

Additional file 1: Supplementary information

This material carries the least weight of anything in the study and the most caveats, which is why it is here rather than in the main text. It reports the four non-clinical contrast cohorts and the cohort-provenance table. Nothing in the paper’s argument depends on any of it.

S1. The non-clinical contrast cohorts

Four cohorts are not clinical: a synthetic Synthea-inspired cohort [1, 2], UCI Adult Income [3], Folktables ACS-Income (California 2018) [4] and Statlog German Credit [5]. They were run through the identical pipeline and are reported only as a contrast, for two reasons: they are the cohorts on which the fairness toolkits studied here were originally demonstrated, and their inclusion shows that the admissibility sensitivity is not peculiar to health data. They are not evidence about clinical prediction, and no clinical conclusion in this paper rests on them.

Counting all nine cohorts with any verdict rather than the five clinical ones, the number whose verdict is not constant across the five admissibility rules is 3 of 9 for the studentized test, the raw max–min range and the double-corrected z -test; 2 of 9 for Cochran’s Q ; 1 of 9 for the fixed cut-point and the four-fifths rule; and 0 of 9 for the bootstrap-interval reading, which flags nothing anywhere. Under the strict variant, which counts a transition into or out of non-evaluability as a change, every all-cohort count rises by exactly one, because German Credit ($n = 200$ test rows) becomes unevaluable at `m100` for every procedure. The clinical counts reported in Section 3.7 of the main text are unchanged under either variant.

Twelve further demographic partitions of these four cohorts have a measured case-mix magnitude $\hat{\rho}$, computed and stored by the same code as the 21 clinical partitions of Table 1 of the main text, but they are excluded from that table by design. Two of them are the only partitions anywhere in the study whose case-mix attributable

fraction is estimated tightly (Adult Income `age_group`, 0.34 [0.26, 0.45]; Adult Income `sex`, 0.18 [0.06, 0.29]), which is a consequence of their sample sizes and not a reason to prefer them as evidence about clinical models.

S2. Software versions

Results in this manuscript were produced under the following versions, five of which are recorded and two of which are not. The recorded five are in `results/comparator_runtime_env.json` and `results/model_provenance.json` in the archived repository, and are given here so that a reader can reproduce the environment without opening those files.

Table 1 Supplementary Table S1. Software version manifest.

Component	Version	Used for
Python	3.11.7	all analyses
NumPy	1.26.4	numerics throughout
SciPy	1.15.3	quadrature, distributions, optimisation
pandas	3.0.2	result assembly
scikit-learn	1.8.0	logistic-regression pipeline, splits, AUROC
XGBoost	3.2.0	the seven gradient-boosted cohort models
fairlearn	0.13.0	four-fifths ratio (<code>MetricFrame.ratio</code>)

Platform: Windows, x86-64, 8 physical cores; NumPy built against OpenBLAS 0.3.23. BLAS thread-count environment variables were left unset.

`PYTHONHASHSEED` is pinned for the whole process tree, and every geometry’s seed word is a CRC-32 digest of its name rather than Python’s salted string hash, so simulation cells are reproducible bit for bit (Section 2.3 of the main text).

Python, NumPy, SciPy and pandas versions are taken from the stored runtime environment record; the XGBoost version is taken from the stored model-provenance record; the scikit-learn and fairlearn versions are those of the analysis environment, are not independently recorded in any result file, and should be treated as unverified.

Table 2 Supplementary Table S2. Cohorts, model provenance and partition geometry at the reference admissibility rule m30. Testbeds, not populations: survey cohorts are complete-case extracts analysed without survey weights, strata or primary sampling units, and no population inference is drawn from any of them. GBT = gradient-boosted trees; LR = logistic regression.

Cohort	Clin.	<i>n</i> test	Prev.	Model	Feat.	Part.	Levels
UCI Heart Disease	yes	61	0.541	GBT	13	0	1
UCI Diabetes 130	yes	19,890	0.116	GBT	14	3	16
NHIS 2024	yes	1,950	0.075	GBT	19	5	19
NHIS 2023	yes	5,423	0.112	GBT	14	4	15
NHANES 2021–2023	yes	820	0.130	GBT	14	4	13
BRFSS 2024	yes	8,978	0.211	GBT	19	5	24
Synthea-inspired	no	2,000	0.300	GBT	20	4	14
UCI Adult Income	no	9,045	0.248	LR	10	3	9
Folktables ACS-Income	no	4,000	0.410	LR	10	3	6
Statlog German Credit	no	200	0.700	LR	11	2	4

UCI Heart Disease retains only one admissible level under m30, so no partition contributes two levels and Eq. 1 of the main text has no value. Its row is retained because non-evaluability is a reportable outcome, not a missing number.

Correcting an error in our earlier analysis: the cohort models are *not* all gradient-boosted trees. Seven of ten are; Adult Income, ACS-Income and German Credit are logistic regression.

Source: `model_provenance.json`.

S3. Cohorts and model provenance

List of abbreviations

AUROC: area under the receiver operating characteristic curve; BRFSS: Behavioral Risk Factor Surveillance System; EHR: electronic health record; GBT: gradient-boosted trees; LR: logistic regression; NHANES: National Health and Nutrition Examination Survey; NHIS: National Health Interview Survey.

References

- [1] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients

139 and the synthetic electronic health care record. Journal of the American Medi-
140 cal Informatics Association. 2018;25(3):230–238. [https://doi.org/10.1093/jamia/](https://doi.org/10.1093/jamia/ocx079)
141 [ocx079](https://doi.org/10.1093/jamia/ocx079).
142
143
144
145 [2] Bellibatlu RR.: RISED healthcare evaluation dataset. Hugging Face
146 Datasets. Available from: [https://huggingface.co/datasets/Rohithreddybc/](https://huggingface.co/datasets/Rohithreddybc/rised-healthcare-eval-dataset)
147 [rised-healthcare-eval-dataset](https://huggingface.co/datasets/Rohithreddybc/rised-healthcare-eval-dataset).
148
149
150
151 [3] Kohavi R. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree
152 hybrid. In: Proceedings of the Second International Conference on Knowledge
153 Discovery and Data Mining (KDD). AAAI Press; 1996. p. 202–207.
154
155
156 [4] Ding F, Hardt M, Miller J, Schmidt L. Retiring Adult: New datasets for fair
157 machine learning. In: Advances in Neural Information Processing Systems 34
158 (NeurIPS 2021); 2021. p. 6478–6490. Available from: [https://proceedings.neurips.](https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbf3c450059-Abstract.html)
159 [cc/paper/2021/hash/32e54441e6382a7fbacbbbf3c450059-Abstract.html](https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbf3c450059-Abstract.html).
160
161
162
163
164 [5] Hofmann H.: Statlog (German Credit Data) dataset. UCI Machine Learn-
165 ing Repository. Available from: [https://archive.ics.uci.edu/dataset/144/statlog+](https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data)
166 [german+credit+data](https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data).
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184