that it makes no distributional assumption, which cannot both be true. The identity is distribution free. The two-parameter plug-in is what was Gaussian, and its cost is now measured rather than assumed away.

The common thread in the first three is not test choice. Each time, a quantity that could have been measured was instead assumed: that the raw permutation test was well behaved, that count-based stability was stability, that $\rho = 2.0$ was ordinary. Each assumption was load-bearing. Pre-registering any of those analyses would have preserved its error intact. What caught all five was recomputing the thing itself, checking an estimator against a case where its answer was known in advance, and refusing to let a withdrawn quantity keep its downstream product.

2 Methods

2.1 Setting, statistic and admissibility

A held-out test set of n patients carries model scores s_i , binary outcomes y_i , and demographic partitions $\mathcal{C}$ (age band, sex, race and ethnicity, insurance status). A partition c divides the cohort into disjoint levels g ; $\hat{A}_g$ is the empirical AUROC in

507 level g . The statistic reported in current practice is

508

509

510

511

512

513

514

515

516

$$\Delta_c = \max_{g \in c} \hat{A}_g - \min_{g \in c} \hat{A}_g, \quad \Delta = \max_{c \in \mathcal{C}} \Delta_c, \quad (1)$$

the maximum over partitions of the within-partition max-min range, taken over *admissible* levels only.

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

A level is admissible if it meets a size criterion and contains **at least two positive and two negative** outcomes; a partition contributes to Eq. (1) only if at least two of its levels are admissible. Five size criteria are swept: at least 20, 30, 50 or 100 rows (`m20`, `m30`, `m50`, `m100`), and at least 10 observations in each outcome class (`ev10`). `m30` is the reference throughout, being the commonest choice in practice. The two-of-each floor is what the code enforces (`null_reference.py`, `_rule_admits`) and is the minimum for an estimable within-level AUROC; earlier versions of this manuscript stated it as one of each, which understated the rule. The distinction is not cosmetic. Of the three levels excluded here for degenerate labels, two carry no positives at all (Diabetes 130 `age_group=[0-10]`, $n = 31$; BRFSS 2024 `age_group=18-24`, $n = 44$) and would fail either rule, but the third, the synthetic cohort's `age_group=18-44`, $n = 371$, carries *one* positive and 370 negatives. Its AUROC is computable and useless; it is excluded by the two-of-each floor and would have been admitted under the rule as previously described.

541 2.2 The seven procedures

All seven consume the same frozen (s, y) and the same partition geometry under the same admissibility rule, so they differ in the statistic and its inference, not in the inputs. Subgroup AUROC variances are DeLong's structural-component variances throughout [60].

(1) *Fixed 0.05 cut-point.* Flag the cohort if $\Delta \geq 0.05$. No reference distribution.

(2) *Four-fifths ratio rule.* Flag the cohort if $\min_g \hat{A}_g / \max_g \hat{A}_g < 0.80$ in the worst partition [7]. Computed with `fairlearn` 0.13.0 (`MetricFrame.ratio, method="between_groups"`) [4]. No reference distribution.

(3) *Raw max-min range permutation test.* Δ from Eq. (1), calibrated against a permutation null at the cohort's own geometry. Conditional on (s, y) , subgroup labels are permuted *within outcome classes*, positives among positives, negatives among negatives, which holds every level's size and prevalence fixed while forcing equal true AUROC across levels. The statistic is recomputed under the same admissibility rule on each draw and a one-sided permutation p -value is reported [61]. Nothing is refitted, so the statement concerns a given fitted model on a given split.

(4) *Studentized permutation test* [8]. The statistic is the maximum over admissible level pairs of $|\hat{A}_i - \hat{A}_j| / \sqrt{\hat{V}_i + \hat{V}_j}$, with inference from the same stratified permutation draws, seed and B as procedure (3), combined by single-step max- T over pairs sharing one permutation pass [62]. Two adaptations of the published method are stated explicitly. The permutation is stratified within outcome class rather than free, because a free permutation resamples each level's event count and leaves many subgroup AUROCs unestimable; stratification also makes the draws identical to those of procedure (3), so the two differ in the statistic alone. And the published two-group test is extended to many levels by max- T . A variance floor of 10^{-12} is applied: a level pair whose summed DeLong variance falls at or below it carries no estimable sampling variability, so the pair is dropped from the max- T family rather than given an infinite statistic. This changes the family over which the familywise rate is controlled, so it is counted rather than merely handled. No pair is dropped for the ten cohorts' observed statistics; the floor does bind, however, among the permutation replicates that build the reference distribution, in six (cohort $\times$ rule) cells: NHANES 2021–2023 under `m20`, `m30` and `m50` (33 of 180,000 pair-evaluations dropped in each) and under `ev10` (8 of 100,000), and NHIS 2023 under `m20` and `m30` (8 of 290,000 each). Because the max- T family defines

the reference distribution against which the observed statistic is judged, these drops belong in this accounting even though they never touch the ten cohorts' own statistics. (5-7) *Double-corrected variance* [9]. The estimator $V_{dc} = S^2 - \bar{v}_k$ subtracts the mean sampling variance of the subgroup AUROCs from their empirical variance, per partition. Because the published estimator targets a variance while current practice reports a range, three inferential readings of it are evaluated: (5) a closed-form z -test on V_{dc} ; (6) Cochran's Q on the same inputs [63], which weights levels by precision rather than equally; and (7) a parametric bootstrap interval on V_{dc} . **Procedures (4) to (7), and both alternatives of Section 2.4, are our implementations.** No published package implements either method for this setting. Each carries a stated adaptation: the published two-group studentized test is extended to many levels by single-step max- T , the permutation is stratified within outcome class rather than free, and the variance floor removes unestimable pairs from the max- T family. Every result we report for these procedures is therefore a property of our implementation and not of the original authors' claims, and we do not repeat this caveat at each table; it applies throughout. Procedures (1) and (2) state no error rate, so their rejection frequencies are reported throughout as *implied false-flag rates* rather than as Type I error against a claimed level.

2.3 Three null families

Type I error was estimated at nominal $\alpha = 0.05$ from 1,000 independent simulations per cell with $B = 999$ permutation replicates within each, every geometry run under both the `m30` and `ev10` rules. Datasets are a pure function of (geometry, replicate, seed); the geometry's seed word is a CRC-32 digest of its name rather than Python's salted string hash, and `PYTHONHASHSEED` is pinned for the whole process tree, so every cell is reproducible bit for bit.

Simple null (6 geometries, 12 cells).

Subgroup membership is assigned independently of (s, y) , so subgroup score distributions and true AUROC are identical by construction and full exchangeability holds. The six geometries vary the number of levels (3 to 10), how unequal they are (balanced through to sizes 1,100/500/200/140/60), the number of partitions the maximum runs over (1 to 5), and prevalence (down to 0.075). Unless stated otherwise $n = 2,000$, prevalence 0.20, true AUROC 0.75.

Composite equal-AUROC null (11 geometries, 22 cells).

Every level has the same true AUROC but a different score distribution, produced by a level-specific strictly monotone map applied to the scores within that level. AUROC is rank-based, so such a map leaves each level's true AUROC exactly unchanged while the score distributions diverge; exchangeability fails and an unstudentized permutation test has no validity guarantee. This is the null studentization is designed for. The set spans three transform families ($s \mapsto s^a$, $s \mapsto \text{expit}(a \logit s)$, and a two-segment piecewise-linear map), 1, 3 and 5 partitions, $n \in \{500, 2000, 10000\}$ and prevalence $\in \{0.05, 0.20, 0.50\}$. Each dataset is checked for transform-induced ties, which would silently move the true AUROC and turn a null cell into a power calculation. What this family does *not* cover is unequal true AUROC, which is the point of the third. The simple and composite families together give the 34 cells of Table 4.

Case-mix null (6 geometries, and a 16-point sweep).

Here the true subgroup AUROCs are deliberately unequal and the model is nevertheless perfectly fair. A single linear predictor $L = b_0 + x$ is drawn with a per-level mean and standard deviation; the outcome is $y \sim \text{Bernoulli}(\text{expit}(L))$; and the model score *is* $\text{expit}(L)$. The model is therefore the exact data-generating probability: Bayes-optimal, perfectly calibrated in every subgroup, one coefficient vector for everybody, no term distinguishing subgroups. The intercept b_0 is solved by quadrature so the

level-weighted prevalence hits its target exactly, and each level’s true AUROC is computed by quadrature rather than simulated, so the reported gaps carry no Monte Carlo error; this quadrature step exploits the same closed-form relation between AUC and the linear predictor’s distribution that Austin and Steyerberg [64] give for a single population.

The whole of a spread-driven geometry is one number: ρ , the ratio of the largest to the smallest per-level standard deviation of the linear predictor. Because it is one number, it can be swept, and it can be measured in real data; both are done here. The sweep holds everything else fixed, three levels of equal size, $n = 2,000$, prevalence 0.20, per-level standard deviations $\rho^{-1/2}/1/\rho^{1/2}$ rescaled to unit geometric mean, and moves ρ across 16 values from 1.0 to 3.167, at 1,000 simulations and $B = 999$ per point under `m30`. The true AUROC gap it induces rises from 0 at $\rho = 1$ to 0.198 at $\rho = 3.167$, computed by quadrature. A sensitivity sweep additionally forces every level to the same true prevalence, which changes the curve little (raw max–min permutation 0.941 versus 0.918 at $\rho = 2.0$), so the effect is not a prevalence artefact.

The six named geometries are listed in Table 2 and reported at `m30`. Their ρ values are 1.0 (location only), 1.222 (mild), 2.0 (moderate) and 3.167 (strong). They are not the same runs as the sweep: they carry different geometry seed words, and at a given ρ their per-level standard deviations differ from the sweep’s unit-geometric-mean parameterisation (mild is 0.9/1.0/1.1, the sweep’s $\rho = 1.222$ point is 0.905/1.000/1.105). Where the two meet they therefore agree only to within Monte Carlo error: raw max–min permutation 0.918 versus 0.920 at $\rho = 2.0$ and 0.141 versus 0.162 at $\rho = 1.222$, against a Monte Carlo standard error near 0.011 on each. Both are reported; neither is a replication of the other. Level size is not the operative constraint for five of the six geometries, but it is for `casemix_location_3`: its level 0 sits at a true prevalence of 0.022 (about 15 of 668 rows are events). The dropped level is the one with the *highest* true AUROC, i.e. the level generating the gap, so `m30` and `ev10` do not agree on this

geometry (`fixed_threshold_005`: 0.642 vs. 0.589; raw max–min permutation: 0.226 vs. 0.214).

Monte Carlo precision.

With 1,000 simulations the Monte Carlo standard error of a rate near 0.05 is about 0.0069; across all cells its median was 0.00686 and its maximum 0.01581. Because Table 4 reports a *maximum* over 34 cells, the comparison for the worst cell is not 0.05 but the 95th percentile of the maximum of 34 such rates from a correctly sized procedure. Treating the 34 cells as independent gives 0.072. They are not independent: each of the 17 geometries contributes one cell under `m30` and one under `ev10`, computed on the *same* simulated datasets, so the two are positively correlated. Modelling that pairing, 17 independent geometries, with the `m30` and `ev10` cells within a geometry treated as perfectly correlated, gives 0.0700 instead. A worst cell inside the paired-structure band is not evidence of miscalibration and we do not treat it as such.

2.4 Two calibration-based alternatives, and a positive control

Existing guidance recommends reporting subgroup calibration and clinical utility instead of a discrimination gap [58, 59], and van Klaveren et al. [45] give model-based concordance as the case-mix-corrected benchmark for discrimination. Neither recommendation had been measured against the failure mode documented here. We therefore ran two calibration-based quantities through the same simulation harness, on the same replicates and the same admissible levels as the seven procedures.

(A) *Subgroup Cox calibration test.* Within each admissible level, the two-parameter recalibration $\text{logit } \Pr(y = 1) = a + b \text{logit } s$ is fitted by Newton–Raphson and tested against $(a, b) = (0, 1)$ by a likelihood ratio statistic on 2 degrees of freedom against the offset model $\text{logit } \pi = \text{logit } s$. This is the standard two-parameter recalibration assessment, applied within subgroup. Per-level p -values are Holm-adjusted across the levels of the partition [65] and the cohort is flagged if the smallest adjusted p falls

below $\alpha = 0.05$. This is a test, with a reference distribution and a nominal level, so its rejection rates are comparable with those of procedures (1)–(7).

This is a test of *weak* calibration, and that is the weakest rung that counts as calibration at all. In Van Calster et al. [66]’s hierarchy, a correct calibration intercept and slope, the two parameters this test examines, is weak calibration; moderate calibration, the property that an event rate of $R\%$ obtains among patients with predicted risk $R\%$, is strictly stronger, and strong calibration is stronger again and unattainable in practice. A model can pass a two-parameter test and still be badly miscalibrated over regions of the risk scale, so nothing below should be read as evidence that a model passing this test is calibrated in the sense that matters clinically. We use the two-parameter form because it is what practice uses and because it has a reference distribution; it is a lower bound on what a calibration assessment should be.

(B) *Model-based-concordance excess*. For each admissible level we compute model-based concordance from the model’s own predicted risks [45], the concordance the model *implies* for the case mix it was applied to, if the model were correct, and subtract the level’s observed AUROC. The reported quantity is the maximum of that difference over levels. **This is a descriptive statistic in AUROC units, not a test.** It has no threshold, no reference distribution and no error rate in the simulation, and it is reported as such throughout; the summary we give for it is a mean over replicates, never a rejection rate.

Neither quantity is proposed here and neither is new: (A) is standard calibration assessment applied by subgroup, (B) is van Klaveren et al. [45] applied by subgroup. We measured two existing recommendations; we did not invent a procedure. Both assume that within-subgroup calibration is achievable and checkable, which in a level carrying a few dozen events it is not, and that assumption is a real limit on both.

<i>Positive control: matched case-mix and unfairness arms.</i>	829
	830
A procedure that never fires also has a nominal false-flag rate, so a low rate under	831
case mix is worthless on its own. We therefore built matched pairs in which the <i>same</i>	832
true subgroup AUROC gap is produced by two different mechanisms. In the case-	833
mix arm, level 0's linear predictor is compressed relative to the others and the model	834
remains the exact data-generating probability, so the gap is entirely case mix and	835
the excess over what the model implies is exactly zero. In the unfairness arm, level	836
0's outcome is partly driven by a covariate the model does not use (weight w), so	837
the model genuinely delivers less discrimination in that level than its own predictions	838
imply, by exactly the size of the gap. The two arms are constructed to agree on true	839
AUROC by level to four decimal places at each of three gap sizes (0.050, 0.123, 0.200).	840
A third geometry combines both mechanisms at a total gap of 0.123, of which 0.076 is	841
genuine excess. Two further geometries make the model unfair with <i>no</i> AUROC gap	842
at all: a subgroup-specific calibration slope of 2.0, and a subgroup-specific intercept	843
shift of one logit. All are run at 1,000 simulations, $B = 999$, m30.	844
	845
	846
	847
	848
	849
	850
	851
	852
	853
	854
2.5 An attempted decomposition of the observed gaps, and	855
how it was checked	856
	857
We attempted to estimate, for every partition of every cohort, how much of the	858
<i>observed</i> subgroup AUROC gap the model's own case mix accounts for. The extreme	859
level pair is selected once on the observed data and held fixed; model-based con-	860
cordance is computed for each of the two levels from the raw predicted risks, and	861
their difference is the case-mix-implied gap. Two summaries follow: the residual gap ,	862
observed minus implied, in AUROC units; and the case-mix fraction , implied divided	863
by observed. Intervals are percentile bootstrap, $B = 2,000$, resampling rows within	864
(level, outcome) and refitting every step of the estimator inside each replicate, with	865
seeds pinned per (cohort, rule, partition).	866
	867
	868
	869
	870
	871
	872
	873
	874

875 *The unit of analysis is 33 distinct partitions, not 64.*

876
877 The estimator runs on (cohort $\times$ rule $\times$ partition) combinations, of which there are
878 33 under m30 and 31 under ev10. Those are not 64 independent quantities. The ev10
879 set is a subset of the m30 set, so the union is **33 distinct cohort $\times$ partition**
880 **cells**, and on the 31 cells present under both rules the two measurements of the case-
881 mix attributable fraction, the decomposition's own estimator, correlate at $r = 0.997$.
882 Counting them as 64 would double-count 31 cells: exactly the double-count we refused
883 when constructing the reference band of Section 2.3, where 34 cells are treated as
884 17 paired geometries for the same reason. Every count in this analysis is therefore
885 reported out of 33.

891
892
893 *A check with a known answer.*

894
895 Before reading any decomposition we asked what the fraction would equal if case
896 mix played no part at all. Cox recalibration, the two-parameter fit of Section 2.4,
897 applied within a level, is a strictly monotone transform of the level's scores. AUROC
898 is rank-based, so recalibration leaves the level's observed AUROC *exactly* unchanged
899 while forcing the predictions to be calibrated there, which is precisely the condition
900 under which model-based concordance reproduces observed AUROC. Computing the
901 numerator from recalibrated predictions therefore drives it onto the denominator by
902 construction, whatever the case mix. This check is reported in Section 3.2; it is what
903 identified the withdrawn 93% figure, and it is the one durable result of this analysis.

904
905 Because the extreme pair is fixed from the observed data, all intervals below are
906 conditional on that selection and do not cover the uncertainty in which levels are
907 extreme, which makes them narrower than an unconditional interval would be.

2.6 Cohorts, models and the admissibility sweep

Six clinical cohorts were used as testbeds: UCI Heart Disease (Cleveland) [67], UCI Diabetes 130-US Hospitals [68], NHIS 2024 [69], NHIS 2023 [70], NHANES 2021–2023 [71] and BRFSS 2024 [72]. A synthetic cohort from a Synthea-inspired process [73, 74] and three non-clinical contrast cohorts, UCI Adult Income [75], Folktables ACS-Income (California 2018) [76] and Statlog German Credit [77], were run identically but are reported in Additional file 1, together with the cohort-provenance table (Supplementary Table S2), because the question this paper asks is what a clinical audit faces. Every result in the main text is a clinical result unless stated otherwise.

These are testbeds, not populations. The clinical set is one encounter-level EHR extract, four unweighted complete-case survey extracts analysed *without* survey weights, strata or primary sampling units, and one 1988 dataset of 303 rows. **No population-level epidemiological inference is intended or supported by anything in this paper.** The selection is deliberate: sparse-event subgroups are the regime in which admissibility bites, and an unweighted extract is exactly what a practitioner audits when they run one of these toolkits on a public file.

Seven of the ten models are gradient-boosted tree ensembles [78] (`XGBClassifier`: `n_estimators` 200, `max_depth` 4, `learning_rate` 0.05, `subsample` 0.80, `colsample_bytree` 0.80, `random_state` 42), six of them at `eval_metric` `logloss`; the synthetic cohort’s model is built by a separate routine that does not record one. The other three, Adult Income, ACS-Income and German Credit, are logistic regression, a scikit-learn [79] pipeline of `StandardScaler` and `LogisticRegression` at library defaults with `max_iter` 1000 and `random_state` 42. Every model is fitted on an 80/20 random split within the same cohort, seed 42 throughout; no model crosses a site, time or population boundary. On Diabetes 130, where the analytic unit is an encounter, the split is grouped on the patient identifier to avoid leakage [80]. Model quality is not at issue and is not reported: all seven procedures see the same scores,

967 so any difference between them is a property of the procedure. Versions of Python,
968 NumPy, SciPy, pandas, scikit-learn, XGBoost and fairlearn are listed in Additional
969 file 1, Supplementary Table S1.
970

971 For the cohort analyses the default is $B = 10,000$ permutation draws; the simu-
972 lation uses $B = 999$ per replicate because $1,000$ replicates $\times 34$ cells $\times 7$ procedures
973 makes the larger value impractical. A permutation p -value can never be zero and none
974 is printed as such; the smallest attainable value is reported as $p \leq 1/(B+1)$. All com-
975 parator results use the *joint* permutation scheme, in which one permutation pass is
976 shared across partitions within a cohort. The independent scheme, in which each par-
977 tition is permuted separately, was also run and changes no cohort’s verdict for either
978 permutation procedure under any rule (Table 5). We make no claim that this compar-
979 ison tests a dependence assumption: both schemes combine cohorts under the same
980 independence assumption and neither is a diagnostic for the other.
981

982 Verdict stability is measured at the level at which the decision is actually made,
983 the cohort, and by *identity* rather than count: a count of flagged cohorts is invariant to
984 any permutation of the flagged set, so it can be constant while every cohort in the set
985 changes. We therefore report which cohorts change verdict across the five admissibility
986 rules, in two variants: one ignoring rules under which a cohort is not evaluable, one
987 treating non-evaluability as a distinct verdict. We do not use the count-based “does
988 the conclusion invert” metric from our earlier analysis: only a procedure with a zero
989 baseline can fail it, so it confounds power level with rule sensitivity. Because five
990 clinical cohorts are tested, cohort-level results are also reported after Holm’s step-
991 down correction [65] across cohorts. Every table is generated from stored result files
992 rather than transcribed.
993

1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012

2.7 Subgroup AUROC in closed form

The simulation of Section 2.3 computes each level’s true AUROC by quadrature. That is a numerical answer for particular geometries; the general relation it instantiates can be stated directly, and stating it converts the sweep from something that establishes a result into something that calibrates its magnitude.

Let the model be well specified in level g : the score S it reports is the true conditional event probability, so $Y \mid S \sim \text{Bernoulli}(S)$. Write $\pi_g = \mathbb{E}[S]$ for the level’s prevalence and $\Delta_g = \mathbb{E}|S - S'|$ for the Gini mean difference of its risk distribution, S and S' independent copies. Then

$$A_g = \frac{1}{2} + \frac{\Delta_g}{4\pi_g(1 - \pi_g)}. \quad (2)$$

No distributional assumption enters the identity: not normality, not a logit link, only well specification. We verify that directly, reproducing Eq. (2) against the definition of AUROC to within 7×10^{-14} on eight risk-distribution shapes including Laplace, lognormal, bimodal, uniform, t_3 and chi-square. Equation (2) is the population form of van Klaveren et al. [45]’s model-based concordance; in the rare-outcome limit it reduces to Austin and Steyerberg [64]’s binormal c-statistic expression $\Phi(\sigma_g/\sqrt{2})$, which the sweep already uses implicitly.

The assumption is in the plug-in, not in the identity, and the two must not be conflated. Evaluating Eq. (2) requires π_g and Δ_g . Recovering them from a level’s linear-predictor mean and standard deviation alone, as $L \sim N(\mu_g, \sigma_g^2)$ with $S = \text{expit}(L)$, makes both one-dimensional integrals in (μ_g, σ_g) and is what imports normality. Two parameters are not a sufficient summary of a risk distribution: at a linear-predictor standard deviation of 0.600 a Gaussian gives $A_g = 0.658$ while a two-atom distribution at the same standard deviation gives 0.646. An error of 0.012 per level is not negligible against implied gaps whose median is near 0.02, so we compute both the Gaussian

1059 plug-in and an *empirical* evaluation that takes π_g and Δ_g straight from the level's
1060 observed fitted scores and assumes no shape at all.

1062 The consequence is the whole of the paper's mechanism in one line. A_g is a func-
1063 tion of the level's own risk distribution, so a single model with one coefficient vector
1064 produces different true AUROC in every level whose predictor distribution differs,
1065 and the equal-AUROC null is false for that model by construction. *Consistency check:*
1066 evaluating Eq. (2) at the 11 named case-mix geometries reproduces the 45 per-level
1067 true AUROC values that Section 2.3's independent trapezoid quadrature returns, to
1068 a maximum absolute discrepancy of 6.13×10^{-9} , which is that rule's own truncation
1069 error.

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

2.8 Measuring case-mix magnitude in the cohorts, and computing what it implies

For every demographic partition of every cohort, at the reference rule `m30`, we take the fitted model's linear predictor, the logit of the predicted probability, which *is* the linear predictor exactly for the three logistic models and is the raw margin for the seven gradient-boosted ones, and compute its mean and standard deviation within each admissible level. The partition's measured case-mix magnitude $\hat{\rho}$ is the largest of those level standard deviations divided by the smallest. Twenty-one partitions of the five evaluable clinical cohorts have two or more admissible levels at `m30`; UCI Heart Disease contributes none, so nine of the ten cohorts appear at all and five clinical cohorts contribute the 21 partitions. Twelve further partitions of the four non-clinical cohorts are computed and stored but are not used here, because the question is what a clinical audit faces.

From measured geometry to implied gap.

Each partition’s implied case-mix gap is then computed *from that partition’s own geometry*, by evaluating Eq. (2) at each admissible level and taking the max-minus-min range. The result is the subgroup AUROC gap a perfectly fair, well-specified model would exhibit on that partition’s measured predictor distribution, in AUROC units, on the same scale as the observed gap it sits beside. **The implied gap is not a function of $\hat{\rho}$** and is not obtained by converting it. Its inputs are the per-level quantities, not their ratio, and the difference is visible in the table: three partitions at $\hat{\rho}$ of 1.140, 1.140 and 1.145 imply 0.010, 0.014 and 0.012. $\hat{\rho}$ is reported beside it as a screening summary, not as its argument.

We compute the implied gap under both evaluations of Section 2.7, the Gaussian plug-in from $(\hat{\mu}_g, \hat{\sigma}_g)$ and the empirical evaluation from the level’s observed scores, and we report both. Because the linear predictor is a property of the fitted model rather than of the cohort, we also evaluate the closed form on every one of the 24 refit specifications, and we build a permutation reference in which the levels of a partition are drawn from one common predictor distribution, so both $\hat{\rho}$ and the implied gap can be read against the noise floor of a max-minus-min statistic at that partition’s own level sizes. That reference matters because $\hat{\rho}$ is a maximum over a minimum and the implied gap a maximum minus a minimum: both exceed their no-effect value with probability one whether or not any case mix is present, so neither is evidence of anything until it is compared with its own null. The generating scripts are `recompute/casemix_implied_gap.py`, `casemix_implied_gap_robustness.py` and `casemix_implied_gap_null.py`, and their outputs are the CSVs of the same names.

A mapping we tried first, and withdrew.

An earlier version of this manuscript instead read each $\hat{\rho}$ onto the 16-point sweep by linear interpolation and reported the resulting rejection rate as an *induced false-flag*

1151 *rate*. That mapping is withdrawn and no number from it survives in this manuscript.
1152 It failed on three counts. It borrowed one geometry for every partition: the sweep fixes
1153 all level *means* at zero, while real partitions differ in level mean by up to 4.48 logit
1154 units, and our own `casemix_location_3` geometry has $\rho = 1.0$ exactly, every level at
1155 identical spread, and still induces a true gap of 0.017 and a false-flag rate of 0.226,
1156 which no ρ -indexed mapping can express. The interpolation rule was load-bearing
1157 and unreported: nearest-neighbour rather than linear interpolation moves the median
1158 from 0.094 to 0.067. And seven of the 21 partitions bracketed into the single sweep
1159 segment $[1.0, 1.1]$, whose two endpoints differ by 0.011 against a Monte Carlo standard
1160 error of 0.021 on that difference, so the interpolated value there was reading noise; an
1161 eighth, Diabetes 130 `age_group` at $\hat{\rho} = 3.304$, lay beyond the sweep's last node and
1162 was clamped rather than interpolated at all. Equation (2) needs no interpolation, no
1163 borrowed geometry and no simulation, which is why it replaces the mapping rather
1164 than repairing it.

1174
1175 ***What the implied gap is not.***
1176

1177 The empirical evaluation of Eq. (2) rests on one assumption that a real cohort does not
1178 satisfy exactly, that the model is well specified in the level; the Gaussian plug-in adds a
1179 second. The implied gap is therefore a model-based *reference value*, what a fair model
1180 with this predictor geometry would produce, and not an estimate of how much of the
1181 observed gap is case mix. It carries no interval. **We report no observed-minus-**
1182 **implied contrast, and the reason is worth stating.** Because the data-generating
1183 process has no noise term, Eq. (2) predicts which level ranks best as well as by how
1184 much, and that directional prediction is weak: the level it puts at the top and the level
1185 it puts at the bottom both match the observed extremes in only 7 of the 21 clinical
1186 partitions, against roughly 5 expected if the two orderings were unrelated. Across
1187 the 87 admissible clinical levels the implied and observed AUROC correlate at 0.852
1188 with a median absolute difference of 0.031 and a range of -0.233 to $+0.128$, which

is the size of the well-specification failure. Two unsigned ranges that often point in opposite directions cannot be subtracted and called a residual. Section 3.2 is where we attempted a decomposition and reports why it does not stand.

3 Results

3.1 How much case mix there is, and what gap it implies

Table 1 is the headline result, and the first thing to read in it is that the answer is a distribution.

The measured magnitude, reported as a range.

Across the 21 demographic partitions of the five evaluable clinical cohorts, the measured case-mix magnitude $\hat{\rho}$ has median **1.145** in the published fit. That single fit is not the quantity of interest, so we refitted every cohort under 4 model classes $\times$ 6 seeds, 24 specifications in all. The median $\hat{\rho}$ across those refits runs **1.132–1.223**, and the published 1.145 sits near the low end of that interval. **The maximum does not survive refitting and we do not report a value for it.** Diabetes 130 age_group is the argmax partition in 23 of the 24 specifications, so *which* partition carries the most case mix is stable; but its $\hat{\rho}$ runs from **1.999 to 6.388** across those specifications, with model classes disagreeing without overlapping, so *how much* is not estimated at all. The published 3.304 is one draw from that range and should never be quoted as a point. Every one of the 488 measured (partition, specification) cells exceeds 1.0; the grid has 504 cells and 16 are not measured, because `nhis2024 | insurance` loses admissibility under 12 specifications and `nhanes2123 | insured` under 4, and the smallest measured value is 1.0009.

That universal exceedance is not evidence, and we withdraw the reading we previously put on it. $\hat{\rho}$ is a maximum over levels divided by a minimum over levels, so it exceeds 1.0 with probability one whether or not any case mix is present.

1243 Against a permutation reference in which every level is a draw from one common
 1244 predictor distribution, the noise floor at these level sizes has median 1.065 across the
 1245 21 partitions and runs from 1.010 to 1.215. **Nine of the 21 partitions exceed their**
 1246 **own null’s 95th percentile, and three sit below their own null’s median:**
 1247 BRFSS 2024 income, BRFSS 2024 health plan and NHIS 2024 race have *less* spread
 1248 heterogeneity than pure sampling noise would produce at their level counts. Case mix
 1249 is present and often substantial, but it is not universal, and no partition’s $\hat{\rho}$ should be
 1250 read without its null beside it.
 1251
 1252 *What that magnitude implies, computed at each partition’s own*
 1253 *geometry, and reported as a range.*
 1254
 1255 Evaluating Eq. (2) at each partition’s own admissible levels gives the subgroup
 1256 AUROC gap a *perfectly fair, well-specified* model would exhibit there. In the pub-
 1257 lished fit its median is 0.022 and its range 0.003 to 0.173. That fit is not the quantity
 1258 of interest either, and the implied gap inherits the exposure that $\hat{\rho}$ has, because it is
 1259 computed from the same fitted linear predictor. **Across the same 24 specifications**
 1260 **the median implied gap runs 0.013–0.030 and the maximum runs 0.131–**
 1261 **0.219, so we quote a point value for neither.** The published 0.022 and 0.173 are
 1262 single draws from those ranges. The maximum is markedly better behaved than $\hat{\rho}$ ’s:
 1263 a factor of 1.7 across specifications against a factor of 3.2, because the closed form
 1264 saturates where the spread ratio does not.
 1265
 1266 **The count reaching the 0.05 convention does not survive at all.** In the
 1267 published fit five of the 21 partitions reach or exceed 0.05, the conventional threshold
 1268 at which a disparity is declared, on case mix alone with no unfairness present anywhere
 1269 in the model. Across the 24 refits that count runs from **2 to 7** of the 20 or 21 partitions
 1270 a specification measures, and the four model classes do not agree: the penalised logistic
 1271 regression gives 2 to 5, the random forest 5 to 7. Under the **ev10** admissibility rule
 1272 rather than **m30** the published-fit count is 4 of 19 and the median implied gap is 0.015.
 1273
 1274
 1275
 1276
 1277
 1278
 1279
 1280
 1281
 1282
 1283
 1284
 1285
 1286
 1287
 1288

We therefore report the count as a range, 2 to 7 of 19 to 21, and no longer state that five partitions reach the convention.

Two further readings are needed before any of these numbers is used. The empirical evaluation of Eq. (2), which assumes no shape, agrees with the Gaussian plug-in on the distribution but not on the partition: the two give medians of 0.022 and 0.022 and maxima of 0.173 and 0.202, but they differ by a median of 0.011 and by up to 0.037 partition by partition, and NHIS 2023 insured moves from 0.007 to 0.043 between them. And against the permutation reference the implied gap has a noise floor of median 0.010 across the 21 partitions, so roughly half of the median implied gap is what a max-minus-min statistic returns with no case mix at all. **Nine of the 21 partitions exceed their own null’s 95th percentile.** All five of the partitions that reach 0.05 in the published fit are among those nine, so the tail of the distribution is real even though its median is largely noise; for BRFSS 2024 race, the weakest of the five, the null median is 0.032 against an observed 0.060.

Three features of the table are worth reading together with $\hat{\rho}$ rather than through it. The implied gap is not a function of $\hat{\rho}$ and is not monotone in it: BRFSS 2024 sex, at $\hat{\rho} = 1.113$, implies 0.024, while NHANES 2021–2023 race, at $\hat{\rho} = 1.316$, implies 0.022. It is also confounded with level count, in the same way and for the same reason that $\hat{\rho}$ is: across the 21 partitions the correlation between level count and the implied gap is 0.661, the nine two-level partitions have a median implied gap of 0.010 against 0.031 for the twelve with more, and the partition carrying the maximum, Diabetes age band, is the one cut into nine native decade bins where the other cohorts use four to six. The worked example of Section 3.9 is built on that partition, and the confound belongs beside it. Level count and level means both matter, and NHIS 2024 age band shows why: its levels differ in linear-predictor mean by 4.48 logit units, the largest spread anywhere in the study, which the spread ratio alone does not see. This is the

1335 concrete reason the withdrawn mapping of Section 2.8 could not work: $\hat{\rho}$ is a useful
 1336 screening summary, not a sufficient statistic for a partition's case-mix geometry.
 1337
 1338
 1339 *Age partitions carry more case mix than sex partitions, and the ordering*
 1340 *survives refitting.*
 1341
 1342 The five `age_group` partitions occupy the top of the $\hat{\rho}$ distribution and the five `sex` or
 1343 `gender` partitions the bottom: in the published fit age runs from 1.285 to 3.304 and
 1344 sex from 1.022 to 1.145, with no overlap, though the upper end of the age range is the
 1345 value we withdraw above and should be read as a range. Race and ethnicity partitions
 1346 fall in between, from NHIS 2024 at 1.051 to BRFSS 2024 at 1.335. The ordering is
 1347 not an artefact of one fit. Across the 24 refit specifications, age exceeds sex in **116**
 1348 of **120** (specification, cohort) comparisons, and the concordance of the full partition
 1349 ranking across specifications has Kendall's $W = 0.692$ over 19 partitions against a
 1350 random-ranking mean of 0.041.
 1351
 1352
 1353 **Three facts scope that ordering and all three belong beside it.**
 1354
 1355 *First, it is confounded with level count, and refitting cannot test the confound.*
 1356 Every sex partition here has $k = 2$ levels; every age partition has $k = 4$ to 9. Across
 1357 the 21 clinical partitions, Spearman's correlation between level count and $\hat{\rho}$ is **0.582**
 1358 ($p = 0.006$). More levels give more opportunity for the extreme pair of standard
 1359 deviations to separate, so a max-over-levels ratio is expected to rise with k under
 1360 any model. Refitting leaves this untouched, because level count is a property of the
 1361 partition and not of the model: all 24 specifications see the same k . We cannot separate
 1362 "age is a stronger driver of predictor spread" from "age partitions are cut more finely,"
 1363 and we do not claim to.
 1364
 1365 *Second, it fails under a change of admissibility rule.* Under `ev10` rather than `m30`,
 1366 NHIS 2024 `age_group` falls from rank 19 of 21 to rank **7 of 19**, its $\hat{\rho}$ dropping from
 1367 1.362 to 1.073. The rule change drops that partition's sparsest levels, which are the
 1368 ones carrying its extreme spreads. A pattern that moves this much on a researcher
 1369
 1370
 1371
 1372
 1373
 1374
 1375
 1376
 1377
 1378
 1379
 1380

degree of freedom that Section 3.7 shows is itself unjustified in practice is not a stable finding about attributes.

Third, the ordering is coarse and nothing finer than it survives. Pairwise Spearman correlation of the partition ranking between specifications has median **0.701**, with **49%** of pairs below 0.7 and the fifth percentile at 0.417. **We therefore make no claim about the ordering of individual partitions**, only about the age-versus-sex contrast, which is the one comparison coarse enough to survive.

A plausible explanation is that risk models draw heavily on age-associated physiology, so predicted risk spreads out very differently across age bands and very similarly across sex categories; this is consistent with the fall in CHARGE-AF discrimination from 0.72 to 0.57 across age strata within one model [47], but we did not test it, and given the level-count confound we report the pattern as an empirical regularity in five fitted models rather than as a mechanism. The practical consequence is correspondingly modest. A subgroup AUROC gap found on an age partition in these cohorts sits where the case-mix-implied gap is largest, four of the five age partitions implying 0.05 or more on their own under `m30` and three of five under `ev10`, so it should not be treated as evidence of anything until the case-mix distribution has been inspected. This is a statement about which *audit results* deserve weight. It is not a statement that any cohort’s model is fair, nor that any cohort is free of subgroup differences.

3.2 A negative result: the observed gaps cannot be decomposed

The previous section concerns what a *fair* model would show. The different question, of a *real* observed gap, what share is genuine and what share is case mix?, is the one an auditor actually wants answered, and this section reports that we could not answer it.

1427 *A methodological warning that does stand: recalibration makes the*
 1428
 1429 *decomposition tautological.*
 1430
 1431 Our first attempt reported that case mix explains a median 93% of observed gaps.
 1432 It was computed from a fraction whose numerator had first been Cox-recalibrated
 1433 within each level. Cox recalibration is a strictly monotone transform, so it leaves the
 1435 level's observed AUROC *exactly* unchanged, verified to within 10^{-12} , while forcing
 1436 the predictions to be calibrated there, and model-based concordance of calibrated
 1437 predictions is then obliged to reproduce that same observed AUROC. It does: across
 1440 all **224** admissible level rows the median absolute difference from observed AUROC
 1441 is **0.0030** for the recalibrated quantity, with correlation $r = \mathbf{0.9956}$, against 0.0162
 1442 and $r = 0.935$ for the raw one. The numerator was being driven onto the denominator
 1445 by construction, so the fraction measured the recalibration step and not case mix.
 1447 This is a general warning, not a bug report on our code, and it is the durable result
 1448 of this analysis. Any attributable-fraction estimator that recalibrates its model-implied
 1450 benchmark within the subgroup it is benchmarking will reproduce that subgroup's
 1451 observed discrimination by construction and report a fraction near 1, whatever the
 1452 case mix. The step looks like a correction and is a tautology. We record it because
 1455 the arrangement is natural enough that we chose it ourselves, and nothing about the
 1456 code's correctness, its tests or its reproducibility would have revealed the defect; what
 1457 revealed it was asking what the quantity would equal if case mix played no part at all.
 1460
 1461 *The corrected fraction is also withdrawn: the decomposition is not*
 1462
 1463 *reliably estimable at these sample sizes.*
 1464
 1465 Repairing the estimator gave a median of about 29%, and we do not report that
 1466 number either. The fraction's denominator is an observed AUROC gap, and at these
 1468 sample sizes that gap is frequently indistinguishable from zero: its median across the
 1469 33 distinct partitions is 0.046 and its tenth percentile is 0.010, so the ratio is routinely
 1470 a small number divided by a smaller and noisier one. The consequences are not subtle.
 1471
 1472

The fraction is reportable at all, operationalised as a bootstrap 2.5th percentile of the observed gap exceeding zero, in only **15 of the 33 distinct partitions**, and among those the median 95% interval is 0.72 wide across both admissibility rules and 0.98 wide at m30 alone; 11 of 27 rule-cells exceed 1.0 in width and 4 exceed 2.0. Where the gate does not bind the estimator degenerates outright: ACS-Income **sex** has an observed gap of **0.0040** and an attributable fraction of -8.61 . A quantity that can take that value on real data at this sample size is not estimating a share of anything. We therefore report no attributable fraction and no claim about what share of any observed gap is case mix.

The residual gap is a difference, not a ratio, and we restore it.

An earlier version withdrew the residual gap, observed minus model-implied, along with the fraction. That went too far. The pathology above is a near-zero noisy denominator, and a difference on the AUROC scale has no denominator to be pathological. Its intervals are correspondingly usable: median width **0.109** at m30 against 0.98 for the fraction on the same partitions, and **10 of the 33** partitions have a residual gap whose 95% lower bound exceeds zero. Two caveats have to travel with that tally and they are not small. The intervals are conditional on the extreme level pair selected from the observed data, so the count is anti-conservative. And **8 of the 33 partitions have a negative model-implied gap**, ACS-Income **sex** at -0.0345 , NHIS 2024 **age_group** at -0.0200 and Diabetes 130 **race** at -0.0025 among them, which means the benchmark is below the observed value for a reason the estimator cannot explain; 4 of the 10 significant residuals sit on one of those. Restricting to clinical partitions with a non-negative implied gap and a positive lower bound leaves **three of the 21**: BRFSS 2024 **race**, NHIS 2024 **income** and the NHANES age band. Three of 21 is a modest result and we report it as one. It is not a decomposition, it is not a share, and it does not identify unfairness; it says only that in three partitions the observed gap

1519 exceeds what the model-based benchmark accounts for by more than the conditional
1520 interval covers.

1522

1523 *This is the thesis, not an exception to it.*

1524

1525 The paper’s claim is that a subgroup AUROC gap is not interpretable as fairness evi-
1526 dence because it mixes a genuine component with a case-mix component in unknown
1527 proportion. Our own failure to recover those proportions, with the model in hand,
1528 the data in hand, the estimator the literature supplies and 2,000 bootstrap replicates
1529 per partition, is the strongest available evidence for exactly that claim. An auditor
1530 with less than we had will not do better. The right response to an undecomposable
1531 gap is not a better decomposition but a report that does not depend on one, which is
1532 what Section 3.1 and the Discussion recommend: the subgroup AUROCs with their
1533 intervals, the case-mix distribution beside them, and no hypothesis test of equality.
1534 Nothing here says any cohort’s model is fair, and nothing here says any cohort’s model
1535 is unfair; that is the point.

1543

1544

1545 3.3 The flag rate is a steep function of case-mix magnitude

1546

1547 Figure 1, Table 2 and Figure 2 give the curve that Section 3.1 reads against. The
1548 model in every one of these simulations is the data-generating probability itself: one
1549 coefficient vector, Bayes-optimal, perfectly calibrated in every subgroup, no unfairness
1550 of any kind present. Subgroups differ only in the distribution of the predictor they are
1551 drawn from, so every rejection is a false alarm about fairness.

1552 The raw max–min range permutation test’s false-flag rate rises from 0.056 at $\rho =$
1553 1.0, where it is correctly sized, through 0.141 at $\rho = 1.222$, 0.481 at $\rho = 1.5$ and 0.918
1554 at $\rho = 2.0$, to 1.000 at $\rho = 3.0$ and above. The other procedures track it closely: at
1555 $\rho = 2.0$ the studentized test gives 0.915, the double-corrected z -test 0.938 and the fixed
1556 cut-point 0.988. The curve is steepest between $\rho = 1.2$ and $\rho = 2.0$, which is precisely

1563

1564

the interval in which real partitions sit, so small differences in case-mix magnitude between two audits translate into large differences in how often each is misled.

The 0.918 rate at $\rho = 2.0$ is one point on this curve, and it should be read as such. The median measured partition sits at $\hat{\rho} = 1.145$, well below it. An earlier draft of this work took the $\rho = 2.0$ figure as its headline and described that geometry as ordinary. It is not ordinary, and the correction is recorded in the Disclosure. What survives is stronger than a point: the flag rate is a known, steep, monotone function of a quantity that can be measured in any cohort before an audit is run. We do not, however, convert a partition's $\hat{\rho}$ into a position on this curve (Section 2.8); the curve establishes that the failure is steep and real, and Table 1 reports what each partition's own geometry implies.

The failure is not a small-sample artefact and does not go away with more data; it gets worse. At the named moderate geometry ($\rho = 2.0$, true gap 0.123) the raw max-min range permutation test flagged in 0.920 of simulations, the studentized test in 0.919, the double-corrected z -test in 0.946, Cochran's Q in 0.926 and the fixed 0.05 cut-point in 0.991. Holding that true gap fixed and raising n from 2,000 to 10,000, all five of those procedures reached **1.000**, a perfectly fair model flagged in every one of a thousand simulations, and the bootstrap-interval reading moved from 0.013 to 0.972. This is what consistency looks like when the estimand is wrong: the procedures are consistent estimators of a subgroup AUROC difference, that difference is real, and it is not evidence of unfairness. It also means that a partition's case-mix-implied gap is a lower bound on the trouble it causes: the implied gap does not depend on sample size, but the probability that a procedure fires on it rises with sample size without limit.

Two procedures stayed low and neither is a counterexample. The four-fifths rule flagged at 0.119 at the moderate geometry and *fell* to 0.006 at $n = 10,000$, a diagnosis rather than a defence: a ratio of two AUROC values sits near 1.0 for any pair of usable AUROC values, so the rule is too insensitive to detect anything. It reached 0.791 at

the strong geometry, so it is not immune, merely slow; the bootstrap interval behaved similarly (0.013, 0.757, 0.972). Adding partitions does not help either: the moderate geometry as the first of three partitions, the other two pure noise, still gave 0.852, 0.864, 0.918 and 0.993 for the four leading procedures. And the failure begins below the conventional threshold of concern. At the mild geometry, whose true gap of 0.036 is *below* the 0.05 convention used to declare a disparity, the studentized test flagged at 0.164, the z -test at 0.208 and the fixed cut-point at 0.506. That geometry is $\rho = 1.222$, a case-mix magnitude that 9 of the 21 measured partitions reach or exceed, so this is the part of the curve most audits are actually standing on.

3.4 Under a well-specified data-generating process, the calibration test does not inherit the failure

The alternatives this literature recommends had, until now, never been measured against the failure this paper documents. Table 3 and Figure 1 measure them, on the same 16-point sweep, the same replicates and the same admissible levels as the seven procedures. The scope of that measurement is narrow and we state it before the numbers.

Scope.

Every cell in which the calibration test was run comes from the case-mix data-generating process of Section 2.3, in which the model score *is* the true conditional event probability. That is the condition under which the null it tests, “this model is calibrated within each subgroup”, is true by construction. **We have not evaluated the calibration test where scores are uncalibrated**, which includes the 17 equal-AUROC null geometries of Table 4: those harnesses build composite nulls by applying level-specific monotone maps to the scores, which preserves each level’s AUROC exactly and destroys calibration by design, so they cannot serve as a size

check for a calibration test and we did not run one on them. Everything below there-
fore reads: *under a data-generating process in which the score is the true conditional
probability, the calibration test holds nominal size across case-mix magnitudes.* It does
not read that the test is well behaved in general.

Within that scope, the size result is clean.

Across the full sweep, while the raw max–min permutation test climbs from 0.056 to
1.000 and the fixed cut-point from 0.295 to 1.000, the calibration test’s false-flag rate
stays at nominal at every one of the 16 points: 0.056, 0.047, 0.046, 0.053, 0.051, 0.032,
0.041, 0.061, 0.051, 0.066, 0.057, 0.035, 0.048, 0.057, 0.045, 0.052. It has no trend in
 ρ ; the whole range, 0.032 to 0.066, sits within Monte Carlo error of 0.05 at 1,000
simulations (standard error about 0.0069), and the mean over the 16 points is 0.0499.
The prevalence-equalised sensitivity sweep gives the same answer (0.039 to 0.060 across
the same 16 points). The structural reason is the contrast with every AUROC-gap
procedure in Table 2: the null “this model is calibrated within each subgroup” remains
true for a well-specified model when subgroups differ in case mix, whereas the null
“true AUROC is equal across subgroups” does not.

The direction of that immunity is not universal.

van Amsterdam [50] show that which of discrimination and calibration is stable under
case-mix shift depends on the causal direction of the prediction task: when prediction
runs in the causal direction, typical of prognosis, calibration is the stable one and
discrimination is not: the arrangement our simulations instantiate. When prediction
runs in the *anti-causal* direction, typical of diagnosis, the relationship reverses and it
is calibration that moves. Diagnostic models are a large share of the clinical prediction
models these audits are run on. Our result is therefore specific to the causal direction
we simulated, and should not be carried across to a diagnostic setting without checking
that first.

1703 *Model-based-concordance excess is flat too, but it is not a test.*
1704
1705 The maximum over levels of model-based concordance minus observed AUROC stays
1706 between 0.018 and 0.021 AUROC units at every point of the sweep, with no trend in
1707 ρ , against a true AUROC gap that grows from 0 to 0.198 over the same range. The
1708 quantity does exactly what van Klaveren et al. [45] intends: it does not manufacture a
1709 signal out of case mix. **We report it as a statistic and not as a rejection rate,**
1710 because in this harness it has no threshold, no reference distribution and no error rate;
1711 the flat value near 0.019 is a small finite-sample positive bias, not a false-flag rate
1712 below nominal.
1713
1714
1715 *A flat rate is worthless without power, so we checked, and the power is*
1716 *modest.*
1717
1718 A procedure that never fires also has a nominal false-flag rate. The matched-pair
1719 positive control of Section 2.4 holds the true AUROC gap fixed and varies only the
1720 mechanism. **All three matched pairs are reported here, including the one**
1721 **where the test does poorly.** At a true gap of 0.050 the calibration test flags the
1722 case-mix arm at 0.049 and the genuinely unfair arm at **0.078**, which is barely above
1723 its own size and no use as a detector. At 0.123 the two are 0.051 and 0.293. At 0.200
1724 they are 0.047 and 0.502. On the mixed geometry, where a total gap of 0.123 is 0.076
1725 genuine excess, it flags at **0.092**. So the test reaches 50% power only at an excess of
1726 0.200, and below about 0.123 it is close to blind. It responds to unfairness rather than
1727 to the gap, which is the property we are reporting, but it responds weakly.
1728
1729 **Coverage of that comparison.** Five of the seven AUROC-gap procedures have
1730 positive-control rows: the fixed cut-point, the four-fifths rule, the double-corrected
1731 z -test, and the raw and studentized permutation tests. Cochran's Q and the bootstrap-
1732 interval reading produced no evaluable replicates on these geometries and are absent
1733 from Panel B of Table 3. The statement below is therefore about five procedures, not
1734 seven, and we have no positive-control evidence either way for the other two. Across

those five, the case-mix and unfairness arms are indistinguishable: 0.948 against 0.950
for the raw permutation test, 0.946 against 0.947 for the studentized test, 0.970 against
0.967 for the double-corrected z -test, and 1.000 against 1.000 at a gap of 0.200 for
four of the five, the exception being the four-fifths rule, which is 0.974 against 0.970
there. They have no discriminant validity in either direction.

The converse case is starker. On the two geometries where the model *is* unfair,
a subgroup-specific calibration slope of 2.0 and a subgroup-specific intercept shift of
one logit, but the true AUROC gap is exactly zero, the **three** inferential AUROC-
gap procedures with positive-control rows range from just below nominal to well
above it (0.043 to 0.078), the four-fifths rule at 0.000 and the fixed cut-point at
0.266 to 0.313, while the calibration test flags at **1.000**. **Those two geometries**
are correctly-specified alternatives for the calibration test, not evidence
of general sensitivity: they perturb the intercept and the slope, which are the two
parameters the test estimates, so a rate of 1.000 there is the test detecting exactly
what it was built to detect and says nothing about its sensitivity to other forms of
unfairness. What the pair of results does establish is the complementary point about
the gap tests, which is that they miss real unfairness when it does not happen to move
AUROC.

Model-based-concordance excess is the weaker of the two on this control. It sepa-
rates the matched arms only slightly better than chance: an area under the curve for
distinguishing a case-mix replicate from an unfairness replicate of 0.590 [0.565, 0.614]
at a gap of 0.123 and 0.545 [0.520, 0.570] at 0.200, against 0.743 [0.722, 0.765] and
0.840 [0.823, 0.857] for the calibration test and 0.491 to 0.514 for every AUROC-gap
procedure. The honest summary is that within the scope stated above the calibration
test does not respond to case mix and does respond, weakly, to unfairness, while model-
based-concordance excess does not respond to case mix and barely detects unfairness
at all. We report both rather than only the one that performs better.

1795 ***What this does and does not establish.***

1796

1797 These are two existing recommendations [45, 58, 59] measured against a failure mode

1798

1799 they were not previously tested on. We are not proposing a procedure and we are not

1800

recommending the Cox calibration test as a fairness test. Four things bound the claim.

1801

1802 It is a weak-calibration test, the lowest rung of Van Calster et al. [66]’s hierarchy.

1803

1804 Its immunity is specific to the causal direction we simulated and reverses in the anti-

1805

causal direction typical of diagnosis [50]. Calibration is in any case a weak fairness

1806

1807 criterion, satisfiable by discarding information from an existing classifier [57], and it

1808

1809 cannot be combined with equal error rates at unequal base rates [55, 56]. And it

1810

1811 assumes within-subgroup calibration is estimable, which in a level with a few dozen

1812

1813 events it is not. What survives is one comparative sentence: under the conditions that

1814

break the AUROC gap, this quantity does not break.

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

1836

1837

1838

1839

1840

3.5 Studentization repairs location, not spread

The one case-mix geometry in which the procedures separate is the one that is not

really about discrimination. When the three levels differ in the predictor’s *location*

only, at equal spread, the true AUROC gap is just 0.017, the subgroups have differ-

ent risk, not different separability, and the raw range test flagged at 0.226 while the

studentized test flagged at 0.044. Studentization does exactly what the theory says

[8, 17, 22]: it restores approximate validity when score distributions differ but true

AUROC’s do not. That repair does not extend to the case that matters. At the mod-

erate spread geometry ($\rho = 2.0$) the raw range flagged at 0.920 and the studentized

test at 0.919, a difference of one simulation in a thousand and well inside Monte Carlo

error; on the sweep at the same ρ the two give 0.918 and 0.915. The sweep shows the

same thing at every magnitude in between: the studentized and raw curves are within

0.008 of each other at all 16 points. When case mix genuinely changes the true sub-

group AUROC, a better-calibrated test of the equal-AUROC null simply detects the

true difference more reliably; a test cannot be rescued by improving its calibration against a null that a fair model does not satisfy.

3.6 Under equal-AUROC nulls the procedures differ sharply, and it does not help

Table 4 reports Type I error over the 34 cells in which every subgroup has the same true AUROC. Here the procedures separate cleanly. The fixed 0.05 cut-point rejected at mean 0.7997 and 0.6800 under the simple and composite nulls, worst cell **0.999**: on some geometries a screen that fires essentially every time nothing is wrong. The fourth-fifths rule rejected at 0.1056 and 0.0384, worst cell **0.466**. The double-corrected z -test gave 0.0948 and 0.0927, worst cell 0.168; Cochran’s Q 0.0808 and 0.0580, worst cell 0.207; and the bootstrap interval failed in the opposite direction at 0.0038 and 0.0006, so three inferential readings of one estimator span rejection rates from 0.0006 to 0.0948 on the same data. The two permutation procedures held nominal size: the studentized test 0.0486 and 0.0480, worst cell 0.069; the raw range 0.0490 and 0.0293, worst cell 0.058. Both worst cells lie inside the reference band for a maximum over 34 cells (0.0700 under the paired `m30/ev10` structure; 0.072 if the 34 cells are treated as independent), so neither is evidence of miscalibration; the studentized test’s worst cell sits at 0.069 against the relevant 0.0700 band, not against the wider, independence-based 0.072, so the margin is thin rather than comfortable. The raw range’s composite-null mean of 0.0293 is the expected conservatism of an unstudentized statistic where exchangeability fails.

The ranking in Table 4 is bounded in scope and we do not ask it to bear weight, for the reason given in Section 2.2. Munko and Sattler [23] add a second reason specific to this table: standard studentization does not give the correct joint limiting distribution in multiple-testing settings, which undercuts our max- T construction directly, and we have not implemented their covariance correction. A

1887 reader should take from this table that the fixed cut-point and the four-fifths rule are
1888 badly calibrated against the null they nominally test, a coarse conclusion consistent
1889 with a literature going back to 1979 [10, 11, 15]. A reader should not take from it that
1890 any one of the four inferential procedures is preferable to another.
1891

1892 The case-mix result does not depend on the ranking, which is why it survives the
1893 caveat. Section 3.3 shows that every AUROC-gap procedure here fails under case mix,
1894 in the same direction and at similar magnitudes: the two procedures nominal in this
1895 table flag a perfectly fair model at 0.918 and 0.915 once case mix reaches $\rho = 2.0$, and
1896 at 0.141 and 0.135 by $\rho = 1.222$, a magnitude 9 of our 21 measured partitions reach.
1897 No reordering of the seven, and no correction to any one implementation, changes
1898 that; the failure is in the estimand they share. The ordering of procedures by Type I
1899 error under an equal-AUROC null therefore carries no information about the ordering
1900 that matters, and we make no claim about it.
1901

1902 3.7 The admissibility rule is a researcher degree of freedom 1903 1904 1905 1906

1907 This section and the next concern how an audit is assembled rather than what it
1908 concludes, and they carry the least weight of anything reported here: they are demon-
1909 strations on testbeds, with the caveats of Section 2.6 attached to every number. Table 5
1910 gives the per-cohort verdict matrix on the five estimable clinical cohorts across the
1911 five admissibility rules. UCI Heart Disease is never evaluable: no partition retains
1912 two admissible levels, so Eq. (1) has no value. The four non-clinical contrast cohorts,
1913 and the cohort-provenance table, are in Additional file 1; the question here is what a
1914 clinical audit faces.

1915 Read by identity rather than count, no procedure that flags anything is stable.
1916 On the five clinical cohorts the number whose verdict is not constant across the five
1917 rules is 3 of 5 for the studentized test and the double-corrected z -test, 2 of 5 for the
1918 raw range and for Cochran's Q , 1 of 5 for the fixed cut-point and the four-fifths rule,
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932

and 0 of 5 for the bootstrap interval, which flags nothing anywhere. Under the strict variant, which counts a transition into or out of non-evaluability as a change, the clinical counts are unchanged.

Two things follow. First, the procedure whose flagged *count* is most stable is among the *least* stable by identity. The studentized test’s clinical flagged set is {BRFSS 2024, Diabetes 130, NHIS 2024} under **m20**, **m30** and **m50**, {Diabetes 130, NHIS 2023, NHIS 2024} under **m100**, and {BRFSS 2024, Diabetes 130, NHIS 2023} under **ev10**: the same count of three throughout, and three different sets. A count-based metric records that as perfect stability, which is why we do not use one. Second, the only procedure with a constant verdict set is the one that never flags, which is trivial. A rule chosen for statistical convenience, and usually neither justified nor reported, therefore decides which cohorts an audit names.

The raw range test flags no clinical cohort under **m20**, **m30** or **m50**, NHIS 2023 under **m100**, and NHIS 2023 and NHIS 2024 under **ev10**; after Holm correction across the five estimable clinical cohorts it flags none under any rule. That is the one procedure Holm nullifies completely; it is not representative of the other four procedures with a *p*-value. The studentized test survives Holm at Diabetes 130 and BRFSS 2024 under **m20**, **m30**, **m50** and **ev10**, and at Diabetes 130 alone under **m100**. The double-corrected *z*-test and Cochran’s *Q* survive more broadly still: the *z*-test at three clinical cohorts under **m20**–**m50** and at one to three under **m100/ev10**; Cochran’s *Q* at all five clinical cohorts under **m20**–**m50** and at three to four under **m100/ev10**. Table 5 gives the Holm-adjusted decision for every procedure and rule beside the unadjusted one. The independent permutation scheme changes no verdict for either permutation procedure.

3.8 Admissibility decides whose performance is measured

Ten subgroup levels were excluded across the ten cohorts at the reference rule (Figure 3): seven below the minimum size of 30 rows and three for degenerate labels.

Two of those three have no cases at all, which makes AUROC undefined rather than merely imprecise. The third, the synthetic cohort's 18–44 band, has one case in 371 rows: its AUROC is computable and carries no information, and it fails the two-of-each floor of Section 2.1. Six are age bands, one is a sex category, and three are race and ethnicity categories, all three of them minority categories in NHIS 2024 and two of them carrying no cases at all. The pattern is structural, not incidental: an admissibility rule removes the smallest levels first, and in cohorts drawn from a majority population the smallest levels are the minority categories. A rule chosen for statistical convenience therefore decides which populations an audit is capable of saying anything about at all, and it is currently chosen without justification and often without report.

3.9 A worked example

Take one cohort and audit two attributes. In Diabetes 130 the **gender** partition shows an observed subgroup AUROC gap of 0.013, and its measured geometry implies a gap of 0.003 for a fair model, against a noise-only reference of 0.001. The **age_group** partition of the same cohort, same model, same test set, shows an observed gap of 0.216, and its geometry implies between 0.117 and 0.219 across the 24 refit specifications, 0.173 in the published one, against a noise-only reference of 0.017. An audit that reports both gaps against the same 0.05 cut-point, or against the same permutation null, treats them as two readings of one quantity. They are not. On the gender partition a fair model would show essentially no gap, so the observed 0.013 is the model's own. On the age partition a fair model would show something between two and four times the threshold at which a disparity is declared, on case mix alone. The age result carries far less information about fairness than the gender result, and nothing in the audit's own output says so, because the audit does not measure case mix.

We are careful about what this does and does not license. It does not license the conclusion that the 0.216 age gap in Diabetes 130 is case mix. The implied gap is what

a fair, well-specified model would show with this predictor geometry, not an estimate of this model’s case-mix component, and Section 3.2 reports that we could not estimate that component at all. **Nor is a single number available here:** no implied-gap value in this example is stable enough across specifications to quote as a point, which is why the age figure is given as a range. This partition is also the one cut into nine native decade bins where the other cohorts use four to six, and both $\hat{\rho}$ and the implied gap rise with level count, so part of what makes it extreme is how finely it is cut. Nor does it license the conclusion that the model is fair on age, or unfair on gender, or anything about this model, which is fitted to an encounter-level extract as a testbed. What it licenses is a claim about evidential weight. Two gaps of similar nominal size, read through the same test, can carry very different amounts of information, and the difference is computable in advance from the predictor distribution alone.

For the same point in a controlled setting, take a team auditing three age bands of roughly 667 patients each at prevalence 0.20, whose oldest band has predictor values twice as dispersed as the youngest’s. They are standing at $\rho = 2.0$, where the fixed cut-point fires with probability 0.991, the double-corrected z -test with 0.946, a raw permutation test with 0.920, and studentizing it changes that to 0.919. Only the four-fifths rule usually stays quiet, at 0.119, and it stays quiet at a true gap of 0.199 in one simulation in five. The model there is the exact data-generating probability and treats no patient differently on the basis of age.

4 Discussion

The problem is the estimand, not the test

The results above reduce to two statements, and the second is the one that makes the first usable. Subgroup fairness auditing of clinical prediction models tests H_0 : equal subgroup AUROC. That null is false for a perfectly fair model whenever case mix differs across subgroups, and case mix exceeded a noise-only reference in 9 of the 21

Table 1 Measured case-mix geometry in every evaluable clinical partition, the subgroup AUROC gap a perfectly fair model would show there, and what survives refitting and a noise-only reference. $\hat{\rho}$ is the largest over the smallest per-level standard deviation of the fitted model’s linear predictor, over levels admissible under m30. The implied case-mix gap is Eq. (2) evaluated at each level, maximum minus minimum, in AUROC units, under two evaluations: *Gauss.* recovers the level’s risk distribution from its linear-predictor mean and standard deviation, *Emp.* takes it from the observed scores. *Refit range* is the Gaussian value’s range over the 24 refits. *Noise* is the median of a permutation reference in which every level is a draw from one common predictor distribution, so it is what this statistic returns with no case mix; † marks the 9 partitions above that reference’s 95th percentile. Rows are ordered by the Gaussian value, the rule marks its median, age bands are bold. **No implied-gap value is an estimate; none should be quoted as a point.**

Cohort	Partition	Lev.	$\hat{\rho}$	Obs. gap	Implied gap		Refit range	Noise
					Gauss.	Emp.		
NHIS 2023	sex	2	1.022	0.019	0.003	0.002	0.001–0.015	0.002
Diabetes 130	gender	2	1.022	0.013	0.003	0.002	0.001–0.019	0.001
NHANES 21–23	sex	2	1.090	0.020	0.005	0.001	0.001–0.023	0.003
NHIS 2023	insured†	2	1.037	0.018	0.007	0.043	0.000–0.049	0.005
NHIS 2024	race	4	1.051	0.109	0.008	0.027	0.004–0.030	0.020
NHANES 21–23	insured	2	1.140	0.033	0.010	0.005	0.000–0.037	0.006
BRFSS 2024	health plan	2	1.029	0.011	0.010	0.017	0.002–0.037	0.007
NHIS 2024	sex	2	1.145	0.004	0.012	0.001	0.000–0.018	0.004
NHIS 2024	income	5	1.140	0.097	0.014	0.022	0.003–0.038	0.017
BRFSS 2024	income†	7	1.063	0.046	0.015	0.027	0.007–0.078	0.012
NHANES 21–23	race	5	1.316	0.046	0.022	0.020	0.004–0.058	0.018
BRFSS 2024	sex†	2	1.113	0.014	0.024	0.022	0.012–0.037	0.002
NHIS 2023	race	7	1.186	0.129	0.025	0.050	0.022–0.070	0.039
Diabetes 130	race	5	1.240	0.124	0.028	0.033	0.011–0.046	0.018
NHANES 21–23	age band†	4	1.602	0.075	0.034	0.020	0.005–0.112	0.010
NHIS 2024	insurance	2	1.322	0.044	0.035	0.014	0.001–0.061	0.008
BRFSS 2024	age band†	5	1.285	0.068	0.057	0.090	0.021–0.154	0.012
NHIS 2024	age band†	6	1.362	0.328	0.057	0.074	0.019–0.178	0.021
BRFSS 2024	race†	8	1.335	0.224	0.058	0.060	0.028–0.096	0.032
NHIS 2023	age band†	4	1.348	0.128	0.069	0.096	0.029–0.138	0.007
Diabetes 130	age band†	9	3.304	0.216	0.173	0.202	0.117–0.219	0.017

Read the ranges, not the columns. In this fit the implied gap has median 0.022, minimum 0.003, maximum 0.173, and 5 of 21 partitions reach 0.05. None of those four numbers survives refitting as a point. Across 24 specifications (4 model classes $\times$ 6 seeds) the median runs 0.013–0.030, the maximum 0.131–0.219, and the 0.05 count 2–7 of the 20 or 21 partitions a specification measures; under *ev10* the published-fit median is 0.015 and the count 4 of 19. The same applies to $\hat{\rho}$: median 1.145, minimum 1.022, maximum 3.304 here, against a refit median of 1.132–1.223 and a refit maximum of 1.999–6.388 (Section 3.1). NHIS 2024 **insurance** is measured under 12 specifications and NHANES **insured** under 20, so their refit ranges rest on fewer cells.

Every implied-gap value exceeds zero by construction, because a maximum minus a minimum does, so the *Noise* column carries the information. Its median is 0.010, about half the median implied gap; only the 9 partitions marked † clear their own reference’s 95th percentile, and three partitions have a $\hat{\rho}$ below their own null’s median.

The observed gap column is descriptive, not a decomposition: the two orderings agree at both extremes in only 7 of 21 partitions against about 5 expected by chance, and Section 3.2 reports why the decomposition does not stand. Nothing here says any model is fair or unfair and no gap is a population disparity (Section 2.6); UCI Heart Disease contributes no row. Both the implied gap and $\hat{\rho}$ rise with level count, a property of the partition and not the model (correlations 0.661, 0.582), and the maximum row is the only partition cut into nine decade bins.

Eq. (2) itself makes no distributional assumption; the *Gauss.* plug-in does, and the two differ by a median of 0.011 and by up to 0.037. Both assume well-specification, which fails here: implied and observed level AUROC correlate at 0.852 across 87 levels, median absolute difference 0.031. Generated by `casemix_implied_gap{, robustness, null}.py`; $\hat{\rho}$ refits from `sd_ratio_robustness.csv`, decomposition from `model_based_concordance.csv`.

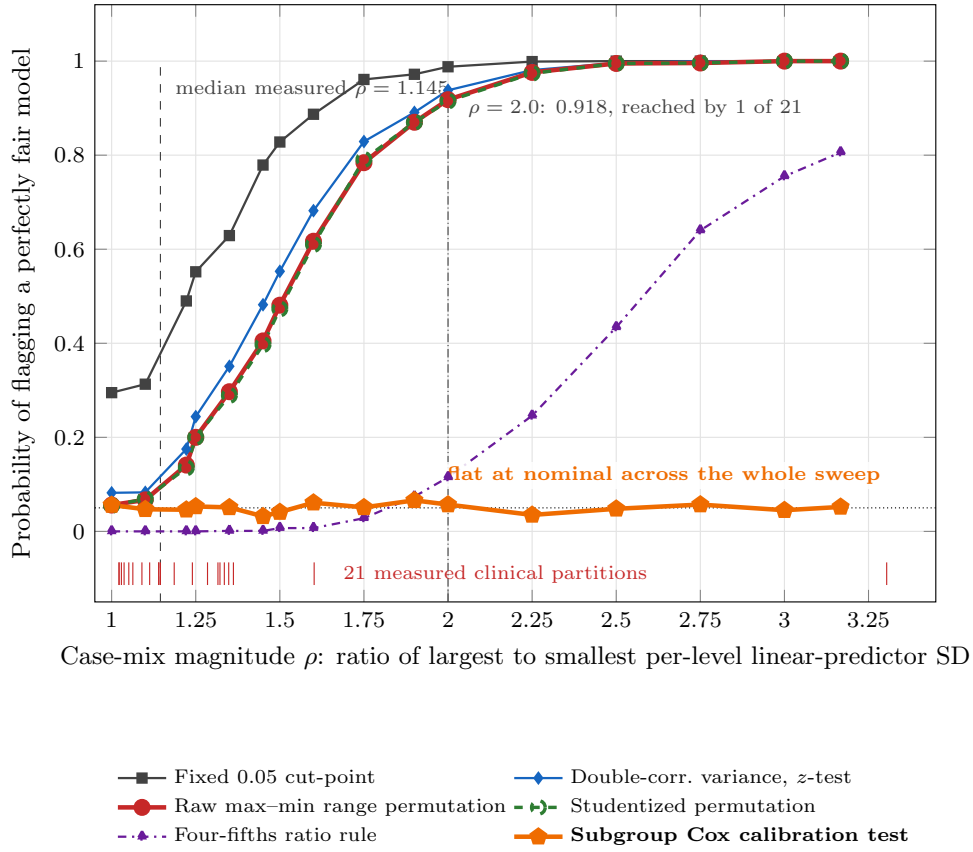


Fig. 1 The false-flag rate is a steep function of case-mix magnitude, and real partitions are spread across the whole of it. Curves: rejection probability for a model that is the data-generating probability, Bayes-optimal, one coefficient vector, perfectly calibrated in every subgroup, as the per-level linear-predictor SD ratio ρ moves from 1.0 to 3.167. Three equal levels, $n = 2,000$, prevalence 0.20, m_{30} , 1,000 simulations and $B = 999$ per point; the induced true AUROC gap runs from 0 to 0.198 and is computed by quadrature. Every rejection is a false alarm about fairness. The dotted horizontal line is the nominal 0.05. Rug marks along the bottom are the 21 measured clinical partitions of Table 1, one tick each, shown to indicate where real partitions sit on this axis. **We do not read a flag rate off this curve for any partition:** the sweep fixes all level means at zero and uses three equal levels at $n = 2,000$, which no real partition matches, and Table 1 instead computes each partition's implied gap from its own geometry (Section 2.8). The dashed vertical is the median measured magnitude, 1.145. The dash-dotted vertical is $\rho = 2.0$, where the raw max-min permutation test flags at 0.918, one point on this curve and not the typical case. Studentizing tracks the raw test to within 0.008 at every point; the four-fifths rule stays low only because it is too insensitive to detect anything. **The subgroup Cox calibration test is the flat orange curve:** it is measured on the same replicates and the same admissible levels, and its false-flag rate stays within 0.032–0.066 across the whole sweep, mean 0.0499, with no trend in ρ . Its flatness is not insensitivity: on a matched positive control at a fixed true gap it separates a genuinely unfair model from a fair one with case mix, where every curve above it cannot (Table 3). Source: `casemix_sweep.csv`.

2163

2164

2165 **Table 2 Flag rates for a perfectly fair model under case-mix heterogeneity.** One shared
 2166 model with identical coefficients, which *is* the data-generating probability: Bayes-optimal and
 2167 perfectly calibrated in every subgroup. Subgroups differ only in the location or spread of the linear
 2168 predictor, which makes their true AUROCs differ. Every entry is therefore a false alarm about
 2169 fairness. Nominal $\alpha = 0.05$; 1,000 simulations per cell, $B = 999$; reference admissibility rule `m30`.
 2170 Column headings give the case-mix magnitude ρ of each geometry, how many of the 21 measured
 2171 clinical partitions of Table 1 reach that magnitude, and the true AUROC gap it induces. **No**
 2172 **column is the typical case:** the median measured partition sits at $\rho = 1.145$, between the first
 2173 and second columns.

	Location only	Mild	Moderate	Moderate 3 part.	Moderate $n=10,000$	Strong
2174 Case-mix magnitude ρ	1.0	1.222	2.0	2.0	2.0	3.167
2175 Real partitions reaching it	21/21	9/21	1/21	1/21	1/21	1/21
2176 True AUROC gap	0.017	0.036	0.123	0.123	0.123	0.199
2177 Fixed 0.05 cut-point	0.6420	0.5060	0.9910	0.9930	1.0000	1.0000
2178 Four-fifths ratio rule	0.0310	0.0000	0.1190	0.1020	0.0060	0.7910
2179 Double-corr. variance, z-test	0.1130	0.2080	0.9460	0.9180	1.0000	0.9990
2180 Double-corr. variance, Cochran Q	0.0910	0.1700	0.9260	0.8600	1.0000	0.9990
2181 Double-corr. variance, bootstrap CI	0.0000	0.0000	0.0130	0.0000	0.9720	0.7570
2182 Raw max-min range permutation	0.2260	0.1620	0.9200	0.8520	1.0000	0.9990
2183 Studentized permutation	0.0440	0.1640	0.9190	0.8640	1.0000	0.9990

2182 Geometries, by per-level predictor standard deviation: *location only*, three levels at means $-2/0/+2$ and
 2183 equal spread 1.0; *mild*, 0.9/1.0/1.1, true AUROC 0.726/0.745/0.762; *moderate*, 0.7/1.0/1.4, true AUROC
 2184 0.684/0.745/0.807; *strong*, four levels, 0.6 to 1.9, true AUROC 0.661 to 0.860. The names are labels for
 2185 positions on the sweep of Figure 1 and carry no claim about which is common; the empirical distribution of
 2186 ρ is in Table 1. *Moderate 3 part.* embeds the moderate geometry as the first of three partitions, the other
 2187 two being pure noise columns. All at $n = 2,000$ and prevalence 0.20 except the $n = 10,000$ column. True
 2188 AUROCs and gaps are computed by quadrature, not simulated, and carry no Monte Carlo error.

2187 The four-fifths rule's low rates are insensitivity, not validity: a ratio of two AUROC values lies near 1.0 for
 2188 any usable pair, and the rule's rate *falls* to 0.0060 as n rises to 10,000 at an unchanged true gap.

2188 Flag rates are from `comparator_type1.csv`; true per-level AUROC values and the closed-form consistency
 2189 check are from `casemix_theory.json`.

2189

2190

2191

2192 clinical partitions we measured, by a wide margin in the tail. No amount of Type I
 2193 calibration under an equal-AUROC null repairs this, because the repair addresses the
 2194 reference distribution while the defect is in the hypothesis.

2196

2197 The second statement is about magnitude, and it is a distribution rather than
 2198 a number, and a distribution whose location we cannot pin down. At the median
 2199 measured partition a fair model would show a subgroup AUROC gap of between 0.013
 2200 and 0.030 on case mix alone depending on which specification is fitted, which is real
 2201 but well under the 0.05 threshold at which a disparity is conventionally declared, and
 2202 about half of which is the noise floor of a max-minus-min statistic. The number of
 2203 partitions clearing that threshold without any unfairness present runs from 2 to 7 of
 2204

Table 3 The recommended alternatives, measured on the same sweep and the same positive control. Panel A: across the full 16-point case-mix sweep, the subgroup Cox calibration test’s false-flag rate is flat at nominal while the AUROC-gap procedures climb to certainty. Panel B: matched pairs holding the true AUROC gap fixed and varying only the mechanism. Nominal $\alpha = 0.05$; 1,000 simulations per cell, $B = 999$, rule `m30`. In every Panel A cell and every case-mix arm of Panel B the model *is* the data-generating probability, so every flag there is a false alarm.

Case-mix	True AUROC	Raw max–min	Fixed 0.05	Cox calib.	MBC excess
ρ	gap	permutation	cut-point	test	(statistic) ^a
1.000	0.000	0.056	0.295	0.056	0.019
1.100	0.017	0.067	0.313	0.047	0.019
1.222	0.036	0.141	0.490	0.046	0.019
1.250	0.040	0.200	0.552	0.053	0.020
1.350	0.054	0.297	0.629	0.051	0.019
1.450	0.067	0.405	0.779	0.032	0.020
1.500	0.073	0.481	0.828	0.041	0.019
1.600	0.085	0.617	0.887	0.061	0.020
1.750	0.100	0.784	0.961	0.051	0.019
1.900	0.114	0.870	0.972	0.066	0.020
2.000	0.123	0.918	0.988	0.057	0.019
2.250	0.143	0.976	0.999	0.035	0.020
2.500	0.160	0.995	1.000	0.048	0.019
2.750	0.176	0.996	1.000	0.057	0.020
3.000	0.189	1.000	1.000	0.045	0.018
3.167	0.198	1.000	1.000	0.052	0.020
Mean over the 16 points				0.0499	0.0194

Panel B. Positive control: same true gap, different mechanism.

Procedure	True gap 0.050		True gap 0.123		True gap 0.200		Unfair, no gap ^b
	case mix	unfair	case mix	unfair	case mix	unfair	
Fixed 0.05 cut-point	0.677	0.684	0.997	0.998	1.000	1.000	0.266–0.313
Four-fifths ratio rule	0.003	0.002	0.294	0.319	0.974	0.970	0.000
Double-corr. variance, <i>z</i> -test	0.377	0.367	0.970	0.967	1.000	1.000	0.076–0.078
Raw max–min range permutation	0.314	0.297	0.948	0.950	1.000	1.000	0.043–0.044
Studentized permutation	0.305	0.296	0.946	0.947	1.000	1.000	0.050–0.058
Subgroup Cox calibration test	0.049	0.078	0.051	0.293	0.047	0.502	1.000

^a **Not a rejection rate.** Model-based-concordance excess is the maximum over levels of model-based concordance minus observed AUROC, in AUROC units, with no threshold, reference distribution or error rate; the column is its mean over 1,000 replicates. Its flatness against a true gap growing to 0.198 is the result. The value near 0.019 is a small finite-sample positive bias and is not a false-flag rate below nominal.

^b Two geometries in which the model is genuinely unfair but the true AUROC gap is exactly zero: a subgroup-specific calibration slope of 2.0, and a subgroup-specific intercept shift of one logit. The range spans the two. **Both perturb the intercept and slope that the calibration test itself estimates**, so they are correctly-specified alternatives for that test and are not evidence of general sensitivity.

Cox calibration test, Panel A: range 0.032–0.066, no trend in ρ , against a Monte Carlo standard error near 0.0069. The prevalence-equalised sweep gives 0.039–0.060, so the flatness is not a prevalence artefact.

Panel B, matched pairs: the case-mix and unfairness arms are constructed to agree on true AUROC to level of one decimal place. In the case-mix arm the model is the case data-generating probability in the unfairness arm level of outcome is partly driven by a covariate the model does not use. Each of the five AUROC-gap procedures shown gives the same rate on both arms, i.e. no discriminant validity. **Cochran’s *Q* and the bootstrap-interval reading produced no evaluable replicates on any positive-control geometry** and are omitted, so Panel B covers five of the seven AUROC-gap procedures. A sixth geometry combining both mechanisms at a total gap of 0.123, of which 0.076 is genuine excess, is flagged by the calibration test at 0.092.

Panel A is from `casemix.sweep.csv`; Panel B is from `casemix.positive.control.csv`.

Table 4 Type I error at nominal $\alpha = 0.05$ under equal-AUROC nulls. Rejection probability where every subgroup has the same true AUROC. 1,000 independent simulations per cell, $B = 999$; 12 simple-null cells and 22 composite-null cells (6 and 11 geometries, each crossed with the `m30` and `ev10` admissibility rules). “Simple” assigns subgroup membership independently of score and outcome; “composite” gives every level the same true AUROC but a different score distribution. **The ranking of procedures in this table should not be read as a ranking of the published methods**, for the reasons given in Sections 2.2 and 3.6. What survives is the coarse fact that two procedures in wide use are badly calibrated even against the null they nominally test, and, independently of any ranking, the finding of Table 2 that *every* AUROC-gap procedure here fails under case mix, so no ordering among them would rescue the family.

Procedure	Simple null (12 cells)		Composite null (22 cells)		Worst cell
	mean	max	mean	max	
Fixed 0.05 cut-point	0.7997	0.9990	0.6800	0.9600	0.999
Four-fifths ratio rule	0.1056	0.4660	0.0384	0.1430	0.466
Double-corr. variance, z-test	0.0948	0.1670	0.0927	0.1680	0.168
Double-corr. variance, Cochran Q	0.0808	0.2070	0.0580	0.0910	0.207
Double-corr. variance, bootstrap CI	0.0038	0.0170	0.0006	0.0080	0.017
Raw max–min range permutation	0.0490	0.0580	0.0293	0.0500	0.058
Studentized permutation	0.0486	0.0690	0.0480	0.0590	0.069

Monte Carlo standard error across the 34 equal-AUROC cells this table reports: median 0.00686, maximum 0.01577. Treating the 34 cells as independent, the 95th percentile of the maximum of 34 such rates from a correctly sized procedure at 1,000 simulations is 0.072; accounting for the paired structure (each of 17 geometries contributes one `m30` and one `ev10` cell computed on the same simulated datasets, hence positively correlated) it is 0.0700. The two permutation procedures’ worst cells (0.058 and 0.069) are inside both bands, but the studentized test’s margin under the relevant paired-structure band is 0.069 against 0.0700, not against the wider independence-based 0.072, and this thin margin is not evidence of miscalibration.

The fixed cut-point and the four-fifths rule state no error rate; their entries are implied false-flag rates, not Type I error against a claimed level.

These are the procedures’ properties under the null they nominally test. Table 2 shows that this ranking does not survive case mix.

Source: `comparator_type1.csv`.

21, and the largest implied gap from 0.131 to 0.219. A summary that reports only the median understates the tail. A summary that reports only the tail, which is what our own earlier draft did by taking $\rho = 2.0$ as the headline, overstates the typical case. A summary that reports either as a point, which is what the version before this one did, overstates what we know.

The mechanism is not subtle and was not discovered here. AUROC is the probability that a randomly chosen case scores above a randomly chosen control from the *same* subgroup [41], so it is a joint function of the model and the subgroup’s covariate distribution; widen a subgroup’s predictor distribution and its cases and controls become easier to tell apart, so its AUROC rises with the model untouched. This point

Table 5 Which clinical cohorts each procedure flags, by admissibility rule, unadjusted and after Holm correction. Identity, not count. Cohort codes: B BRFSS 2024, D Diabetes 130, A NHANES 2021–2023, N3 NHIS 2023, N4 NHIS 2024. UCI Heart Disease is not evaluable under any rule and is omitted. $B = 10,000$, joint permutation scheme, seed 42. “Raw” is the unadjusted per-cohort decision at $\alpha = 0.05$; “Holm” is the decision after Holm’s step-down correction [65] across the five estimable clinical cohorts, applied where the procedure reports a p -value. The last column counts clinical cohorts whose *raw* verdict is not constant across the five rules, of 5 estimable. The four non-clinical contrast cohorts are in Additional file 1.

Procedure		m20	m30	m50	m100	ev10	Clin.
Fixed 0.05 cut-point	Raw	B D A N3 N4	B D A N3 N4	B D A N3 N4	B D A N3 N4	B D N3 N4	1/5
	Holm	n/a	n/a	n/a	n/a	n/a	
Four-fifths ratio rule	Raw	B D N4	B D N4	B D N4	D N4	B D N4	1/5
	Holm	n/a	n/a	n/a	n/a	n/a	
Double-corr., z -test	Raw	B D A	B D A	B D A	D	B D N4	3/5
	Holm	B D A	B D A	B D A	D	B D N4	
Double-corr., Cochran Q	Raw	B D A N3 N4	B D A N3 N4	B D A N3 N4	D N3 N4	B D N3 N4	2/5
	Holm	B D A N3 N4	B D A N3 N4	B D A N3 N4	D N3 N4	B D N3 N4	
Double-corr., bootstrap CI	Raw	—	—	—	—	—	0/5
	Holm	—	—	—	—	—	
Raw max–min range perm.	Raw	—	—	—	N3	N3 N4	2/5
	Holm	—	—	—	—	—	
Studentized permutation	Raw	B D N4	B D N4	B D N4	D N3 N4	B D N3	3/5
	Holm	B D	B D	B D	D	B D	

“—” means no clinical cohort flagged; “n/a” means the procedure reports no p -value, so Holm correction does not apply (Section 2.2). The bootstrap-interval reading flags no cohort under any rule, raw or Holm, which is why its verdict is trivially constant.

The studentized permutation test’s raw decision returns a count of three under every rule, but three *different* sets; a count-based stability metric would record this as perfect stability, which is why we do not use one. Under the strict variant, which treats non-evaluability as a distinct verdict, the clinical counts are unchanged.

Running the independent permutation scheme instead of the joint scheme changes no cohort’s raw verdict for either permutation procedure under any rule. This is not a test of a dependence assumption; both schemes combine cohorts under the same independence assumption.

Holm correction across the five estimable clinical cohorts nullifies the raw max–min range permutation test’s flags entirely, but does not nullify the other three inferential procedures: the studentized test survives at Diabetes 130 and BRFSS 2024 under m20, m30, m50 and ev10, and at Diabetes 130 alone under m100; the double-corrected z -test and Cochran’s Q are unchanged from their raw decisions at every rule shown.

Source: `rule_stability.csv`.

2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346

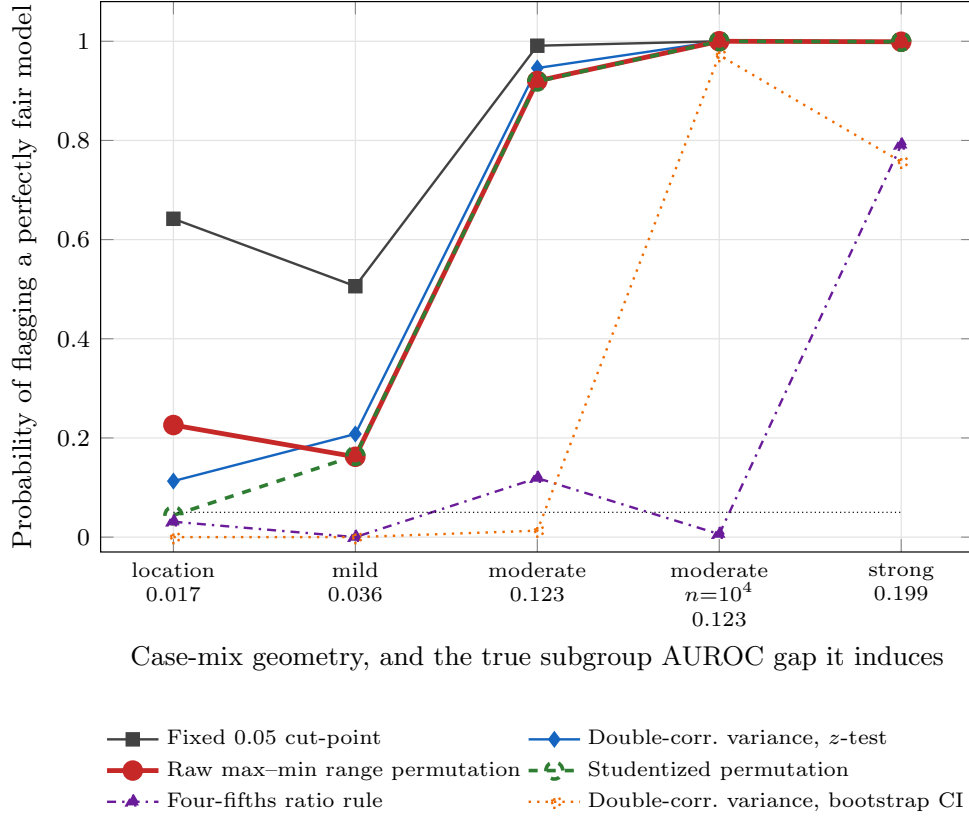


Fig. 2 At sufficient case-mix magnitude every procedure flags a perfectly fair model. The five named geometries, complementing the continuous sweep of Figure 1; their case-mix magnitudes ρ are 1.0, 1.222, 2.0, 2.0 and 3.167, of which the 21 measured clinical partitions reach 1.222 nine times and 2.0 once (Table 1). The model in every cell is the exact data-generating probability, with one coefficient vector, Bayes-optimal and perfectly calibrated in every subgroup; subgroups differ only in the distribution of the linear predictor. The dotted horizontal line is the nominal 0.05. Studentizing repairs the *location-only* geometry (0.226 to 0.044), where the true AUROC gap is a negligible 0.017, and does nothing at the moderate spread geometry, where the raw and studentized tests give 0.920 and 0.919. Going from $n = 2,000$ to $n = 10,000$ at an unchanged true gap of 0.123 drives four procedures to 1.000 and the bootstrap interval from 0.013 to 0.972: the failure is a property of the estimand, so more data makes it worse, not better. The four-fifths rule stays low only because it is too insensitive to detect anything, and reaches 0.791 at the strong geometry. Flag rates are from `comparator_type1.csv`; true per-level AUROC values are from `casemix_theory.json`.

was made for external validation [45, 46, 51]; what Table 2 and Figure 1 add is the magnitude in the fairness setting, the fact that it is confined to no one procedure, and, in Table 1, how much of it there is in cohorts of the kind these audits are run on. The $n = 10,000$ column is what we would ask a sceptical reader to look at first: if

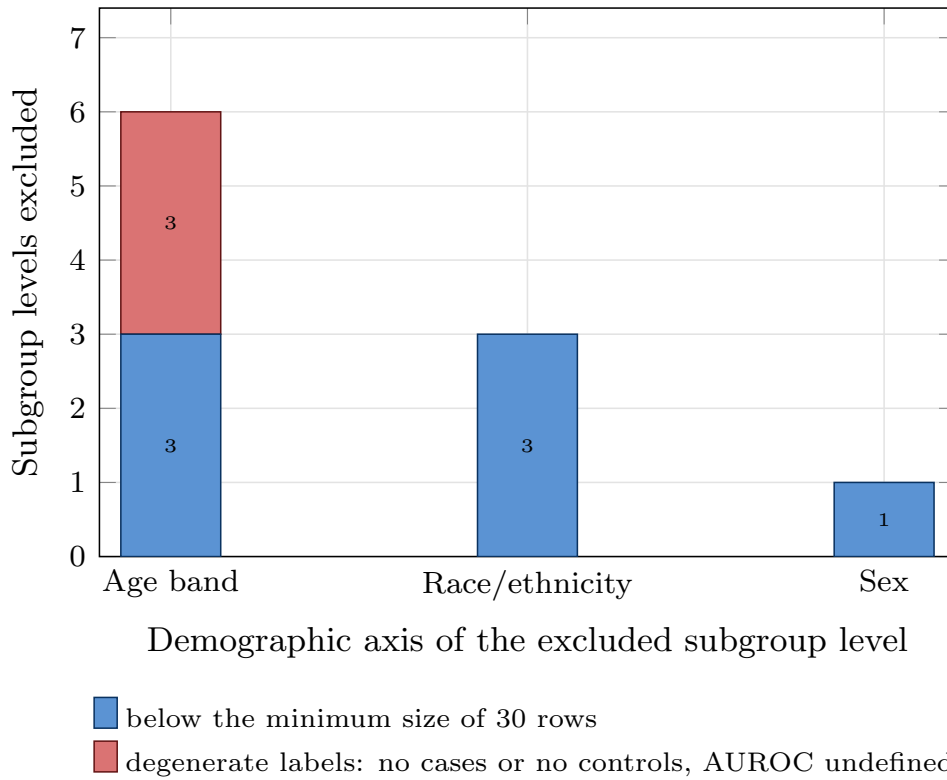


Fig. 3 Admissibility rules preferentially silence minority strata. All ten subgroup levels excluded across the ten cohorts at the reference rule `m30`, by demographic axis and reason. Every one of the three race and ethnicity exclusions is a minority category in NHIS 2024: NH-AIAN ($n = 23$, no cases), NH-AIAN+other ($n = 17$, one case) and NH-Other/Multi ($n = 18$, no cases). The sex exclusion is female patients in UCI Heart Disease ($n = 19$). An admissibility rule removes the smallest levels first; in cohorts drawn from a majority population the smallest levels are the minority categories, so the rule decides whose performance an audit is capable of describing at all. Source: `excluded_subgroups.csv`.

the flagging were noise, more data would reduce it, and instead five procedures go to 1.000 at an unchanged true gap. That is the signature of a consistent estimator of the wrong quantity, and it means the problem cannot be managed by powering the audit better or collecting more patients, both of which make it worse.

2439 **Engaging Pfohl et al.**

2440

2441 Pfohl et al. [52] argue, through causal graphical models, that equal subgroup per-
2442 formance is an unreliable fairness target even under representative sampling and a
2443 correctly specified null, and that under selection bias disaggregated evaluation and its
2444 conditional-independence alternatives may both be invalid absent explicit assumptions
2445 about the bias mechanism. The natural defensive reading of that paper, that this is a
2446 sampling problem, to be handled by better data or sensitivity analysis, is the one our
2447 results close off. Our case-mix simulations have no selection bias, no unrepresentative
2448 sampling, no misspecification, no measurement error and no label bias; the score is
2449 the true conditional probability, and the only thing differing between subgroups is the
2450 covariate distribution, which alone produces flag rates from 0.056 at $\rho = 1.0$ to 1.000
2451 at $\rho = 3.0$. Their argument is therefore stronger than a sampling caveat: the target
2452 is unreliable in the base case. How badly it fails in any given audit is then an empir-
2453 ical question about that audit’s cohort and partition. Table 1 answers it for ours, as
2454 a range rather than a number. A fair model would show a subgroup AUROC gap of
2455 0.013 to 0.030 at the median partition and 0.131 to 0.219 at the worst, so the con-
2456 tamination is small in most of our partitions and larger than the reporting threshold
2457 in somewhere between a tenth and a third of them.

2469 We add one refinement specific to AUROC that a general performance-parity argu-
2470 ment does not reach. AUROC has an uninformative point at 0.5, not at 0.0, so a ratio
2471 scale on it is meaningless: a four-fifths ratio of 0.80 corresponds to one subgroup at
2472 0.500, pure chance, while another sits at 0.625, and also to one at 0.720 while another
2473 sits at 0.900. The rule is simultaneously blind to a subgroup for whom the model is
2474 useless and triggerable by pairs of values that are both clinically acceptable. Nothing
2475 about the number 0.80 survives the change of scale from selection rates, which have a
2476 meaningful zero, to a concordance probability, which does not: an objection separate
2477 from the Type I error objection of Greenberg [10], Boardman [11], Roth et al. [15].

2484

What to do instead

Who these recommendations are for. They are addressed to the people who build, audit and sign off on clinical prediction models: model developers, evaluation and validation teams, and the governance committees that approve deployment. **Clinicians and patients are out of scope as audiences.** We have not studied how a subgroup performance report should be presented to a clinician at the point of care, and we make no claim about what a patient should be told about subgroup performance. Those are questions about communication and consent that this paper’s evidence does not reach, and they matter enough that they should not be answered by inference from a simulation study.

Do not use a subgroup AUROC-gap test as fairness evidence for a clinical prediction model. This applies to all seven procedures studied here, including the two that are correctly sized under an equal-AUROC null and including the one our own earlier analysis recommended. It applies whether the gap is compared to a cut-point, a ratio, or a permutation distribution.

If a gap statistic is reported at all, report it descriptively and next to the case-mix distribution. A subgroup AUROC table is informative about how well the model separates each subgroup’s cases from its own controls, which is a real and reportable property; it becomes misleading the moment it is converted into a hypothesis test whose null is equality. Report each subgroup’s AUROC with its interval, and beside it the distribution of the predicted probability within that subgroup, spread as well as location, so a reader can see whether an observed difference is what a shared, fair model would produce.

Compute the implied case-mix gap beside every observed gap, and weight the observed gap accordingly. The per-level mean and standard deviation of the linear predictor are one line of code on data an audit already has, and Eq. (2) turns them into the gap a fair model would show, in the same units as the gap being reported,

2531 with no simulation. Compute it from the observed scores rather than from their mean
2532 and standard deviation: the identity needs no distributional assumption and the two-
2533 parameter plug-in adds one that moves individual partitions by up to 0.037. Compute
2534 a noise-only reference beside it, by permuting the level labels, because a maximum
2535 minus a minimum is positive whether or not any case mix is present. Our own parti-
2536 tions show what that buys. In these cohorts an audit of a sex or gender partition faces
2537 an implied gap of at most 0.024 in the published fit, so a gap found there is unlikely
2538 to be an artefact of case mix and deserves to be taken seriously on its own terms. An
2539 audit of an age-band partition faces between 0.034 and 0.173 in that fit, and four of
2540 the five clear the 0.05 convention on case mix alone, so a gap found there should not
2541 be reported as a finding without the case-mix geometry beside it.

2542 We put four limits on that advice. The age-versus-sex ordering is confounded with
2543 level count, which we cannot separate from it (Section 3.1). It does not survive a change
2544 of admissibility rule: under `ev10` the age partitions run from 0.013 to 0.173 and the
2545 count clearing the convention falls to three of five, so the age and sex ranges overlap.
2546 The numbers themselves are one specification's, and across 24 the per-partition ranges
2547 are wide enough that no partition's implied gap should be quoted as a point. And it
2548 is a property of 21 partitions in five fitted models, not a law: another cohort, another
2549 feature set or another outcome could put age and sex anywhere, which is exactly why
2550 the advice is to compute the implied gap rather than to memorise our ordering.

2551 *A decision threshold, since an ordering without one is not actionable.* We suggest
2552 the following, and we mark it as a proposal rather than a validated rule because we
2553 have not evaluated its error rates. Where a partition's implied case-mix gap reaches the
2554 threshold at which the audit would declare a disparity, conventionally 0.05, report the
2555 subgroup AUROCs descriptively and do not report a gap test at all: at that geometry
2556 a fair model already produces the effect the test would be detecting. Below that, a
2557 gap test remains uninformative about fairness for the reasons this paper gives, but
2558

the case-mix contribution is at least small enough not to dominate. Under this rule between 2 and 7 of our 21 partitions would be reported descriptively only, depending on which model is fitted, which is itself an argument for computing the rule on the audit’s own model rather than borrowing our count. We have not evaluated what this rule does to detection of genuine unfairness, and it should not be adopted as a standard on our say-so.

Report subgroup calibration and subgroup net benefit instead, and we have now measured the first of these rather than only citing it. This is not a proposal of ours, and we are explicit that it is not. What is new here is not the recommendation but the evidence for it. On the same 16-point sweep that drives every AUROC-gap procedure from 0.056 to 1.000, a subgroup Cox calibration test’s false-flag rate does not move at all: 0.032 to 0.066 across the whole range, mean 0.0499 against a nominal 0.05, with no trend in case-mix magnitude (Table 3, Figure 1). The reason is structural rather than incidental. The null “this model is calibrated within each subgroup” remains true for a fair model when subgroups differ in case mix; the null “true AUROC is equal across subgroups” does not. And the flatness is not the flatness of a test that never fires: on a matched positive control holding the true AUROC gap fixed at 0.123, it flags the case-mix arm at 0.051 and the genuinely unfair arm at 0.293, while the raw permutation test, the studentized test and the double-corrected z -test give indistinguishable rates on the two arms (0.946–0.970 against 0.947–0.967); the four-fifths rule and the fixed cut-point do not (0.294 against 0.319, and 0.997 against 0.998). On two geometries where the model is genuinely unfair but the AUROC gap is exactly zero, the gap procedures sit at nominal and the calibration test flags at 1.000. That is the operating-characteristic evidence the recommendation previously lacked, and it points the same way the guidance already did.

We are deliberately not converting this into a recommendation of our own. A subgroup calibration test inherits the levelling-down objection [81] that attaches to any

2623 equality criterion, says nothing about clinical utility, and assumes within-subgroup
 2624 calibration is estimable, which in a level carrying a few dozen events it is not; it is
 2625 our implementation, not a package; and we have measured it in simulation, not in
 2626 a deployed audit. The claim is comparative and bounded: under the conditions that
 2627 break the AUROC gap, calibration does not break. van Klaveren et al. [45]’s model-
 2628 based concordance behaves similarly on the sweep, its excess over observed AUROC
 2629 stays near 0.019 AUROC units while the true gap grows to 0.198, but it separates
 2630 the matched arms only slightly better than chance (0.545–0.590), so we report it as
 2631 flat rather than as useful. Van Calster et al. [58] review 32 performance measures
 2632 for predictive artificial intelligence in medicine and conclude that the essential set is
 2633 discrimination, a calibration plot, a clinical utility measure such as net benefit with
 2634 decision curve analysis, and the distribution of predicted probabilities by outcome.
 2635 Pfohl et al. [59] make the fairness-specific version of that argument: develop models cal-
 2636 ibrated and useful for every population, with thresholds reflecting patient preferences,
 2637 rather than constraining a performance metric to be equal. Benitez-Aurioles et al. [82]
 2638 assess fairness by subgroup net benefit precisely to avoid the levelling-down incentive
 2639 an equal-performance criterion creates, the failure mode Mittelstadt et al. [81] identify
 2640 as endemic to fairness measures built on narrowing between-group gaps rather than on
 2641 distributive-justice considerations, and Meerwijk et al. [83] found AUROC of limited
 2642 utility for subgroup fairness assessment in resource-limited decision support; decision
 2643 curve analysis has accessible primers [84, 85]. Our contribution is not the alternative
 2644 but the quantified demonstration of why the AUROC-gap route fails, which is the
 2645 part of the case existing guidance states without measuring. Subgroup calibration and
 2646 net benefit are not themselves immune to case mix, net benefit depends on subgroup
 2647 prevalence, so they too belong in a descriptive report beside the case-mix distribution
 2648 rather than in an equality test. Where the question is specifically whether discrimina-
 2649 tion is worse *than it should be given whom the model is applied to*, van Klaveren et al.
 2650
 2651
 2652
 2653
 2654
 2655
 2656
 2657
 2658
 2659
 2660
 2661
 2662
 2663
 2664
 2665
 2666
 2667
 2668

[45] give model-based concordance as the case-mix-corrected benchmark. We applied it to our own cohorts and could not get a stable answer out of it (Section 3.2), which is a caution about what that benchmark can deliver at these sample sizes rather than an objection to the quantity.

What to tell a compliance officer who mandates the four-fifths rule. Compute and report it; refusing a mandated number is rarely a winning position. What must accompany it is that the rule was written for selection rates, which have a meaningful zero, and AUROC does not, so 0.80 has no interpretation on this scale; that its output tracks the quantity of interest in neither direction, as it fired on a perfectly fair model in 79% of replicates at one case-mix geometry ($\rho = 3.167$, a magnitude one of our 21 measured partitions exceeds) while at another its rate *fell* from 0.119 to 0.006 as the sample size rose five-fold at an unchanged true gap; and that its false-positive behaviour on rates has been documented since 1979 [10, 11]. A four-fifths failure is not evidence of unfairness and a pass is not evidence of parity. The constructive substitute is the essential set current guidance already asks for [58], reported by subgroup, with the case-mix distribution beside it and a pre-specified account of what the organisation would do about a genuine difference.

Admissibility, and who it silences

Two secondary findings survive independently of the estimand argument, because they concern how an audit is assembled rather than what it concludes.

The first is that the subgroup-admissibility rule is a live researcher degree of freedom. Table 5 shows that every procedure which flags anything changes *which* cohort it flags as the rule moves. The procedure whose count is most stable, the studentized test at three under every rule, returns three different sets of three. Reporting the rule, the levels it excluded and the resulting minimum detectable effect should be minimum practice, and a cohort in which fewer than two levels per partition survive should be

2715 reported as not evaluable, as UCI Heart Disease is here, rather than assigned a num-
2716 ber. This is a specific instance of the multiverse problem Simson et al. [30] describe,
2717 creating multiplicity of the kind Mandl et al. [31] address.
2718

2720 *Who should receive that documentation.* A rule reported only inside the analytics
2721 team is not reported. The admissibility rule, the list of excluded levels with their
2722 sizes and event counts, and the resulting minimum detectable effect should go to
2723 whoever signs off on the model’s deployment: in most organisations that is the clinical
2724 governance or safety committee that approves the model for use, and where one exists
2725 the institutional review or ethics body that reviewed the underlying data. The point
2726 of naming a recipient outside the team is that an exclusion list is a statement about
2727 which patients the audit cannot describe, and that is a governance question rather
2728 than a statistical one.
2729

2735 The second finding is that admissibility rules are not demographically neutral
2736 (Figure 3). All three race and ethnicity levels excluded here are minority categories
2737 and two had no cases at all. Naming that and stopping is not useful. Three remedies
2738 are available. Hierarchical pooling of small levels, with the shrinkage made explicit,
2739 redistributes information across levels. Oversampling or purposive recruitment within
2740 small strata at the design stage is the only remedy that adds information rather than
2741 redistributing it. Explicit deferral reports the stratum as not evaluable together with
2742 the sample size that would be needed. What should not happen is silent deletion, the
2743 current default, which produces an audit that appears complete while being struc-
2744 turally unable to say anything about the groups most likely to be harmed. None of
2745 these remedies touches the case-mix problem; they address only who is measured.
2746

2753 *We do not evaluate a partial-pooling estimator.* Hierarchical pooling is the remedy
2754 we would expect to be tried first, and a subgroup AUROC could in principle be shrunk
2755 towards a pooled value under a random-effects model. We have not implemented one,
2756 have not measured what shrinkage does to the case-mix contamination documented
2757
2758
2759
2760

here, and make no claim about it. We note only that pooling changes the estimate and not the estimand: a shrunken subgroup AUROC is still a joint function of the model and the subgroup’s covariate distribution, so we would not expect pooling to repair the failure this paper documents. That expectation is untested.

Our own earlier conclusions

We have been wrong about this evidence base five times, and the corrections are stated in full in the Disclosure at the front of the paper rather than repeated here. Three of the five substituted an assumption for a measurement that was available; one implemented an estimator correctly and had it measure itself; one withdrew a quantity and kept its downstream product. What caught all five was recomputing the thing itself, and pre-registration would have preserved every one of them intact.

Limitations

The case-mix result is a simulation result, in the geometries simulated. Six named geometries and a 16-point sweep, at one or three partitions, a single prevalence apart from the varied cells, and a Gaussian linear predictor do not exhaust the space. We claim that these procedures fail under case mix at the magnitudes shown; we do not characterise generally when they do and do not fail.

The implied gap is not a flag rate, and the two should not be conflated. Table 1 reports the gap a fair model would show, in AUROC units. It does not report how often a procedure would fire on that gap, which depends on the partition’s sample size, level sizes and prevalence, none of which Eq. (2) uses. The $n = 10,000$ column of Table 2 shows how strong the sample-size dependence is. A per-partition simulation at each partition’s own size and level structure would give flag rates directly, and we have not run one. Readers who want a probability rather than an effect size will not

2807 find one here, and that is deliberate: the mapping that used to supply it was not
 2808 defensible (Section 2.8).
 2809
 2810 *$\hat{\rho}$ itself is unstable at the top of its range, and so is the implied gap.* Across 24 refits
 2811 $\hat{\rho}$'s maximum runs from 1.999 to 6.388, with model classes disagreeing without overlap,
 2812 so any statement of the form “the worst partition carries $\hat{\rho} = x$ ” is not supported.
 2813
 2814 Which partition is worst is stable; how much worse it is, is not. The implied gap
 2815 inherits this, because it is computed from the same fitted linear predictor: across the
 2816 same specifications its median runs 0.013 to 0.030, its maximum 0.131 to 0.219, and
 2817 the count of partitions reaching 0.05 runs 2 to 7. Its maximum is the better behaved
 2818 of the two, and its count the worse. **No implied-gap value in this paper is an**
 2819 **estimate and none should be quoted as a point.**
 2820
 2821 *We report no decomposition of the observed gaps, and the reasons are worth sep-*
 2822 *arating.* One is precision: the fraction’s denominator is an observed gap that is often
 2823 near zero at these sample sizes, so the ratio is unstable and in one partition takes
 2824 the value -8.61 . One is conditioning: the bootstrap intervals are conditional on the
 2825 extreme level pair selected from the observed data, and do not cover the uncertainty
 2826 in *which* levels are extreme, so they are narrower than an unconditional interval would
 2827 be. One is an assumption we cannot check: model-based concordance assumes the pre-
 2828 dictions are calibrated within the subgroup it is computed on, and on real cohorts
 2829 they frequently are not, so where that assumption fails the case-mix component car-
 2830 ries a bias we do not quantify. On Diabetes 130 the interval is optimistic for a further
 2831 reason, because a patient can contribute several encounter rows to the same test set.
 2832 We therefore treat the decomposition as not estimable here rather than as estimated
 2833 imprecisely, and a reader should not reconstruct a fraction from anything in this paper.
 2834
 2835 *The closed-form implied gap of Table 1 rests on well-specification, which these*
 2836 *cohorts do not satisfy.* Across the 87 admissible clinical levels the implied and observed
 2837 level AUROC correlate at 0.852 with a median absolute difference of 0.031 and a range
 2838
 2839
 2840
 2841
 2842
 2843
 2844
 2845
 2846
 2847
 2848
 2849
 2850
 2851
 2852

of -0.233 to $+0.128$. The Gaussian plug-in adds a second assumption whose cost we now quantify rather than assume away: it moves individual partitions by up to 0.037 against the assumption-free evaluation. The implied gap is a reference value for a fair model with the measured predictor geometry, not an estimate of the case-mix share of the observed gap. We report no observed-minus-implied contrast: the closed form’s directional prediction matches the observed extremes in only 7 of 21 partitions, against about 5 expected by chance, so the two ranges cannot be subtracted. We also report the implied gap against a permutation reference rather than against zero, because it is a maximum minus a minimum and is positive by construction; that reference has median 0.010 , and only 9 of the 21 partitions clear its 95th percentile.

The two alternatives are measured in simulation only, and are our implementations. The flat false-flag rate of the subgroup Cox calibration test is established on the same synthetic sweep as everything else in Table 2, with a Gaussian linear predictor, three equal levels and $n = 2,000$; we have not measured it on the clinical cohorts, and its power under genuine unfairness is established on two constructed mechanisms rather than on any real form of model unfairness. Its assumption that within-subgroup calibration is estimable fails in exactly the small, sparse-event levels that Section 3.8 shows an audit is most likely to drop. We report it as evidence that the recommended alternative does not inherit the failure documented here, not as a validated procedure, and we do not recommend it as a fairness test.

Nothing here is evidence that any cohort is free of subgroup differences, and no such claim is made. The cohort analyses illustrate how verdicts move with analyst choices. They are not findings about the cohorts, for the reason set out in Section 2.6.

Retraining variation is not captured. All models use fixed hyperparameters and seed 42, and every permutation null conditions on the fitted scores, so variation across retraining seeds [27] and the instability of AUROC across resamples [29] do not enter any interval reported here. The one place we do vary the fit is the 24-specification refit

of Section 3.1, and there the variation was large enough to withdraw a number from $\hat{\rho}$ and three more from the implied gap.

No procedure is recommended for this task. We do not recommend the studentized test, the raw range, any of the three readings of the double-corrected variance estimator, the fixed cut-point, or the four-fifths rule. Nor do we recommend measuring $\hat{\rho}$ as a test: it is a descriptive quantity that tells an auditor how much weight a gap statistic can bear, and it has no error rate because it is not an inference. Of the alternatives we point to, subgroup calibration, subgroup net benefit, case-mix-corrected concordance, all are taken from existing guidance and none is proposed here. Two of them now carry measured operating characteristics under case mix (Table 3), which is a narrower claim than a recommendation: it establishes that they do not inherit this particular failure, not that they are adequate fairness criteria. Subgroup net benefit is not evaluated anywhere in this manuscript and we make no claim about it. Nothing in this paper licenses a claim about the deployment, clinical utility or quality of any model.

5 Conclusions

Subgroup fairness audits of clinical prediction models test the hypothesis that true AUROC is equal across subgroups. That hypothesis is false for a perfectly fair model whenever subgroups differ in case mix, and how badly the audit is misled is a steep, monotone function of how much they differ. We map that function with a model that is the exact data-generating probability, one coefficient vector, Bayes-optimal, perfectly calibrated in every subgroup, and subgroups differing only in the spread of the predictor. The raw max-min range permutation test's false-flag rate runs from 0.056 where the spreads are equal, through 0.141 at a spread ratio of 1.222 and 0.918 at 2.0, to 1.000 at 3.0; studentizing tracks it to within 0.008 throughout, and raising n from 2,000 to 10,000 at a fixed true gap drives five procedures to 1.000.

We then measured that spread ratio where it matters, in 21 demographic partitions of our own five evaluable clinical cohorts. Its median is 1.145 in the published fit, and across 24 refits the median runs 1.132 to 1.223. The maximum is not stable enough to quote as a number and we report it as a range, 1.999 to 6.388. Feeding each partition's own measured geometry through the closed form gives the gap a perfectly fair model would show there, and it is no more stable than the ratio. Across the same 24 refits its median runs 0.013 to 0.030, its maximum 0.131 to 0.219, and the number of partitions reaching 0.05, the threshold at which a disparity is conventionally declared, runs from 2 to 7 of 21. Under a different admissibility rule the published-fit count is 4 of 19. Against a permutation reference that holds every level at one common predictor distribution, 9 of the 21 partitions clear the 95th percentile, and the reference's own median is about half the median implied gap. We therefore quote no point value for any of these quantities. Contamination at the median is modest and partly noise. In the tail it is real and on its own large enough to trip the convention, with no unfairness present anywhere in the model.

We could not decompose the observed gaps, and that is a result rather than a gap in the work. Applying the case-mix-corrected benchmark the literature supplies, with the model and data in hand and 2,000 bootstrap replicates per partition, the attributable fraction is reportable in only 15 of 33 distinct partitions and reaches -8.61 in one of them. We withdraw the 93% an earlier version of this analysis reported, whose estimator was near-tautological, and we withdraw the 29% that replaced it. The residual gap, which is a difference rather than a ratio, we do report: three of the 21 clinical partitions carry a positive residual whose conditional interval excludes zero and whose implied gap is non-negative. An observed gap mixes a genuine component with a case-mix component in a proportion we could not recover. That is the precise reason a gap statistic cannot be read as fairness evidence: on a matched control holding the true gap fixed, no AUROC-gap procedure distinguishes a fair model with case

2991 mix from a genuinely unfair one, and none detects unfairness that leaves AUROC
2992 untouched.
2993

2994 We therefore recommend that subgroup AUROC-gap tests not be used as fairness
2995 evidence for clinical prediction models. Where a gap is reported, it should be reported
2996 descriptively, with intervals, alongside the case-mix distribution of each subgroup, and
2997 never as a hypothesis test. What should be reported instead, subgroup calibration
2998 and subgroup net benefit, is what current guidance on evaluating predictive artificial
2999 intelligence in medicine already recommends [58, 59, 82], and we are not proposing
3000 a new measure. We have, however, now measured one of them against this failure: a
3001 subgroup Cox calibration test holds its nominal false-flag rate across the entire case-
3002 mix sweep, mean 0.0499, while separating a genuinely unfair model from a fair one
3003 at the same true gap. The recommendation is no longer only cited; it is evidenced.
3004 An earlier version of this work presented the ratio-2.0 geometry, and the 0.918 rate
3005 it produces, as the ordinary clinical case; measuring the quantity in our own cohorts
3006 showed it is not, and the distribution that replaced it is a more useful result than the
3007 number it cost. Whether any particular model is unfair to any particular subgroup
3008 remains a question this class of test cannot settle, and we do not settle it.
3009
3010

3011 **Declarations**

3012
3013 *Ethics approval and consent to participate.*
3014

3015 Not applicable. All analyses use publicly available de-identified data or synthetic data,
3016 and no human participants were recruited.
3017

3018
3019 *Consent for publication.*
3020

3021 Not applicable.
3022
3023
3024
3025
3026
3027
3028
3029
3030
3031
3032
3033
3034
3035
3036

Availability of data and materials.

All cohorts are public and are obtained from the cited sources: NHIS 2024, NHIS 2023 and NHANES 2021–2023 from the National Center for Health Statistics, BRFSS 2024 from the Centers for Disease Control and Prevention, Heart Disease, Diabetes 130-US Hospitals, Adult Income and Statlog German Credit from the UCI Machine Learning Repository, and ACS-Income from Folktables. The synthetic cohort is deposited at [doi:10.57967/hf/8734](https://doi.org/10.57967/hf/8734) [74]; no real patient records were used in it at any stage. The analysis code, the simulation harness, the per-cohort permutation nulls and the result files underlying every table and figure are archived in the deposit listed below, in the tagged release that corresponds to this manuscript. Every table and figure in the paper can be regenerated from that release using the commands documented in the repository README.

- **Project name:** rised-healthcare-eval
- **Project home page:** <https://github.com/rohithreddybc/rised-healthcare-eval>
- **Archived version:** 10.5281/zenodo.21918249 (<https://doi.org/10.5281/zenodo.21918249>)
- **Operating system(s):** Platform independent. All results reported here were produced on Windows 11 running Python 3.11.7.
- **Programming language:** Python
- **Other requirements:** Python 3.9 or later, with numpy 1.26.4, scipy 1.15.3, pandas 3.0.2, scikit-learn 1.8.0, xgboost 3.2.0 and fairlearn 0.13.0. Pinned versions for the full dependency tree are given in `requirements-lock.txt` in the repository.
- **License:** MIT
- **Any restrictions to use by non-academics:** None

Competing interests.

The authors declare that they have no competing interests.

3083 ***Funding.***

3084
3085 This work received no specific grant from any funding agency in the public, commercial
3086 or not-for-profit sectors. No funder had any role in the design of the study, in the
3087 collection, analysis or interpretation of data, or in writing the manuscript.
3088

3090
3091 ***Authors' contributions.***

3092
3093 RRB conceived the study, designed and implemented the simulation and comparator
3094 code, ran all analyses, and drafted the manuscript.
3095

3096 MS designed and implemented the simulation and comparator code, ran some analy-
3097 ses, drafted and reviewed the manuscript.
3098

3099 YJ contributed to the design of the null geometries and the admissibility sweep and
3100 reviewed the analysis code.
3101

3102 AI contributed to the cohort construction and the interpretation of the case-mix
3103 results.
3104

3105
3106 RJ helped in writing, reviewing, and supervised the study.
3107

3108 All authors read and approved the final manuscript.
3109

3110

3111

3112

3112 ***Acknowledgements.***

3113

3114 Not applicable.
3115

3116

3117

3117 **References**

3118

3119

3120 [1] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al.

3121 TRIPOD+AI statement: updated guidance for reporting clinical prediction mod-
3122 els that use regression or machine learning methods. BMJ. 2024;385:e078378.

3123

3124 <https://doi.org/10.1136/bmj-2023-078378>.
3125

3126

3127

3128

- [2] Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*. 2025;388:e081554. <https://doi.org/10.1136/bmj-2024-081554>.
- [3] Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*. 2019;63(4/5):4:1–4:15. <https://doi.org/10.1147/JRD.2019.2942287>.
- [4] Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, et al. Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft; 2020. MSR-TR-2020-32.
- [5] Panigutti C, Perotti A, Panisson A, Bajardi P, Pedreschi D. FairLens: Auditing black-box clinical decision support systems. *Information Processing & Management*. 2021;58(5):102657. <https://doi.org/10.1016/j.ipm.2021.102657>.
- [6] U S Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Department of Health and Human Services; 2021. Available from: <https://www.fda.gov/media/145022/download>.
- [7] Equal Employment Opportunity Commission and others. Uniform Guidelines on Employee Selection Procedures, 29 CFR 1607.4(D). *Federal Register* 43(166):38290–38315; 1978.
- [8] DiCiccio C, Vasudevan S, Basu K, Kenthapadi K, Agarwal D. Evaluating fairness using permutation tests. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*; 2020. p. 1467–1477.

3175 [9] Lum K, Zhang Y, Bower A. De-biasing “bias” measurement. In: Proceed-
 3176 ings of the 2022 ACM Conference on Fairness, Accountability, and Transparency
 3177 (FAccT). Association for Computing Machinery; 2022. p. 379–389.
 3178
 3179
 3180
 3181 [10] Greenberg I. An analysis of the EEOCC “four-fifths” rule. *Management Science*.
 3182 1979;25(8):762–769. <https://doi.org/10.1287/mnsc.25.8.762>.
 3183
 3184
 3185 [11] Boardman AE. Another analysis of the EEOCC “four-fifths” rule. *Management*
 3186 *Science*. 1979;25(8):770–776. <https://doi.org/10.1287/mnsc.25.8.770>.
 3187
 3188
 3189 [12] Morris SB, Lobsenz RE. Significance tests and confidence intervals for the adverse
 3190 impact ratio. *Personnel Psychology*. 2000;53(1):89–111. [https://doi.org/10.1111/](https://doi.org/10.1111/j.1744-6570.2000.tb00195.x)
 3191 [j.1744-6570.2000.tb00195.x](https://doi.org/10.1111/j.1744-6570.2000.tb00195.x).
 3192
 3193
 3194
 3195 [13] Collins MW, Morris SB. Testing for adverse impact when sample size is small.
 3196 *Journal of Applied Psychology*. 2008;93(2):463–471. [https://doi.org/10.1037/](https://doi.org/10.1037/0021-9010.93.2.463)
 3197 [0021-9010.93.2.463](https://doi.org/10.1037/0021-9010.93.2.463).
 3198
 3199
 3200
 3201 [14] Besse P, del Barrio E, Gordaliza P, Loubes JM. Confidence intervals for testing
 3202 disparate impact in fair learning. *arXiv preprint*. 2018;[https://doi.org/10.48550/](https://doi.org/10.48550/arXiv.1807.06362)
 3203 [arXiv.1807.06362](https://doi.org/10.48550/arXiv.1807.06362).
 3204
 3205
 3206
 3207 [15] Roth PL, Bobko P, Switzer FS. Modeling the behavior of the 4/5ths rule for
 3208 determining adverse impact: Reasons for caution. *Journal of Applied Psychology*.
 3209 2006;91(3):507–522. <https://doi.org/10.1037/0021-9010.91.3.507>.
 3210
 3211
 3212 [16] Watkins EA, Chen J. The four-fifths rule is not disparate impact: A woeful
 3213 tale of epistemic trespassing in algorithmic fairness. In: *Proceedings of the 2024*
 3214 *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*; 2024.
 3215
 3216
 3217 p. 764–775.
 3218
 3219
 3220

- [17] Janssen A. Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem. *Statistics & Probability Letters*. 1997;36(1):9–21. [https://doi.org/10.1016/S0167-7152\(97\)00043-6](https://doi.org/10.1016/S0167-7152(97)00043-6).
- [18] Neubert K, Brunner E. A studentized permutation test for the non-parametric Behrens–Fisher problem. *Computational Statistics & Data Analysis*. 2007;51(10):5192–5204. <https://doi.org/10.1016/j.csda.2006.05.024>.
- [19] Chung E, Romano JP. Exact and asymptotically robust permutation tests. *The Annals of Statistics*. 2013;41(2):484–507. <https://doi.org/10.1214/13-AOS1090>.
- [20] Chung E, Romano JP. Multivariate and multiple permutation tests. *Journal of Econometrics*. 2016;193(1):76–91. <https://doi.org/10.1016/j.jeconom.2016.01.003>.
- [21] DiCiccio CJ, Romano JP. Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*. 2017;112(519):1211–1220. <https://doi.org/10.1080/01621459.2016.1202117>.
- [22] Pauly M, Asendorf T, Konietzschke F. Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data. *Biometrical Journal*. 2016;58(6):1319–1337. <https://doi.org/10.1002/bimj.201500105>.
- [23] Munko M, Sattler P. Covariance correction for permutation statistics in multiple testing problems. *arXiv preprint arXiv:260406915*. 2026;<https://doi.org/10.48550/arXiv.2604.06915>.
- [24] Roelofs R, Cain N, Shlens J, Mozer MC. Mitigating bias in calibration error estimation. In: *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*. vol. 151; 2022. p. 4036–4054. Available from: <https://proceedings.mlr.press/v151/roelofs22a.html>.

- 3267 [25] Nixon J, Dusenberry MW, Zhang L, Jerfel G, Tran D. Measuring calibration in
3268 deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision
3269 and Pattern Recognition Workshops (CVPRW); 2019. p. 38–41. Available from:
3270 <https://arxiv.org/abs/1904.01685>.
3271
3272
3273
3274 [26] Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G.
3275 Subgroup analyses in randomised controlled trials: quantifying the risks of false-
3276 positives and false-negatives. *Health Technology Assessment*. 2001;5(33). <https://doi.org/10.3310/hta5330>.
3277
3278
3279
3280
3281 [27] Amir S, van de Meent JW, Wallace BC. On the impact of random seeds on the
3282 fairness of clinical classifiers. In: Proceedings of the 2021 Conference of the North
3283 American Chapter of the Association for Computational Linguistics (NAACL);
3284 2021. p. 3808–3823.
3285
3286
3287
3288
3289 [28] Libin A, Alessa A, Patel M, Shara N. Evaluating and reducing subgroup disparity
3290 in AI models predicting pediatric COVID-19 test outcomes. In: Proceedings of
3291 the 2024 IEEE International Conference on Big Data. IEEE; 2024. p. 5159–5164.
3292
3293
3294
3295 [29] van Leeuwen FD, van Smeden M, Kent DM, van Klaveren D. Instability of the
3296 AUROC of clinical prediction models. *Statistics in Medicine*. 2025;44(5):e10352.
3297 <https://doi.org/10.1002/sim.10352>.
3298
3299
3300
3301 [30] Simson J, Pfisterer F, Kern C. One model many scores: Using multiverse analysis
3302 to prevent fairness hacking and evaluate the influence of model design decisions.
3303 In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and
3304 Transparency (FAccT); 2024. p. 1305–1320.
3305
3306
3307
3308 [31] Mandl MM, Becker-Pennrich AS, Hinske LC, Hoffmann S, Boulesteix AL.
3309 Addressing researcher degrees of freedom through minP adjustment. *BMC*
3310
3311
3312

- Medical Research Methodology. 2024;24(1):152. <https://doi.org/10.1186/s12874-024-02279-2>. 3313
3314
3315
3316
- [32] Ferrara A, Cozzi F, Perotti A, Panisson A, Bonchi F. Size-adaptive hypothesis 3317
testing for fairness. In: Advances in Neural Information Processing Systems 38 3318
(NeurIPS 2025); 2025. p. 108590–108626. 3319
3320
3321
3322
- [33] Debelak R. FairTree: Subgroup fairness auditing of machine learning models with 3323
bias–variance decomposition. In: Proceedings of the 2026 ACM Conference on 3324
Fairness, Accountability, and Transparency (FAccT); 2026. p. 4794–4810. 3325
3326
3327
3328
- [34] Cureton EE. Rank-biserial correlation. Psychometrika. 1956;21(3):287–290. 3329
<https://doi.org/10.1007/BF02289138>. 3330
3331
3332
- [35] Glass GV. Note on rank biserial correlation. Educational and Psychological Mea- 3333
surement. 1966;26(3):623–631. <https://doi.org/10.1177/001316446602600307>. 3334
3335
3336
- [36] Wendt HW. Dealing with a common problem in social science: A simplified rank- 3337
biserial coefficient of correlation based on the U statistic. European Journal of 3338
Social Psychology. 1972;2(4):463–465. <https://doi.org/10.1002/ejsp.2420020412>. 3339
3340
3341
3342
- [37] Willson VL. Critical values of the rank-biserial correlation coefficient. Educational 3343
and Psychological Measurement. 1976;36(2):297–300. <https://doi.org/10.1177/001316447603600207>. 3344
3345
3346
3347
3348
- [38] Bamber D. The area above the ordinal dominance graph and the area below 3349
the receiver operating characteristic graph. Journal of Mathematical Psychology. 3350
1975;12(4):387–415. [https://doi.org/10.1016/0022-2496\(75\)90001-2](https://doi.org/10.1016/0022-2496(75)90001-2). 3351
3352
3353
3354
- [39] Green DM. General prediction relating yes-no and forced-choice results. The 3355
Journal of the Acoustical Society of America. 1964;36(5):1042. <https://doi.org/> 3356
3357
3358

3359 [10.1121/1.2143307](https://doi.org/10.1121/1.2143307).
 3360
 3361 [40] Birnbaum ZW, Klose OM. Bounds for the variance of the Mann–Whitney
 3362 statistic. The Annals of Mathematical Statistics. 1957;28(4):933–945. [https:](https://doi.org/10.1214/aoms/1177706794)
 3363 [//doi.org/10.1214/aoms/1177706794](https://doi.org/10.1214/aoms/1177706794).
 3364
 3365
 3366
 3367 [41] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver
 3368 operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36. [https:](https://doi.org/10.1148/radiology.143.1.7063747)
 3369 [//doi.org/10.1148/radiology.143.1.7063747](https://doi.org/10.1148/radiology.143.1.7063747).
 3370
 3371
 3372
 3373 [42] Somers RH. A new asymmetric measure of association for ordinal variables.
 3374 American Sociological Review. 1962;27(6):799. <https://doi.org/10.2307/2090408>.
 3375
 3376
 3377 [43] Newson R. Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’
 3378 D and median differences. The Stata Journal. 2002;2:45–64. [https://doi.org/10.](https://doi.org/10.1177/1536867X0200200103)
 3379 [1177/1536867X0200200103](https://doi.org/10.1177/1536867X0200200103).
 3380
 3381
 3382
 3383 [44] van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative
 3384 ability of risk models in clustered data. BMC Medical Research Methodology.
 3385 2014;14(1):5. <https://doi.org/10.1186/1471-2288-14-5>.
 3386
 3387
 3388
 3389 [45] van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance
 3390 measure for risk prediction models in external validation settings. Statistics in
 3391 Medicine. 2016;35(23):4136–4152. <https://doi.org/10.1002/sim.6997>.
 3392
 3393
 3394
 3395 [46] Damen JAAG, Debray TPA, Pajouheshnia R, Reitsma JB, Scholten RJPM,
 3396 Moons KGM, et al. Empirical evidence of the impact of study characteristics on
 3397 the performance of prediction models: a meta-epidemiological study. BMJ Open.
 3398 2019;9(4):e026160. <https://doi.org/10.1136/bmjopen-2018-026160>.
 3399
 3400
 3401
 3402
 3403
 3404

- [47] Kartoun U, Khurshid S, Kwon BC, Patel AP, Batra P, Philippakis A, et al. Prediction performance and fairness heterogeneity in cardiovascular risk models. *Scientific Reports*. 2022;12(1):12542. <https://doi.org/10.1038/s41598-022-16615-3>.
- [48] Howard PD. The effect of sample heterogeneity and risk categorization on AUC predictive validity metrics. *Criminal Justice and Behavior*. 2017;44:103–120. <https://doi.org/10.1177/0093854816678899>.
- [49] Kundu S, Mazumdar M, Ferket B. Impact of correlation of predictors on discrimination of risk models in development and external populations. *BMC Medical Research Methodology*. 2017;17:63. <https://doi.org/10.1186/s12874-017-0345-1>.
- [50] van Amsterdam WAC. A causal viewpoint on prediction model performance under changes in case-mix: discrimination and calibration respond differently for prognosis and diagnosis predictions. *arXiv preprint arXiv:240901444*. 2024;<https://doi.org/10.48550/arXiv.2409.01444>.
- [51] Nieboer D, van der Ploeg T, Steyerberg EW. Assessing discriminative performance at external validation of clinical prediction models. *PLOS ONE*. 2016;11(2):e0148820. <https://doi.org/10.1371/journal.pone.0148820>.
- [52] Pfohl S, Harris N, Nagpal C, Madras D, Mhasawade V, Salaudeen O, et al. Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness. In: *Advances in Neural Information Processing Systems 38* (NeurIPS 2025); 2025. p. 46857–46918.
- [53] Petersen E, Holm S, Ganz M, Feragen A. The path toward equal performance in medical machine learning. *Patterns*. 2023;4:100790. <https://doi.org/10.1016/j.patter.2023.100790>.

- [54] Jones C, Castro DC, De Sousa Ribeiro F, Oktay O, McCradden M, Glocker B. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*. 2024;6:138–146. <https://doi.org/10.1038/s42256-024-00797-8>.
- [55] Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*. 2017;5(2):153–163. <https://doi.org/10.1089/big.2016.0047>.
- [56] Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. In: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). vol. 67 of Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum für Informatik; 2017. p. 43:1–43:23.
- [57] Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On fairness and calibration. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; 2017. p. 5680–5689. Available from: <https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffb2d39ab038d1cd7-Abstract.html>.
- [58] Van Calster B, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *The Lancet Digital Health*. 2025;7(12):100916. <https://doi.org/10.1016/j.landig.2025.100916>.
- [59] Pfohl S, Xu Y, Foryciarz A, Ignatiadis N, Genkins J, Shah N. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*; 2022. p. 1039–1052.

- [60] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845. <https://doi.org/10.2307/2531595>.
- [61] Ramdas A, Barber RF, Candès EJ, Tibshirani RJ. Permutation tests using arbitrary permutation distributions. *Sankhya A*. 2023;85:1156–1177. <https://doi.org/10.1007/s13171-023-00308-8>.
- [62] Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for p -Value Adjustment*. New York: Wiley; 1993.
- [63] Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10(1):101–129. <https://doi.org/10.2307/3001666>.
- [64] Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Medical Research Methodology*. 2012;12:82. <https://doi.org/10.1186/1471-2288-12-82>.
- [65] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979;.
- [66] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*. 2016;74:167–176. <https://doi.org/10.1016/j.jclinepi.2015.12.005>.
- [67] Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*. 1989;64(5):304–310. [https://doi.org/10.1016/0002-9149\(89\)90524-9](https://doi.org/10.1016/0002-9149(89)90524-9).

3543 [68] Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, et al. Impact
3544 of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical
3545 database patient records. *BioMed Research International*. 2014;2014:781670.
3546 <https://doi.org/10.1155/2014/781670>.
3547
3548
3549
3550 [69] National Center for Health Statistics. National Health Interview Survey, 2024
3551 Public-Use Data File (Sample Adult). U.S. Department of Health and Human
3552 Services, Centers for Disease Control and Prevention; 2025. Available from: [https:](https://www.cdc.gov/nchs/nhis/documentation/2024-nhis.html)
3553 [//www.cdc.gov/nchs/nhis/documentation/2024-nhis.html](https://www.cdc.gov/nchs/nhis/documentation/2024-nhis.html).
3554
3555
3556
3557
3558 [70] National Center for Health Statistics. National Health Interview Survey, 2023
3559 Public-Use Data File (Sample Adult). U.S. Department of Health and Human
3560 Services, Centers for Disease Control and Prevention; 2024. Available from: [https:](https://www.cdc.gov/nchs/nhis/documentation/2023-nhis.html)
3561 [//www.cdc.gov/nchs/nhis/documentation/2023-nhis.html](https://www.cdc.gov/nchs/nhis/documentation/2023-nhis.html).
3562
3563
3564
3565 [71] National Center for Health Statistics. National Health and Nutrition Examination
3566 Survey, 2021–2023 (Cycle L) Public-Use Microdata. U.S. Department of Health
3567 and Human Services, Centers for Disease Control and Prevention; 2024. Avail-
3568 able from: [https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?](https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023)
3569 [Cycle=2021-2023](https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023).
3570
3571
3572
3573
3574 [72] Centers for Disease Control and Prevention. Behavioral Risk Factor Surveil-
3575 lance System: 2024 Annual Survey Data. U.S. Department of Health and
3576 Human Services; 2025. Available from: [https://www.cdc.gov/brfss/annual_data/](https://www.cdc.gov/brfss/annual_data/annual_2024.html)
3577 [annual_2024.html](https://www.cdc.gov/brfss/annual_data/annual_2024.html).
3578
3579
3580
3581
3582 [73] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea:
3583 An approach, method, and software mechanism for generating synthetic patients
3584 and the synthetic electronic health care record. *Journal of the American Medi-*
3585 *cal Informatics Association*. 2018;25(3):230–238. <https://doi.org/10.1093/jamia/>
3586
3587
3588

ocx079.	3589
	3590
[74] Bellibatlu RR.: RISED healthcare evaluation dataset. Hugging Face. Hugging	3591
Face Datasets. Available from: https://huggingface.co/datasets/Rohithreddybc/	3593
rised-healthcare-eval-dataset .	3594
	3595
	3596
[75] Kohavi R. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree	3597
hybrid. In: Proceedings of the Second International Conference on Knowledge	3598
Discovery and Data Mining (KDD). AAAI Press; 1996. p. 202–207.	3599
	3600
	3601
	3602
[76] Ding F, Hardt M, Miller J, Schmidt L. Retiring Adult: New datasets for fair	3603
machine learning. In: Advances in Neural Information Processing Systems 34	3604
(NeurIPS 2021); 2021. p. 6478–6490. Available from: https://proceedings.neurips.	3605
cc/paper/2021/hash/32e54441e6382a7fbacbbbf3c450059-Abstract.html .	3606
	3607
	3608
	3609
[77] Hofmann H.: Statlog (German Credit Data) dataset. UCI Machine Learning	3610
Repository. UCI Machine Learning Repository. Available from: https://archive.	3611
ics.uci.edu/dataset/144/statlog+german+credit+data .	3612
	3613
	3614
	3615
[78] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings	3616
of the 22nd ACM SIGKDD International Conference on Knowledge Discovery	3617
and Data Mining; 2016. p. 785–794.	3618
	3619
	3620
	3621
[79] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in	3622
Python. Journal of Machine Learning Research. 2011;12:2825–2830.	3623
	3624
	3625
[80] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-	3626
learning-based science. Patterns. 2023;4(9):100804. https://doi.org/10.1016/j.	3627
patter.2023.100804 .	3628
	3629
	3630
	3631
	3632
	3633
	3634

3635 [81] Mittelstadt B, Wachter S, Russell C. The unfairness of fair machine learning:
3636 Levelling down and strict egalitarianism by default. Michigan Technology Law
3637 Review. 2024;30(1):3. <https://doi.org/10.36645/mtlr.30.1.unfairness>.
3638
3639
3640
3641 [82] Benitez-Aurioles J, Joules A, Brusini I, Peek N, Sperrin M. Understanding
3642 algorithmic fairness for clinical prediction in terms of subgroup net benefit and
3643 health equity. Epidemiology. 2026;37(3):386–396. [https://doi.org/10.1097/EDE.](https://doi.org/10.1097/EDE.0000000000001949)
3644 [0000000000001949](https://doi.org/10.1097/EDE.0000000000001949).
3645
3646
3647
3648 [83] Meerwijk EL, McElfresh DC, Martins S, Tamang SR. Evaluating accuracy and
3649 fairness of clinical decision support algorithms when health care resources are
3650 limited. Journal of Biomedical Informatics. 2024;156:104664. [https://doi.org/10.](https://doi.org/10.1016/j.jbi.2024.104664)
3651 [1016/j.jbi.2024.104664](https://doi.org/10.1016/j.jbi.2024.104664).
3652
3653
3654
3655 [84] Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to inter-
3656 preting decision curve analysis. Diagnostic and Prognostic Research. 2019;3(1):18.
3657 <https://doi.org/10.1186/s41512-019-0064-7>.
3658
3659
3660
3661 [85] Piovani D, Sokou R, Tsantes AG, Vitello AS, Bonovas S. Optimizing clinical
3662 decision making with decision curve analysis: insights for clinical investigators.
3663 Healthcare. 2023;11(16):2244. <https://doi.org/10.3390/healthcare11162244>.
3664
3665
3666
3667
3668
3669
3670
3671
3672
3673
3674
3675
3676
3677
3678
3679
3680

Supplementary Files

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

- [additionalfile1.pdf](#)