

You Don't Need the Black Box: Named ECG Features Rival Foundation Models for Mortality Prediction

Axel Faes

Abstract—Foundation models (FMs) for the 12-lead electrocardiogram (ECG) are reported to extract prognostic signal that named measurements cannot. Under matched modeling, we test whether any real ECG-FM beats a strong, transparent gradient-boosted model over named, clinically grounded features for one-year all-cause mortality from a single ECG on external (CODE-15) and internal (MIMIC-IV-ECG) cohorts, tagging each comparison for pretraining-data overlap. The named model reached an area under the receiver operating characteristic curve (AUROC) of 0.811 externally and 0.773 internally over 231 and 235 named features, with near-ideal calibration (slope 1.02 and 1.01) and precision roughly 2.5 to 3 times prevalence. Against the best frozen-embedding probe of six published ECG-FMs, none out-discriminated the named model: it won 7 of 12 comparisons and tied 5: ECG-JEPA and ST-MEM on both cohorts, including one they never saw, and ECGFounder internally. The pattern tracks encoder strength rather than pretraining overlap. The most transferable result is cross-cohort: trained on one cohort and tested on the other, the named model beats a leakage-clean ECGFounder in both directions, and ECG-JEPA and ST-MEM edge it in one direction but it beats them in the other, so no FM beats it in both; ECG-FM, competitive within a cohort, collapses under transfer. A leakage-clean FM adds a small complementary increment externally (up to about 0.02 AUROC) and near zero internally, though the stronger encoders carry some. Each named prediction carries a per-patient provable sufficient-reason certificate of median length one feature. A single command reproduces every reported number from raw data.

Index Terms—Data leakage, electrocardiography, formal explanations, foundation models, gradient boosting, interpretable machine learning, mortality prediction.

I. INTRODUCTION

DEEP neural networks read prognostic information from the 12-lead ECG that trained cardiologists do not annotate. Networks trained end to end match specialist performance on rhythm classification [16], detect atrial fibrillation from a sinus-rhythm tracing [17], flag contractile dysfunction from a surface recording [18], and forecast new-onset atrial fibrillation before it appears [21]. The most direct statement of the premise this study re-examines is the prediction of one-year all-cause mortality from raw voltage. Raghunath *et al.* [20] reported strong discrimination even on ECGs that clinicians read as normal, attributing the signal to features not apparent on visual inspection. That result, together with the broader claim that the ECG encodes sub-visual physiology [15], [17], framed a decade of work around signal a human cannot name. The arrival of ECG foundation models [1], [14]

extended the premise. Large self-supervised encoders promise to capture this hidden structure once and transfer it to many downstream tasks, and the transfer story has made frozen-embedding probing the default way to report an ECG-FM's value.

The premise has visible cracks that motivate a direct test. Compact, task-specific models rival foundation models across most clinical ECG tasks in recent benchmarks [8], and holistic evaluations that look past the headline metric find the reported gains fragile [9]. Part of the apparent black-box advantage is nameable. Deep networks estimate chronological age from the ECG with high accuracy [19], and a network-estimated ECG age is itself a mortality predictor [22], which places a well-understood covariate, age, inside the “hidden” signal. Interpretable models can match black boxes on high-stakes tasks when the features are chosen well [39], which reframes the choice as an empirical question rather than an accuracy-transparency trade-off. That trade-off has nonetheless been asserted: the claim that explainability necessarily costs accuracy [43] has not been tested against a nonlinear named model competing with real ECG-FMs. A separate and under-examined confound is leakage. When an evaluation cohort overlaps a model's pretraining data, benchmark performance inflates in ways that do not transfer [53]. For ECG-FMs the risk is concrete, because some published encoders were pretrained on cohorts that are also standard evaluation sets, yet a controlled leaked-versus-clean comparison across real encoders has not been reported.

The question matters for deployment, not only for benchmarking. One-year mortality from a single ECG is a screening-grade signal that could route follow-up, and a model that a clinician must act on should be one whose reasons can be examined and, ideally, proven sufficient for the case at hand. If a named model of trusted measurements matches an opaque encoder, the interpretable model is the safer default under distribution shift and the auditable one at the bedside. If instead an encoder carries genuine signal beyond named features, that is worth knowing precisely, so the field can invest in the encoders that add value rather than the ones that inflate on a familiar cohort. Either answer is useful, which is why the comparison deserves the controls that make its result trustworthy.

This study asks one controlled question and answers it with a positive result. For one-year all-cause mortality from a single 12-lead ECG, we compare a transparent gradient-boosted model over named, clinically grounded features against the

best frozen-embedding probe of six real ECG-FMs, on an external and an internal cohort, under matched modeling and with a pretraining-overlap tag on every comparison. The contribution is the conjunction of a strong named baseline that FM benchmarks omit [8], a head-to-head comparison across six real encoders with pretraining overlap tagged as a fairness control [7], [53], a cross-cohort transfer arm in which each model is trained on one cohort and tested on the other, and a per-patient provable sufficient-reason certificate over the named features. The closest neighbor, Lukyanenko *et al.* [24], evaluated gradient-boosted and demographic baselines on the same two cohorts, but ran no FM comparison, no leakage tagging, no transfer arm, and no per-patient certificate. Those are precisely our delta, and joining them into one study turns a familiar “simple baseline is competitive” observation into a controlled account of how a named model stands against real encoders.

The scope of the claim is bounded and stated once here. This is a positive competitiveness study under equal modeling. Equal AUROC is not equal information, so the correct claim is that no FM out-discriminates the named model head-to-head, while a leakage-clean FM contributes a small, disclosed complementary increment of about 0.01 to 0.02 AUROC. Age is the single strongest individual feature, but the ECG waveform features collectively exceed age alone, and the ECG-specific component beyond age is substantial rather than marginal. Within those bounds the finding stands, and it strengthens under transfer: trained on one cohort and tested on the other, the named model beats a leakage-clean encoder in both directions. Section II reviews FM benchmarks, interpretable neighbors, formal certificates, and the leakage literature. Section III describes the cohorts, the six encoders and their probes, the named feature bank, the transparent model, the certificate, and the statistics. Section IV reports the main comparison, the cross-cohort transfer arm, calibration, three scoping contrasts, and the complementarity analysis. Section V discusses the result and states its limitations. Section VI concludes.

II. RELATED WORK

A. ECG foundation models and their benchmarks

Self-supervised ECG encoders now span several pretraining paradigms and are the object of study here. ECGFounder was trained on over ten million recordings with external multi-domain evaluation [1]. ECG-FM is an open encoder in the wav2vec family [2], [10]. HuBERT-ECG adapts masked-unit prediction from speech [3], [11]. ECG-JEPA applies a joint-embedding predictive architecture [4], [12]. A multimodal variant, MERL [6], and further masked-autoencoder encoders complete the design space; the masked-autoencoder ST-MEM [5], [13] is included in our evaluated roster below, whereas MERL is multimodal and requires paired ECG-text report inputs, so it is out of scope for this signal-only frozen-embedding comparison. Evaluation of these encoders has converged on linear or shallow probing of frozen embeddings, a protocol formalized by benchmark suites such as BenchECG and its xECG baseline [7]. The convention isolates representation quality from a fine-tuning budget, which makes it a fair way to compare encoders against each other.

The critical gap is comparative rather than architectural. These benchmarks pit encoders against other encoders and against linear probes, but not against a strong named-feature model built from measurements a clinician already trusts. Al-Masud *et al.* [8] found that compact models beat FMs on most clinical tasks and named the absence of a fair simple baseline as an open problem. Filice *et al.* [9] reached a similar conclusion by looking beyond accuracy to calibration and robustness, where the FM advantage narrows further. Neither benchmark includes a named-feature reference on a mortality endpoint, so the question of whether an FM adds value over named measurements has been left open by exactly the studies best placed to answer it. Our design closes that gap by making a fully named, boosted model the reference point against which every encoder is measured.

B. Interpretable and named-feature neighbors

A parallel line predicts ECG outcomes from named measurements or from a small set of interpretable factors. Van de Leur *et al.* [23] compressed the ECG into 21 interpretable factors via a variational autoencoder and predicted one-year mortality from them, showing that a low-dimensional, inspectable representation retains prognostic value. Kim *et al.* [25] combined named features with demographics for cardiovascular mortality, and a systematic review of externally validated one-to-three-year mortality models found that age and demographics dominate and that simple models rival complex ones [26]. These studies establish that named inputs are competitive, but none compares against a foundation model or certifies its predictions.

The closest neighbor is Lukyanenko *et al.* [24], who benchmarked deep survival models against gradient-boosted and demographic baselines on the same MIMIC and CODE-15 cohorts we use. That work established that a boosted baseline is competitive with deep survival networks, which is the starting point rather than the endpoint of the present study. It did not compare against foundation models, did not control for pretraining leakage, and issued no per-patient certificate. We therefore build an independent named pipeline, from delineation through feature computation to the boosted model, and add exactly those three controls. The comparison is the contribution, and the independence of the pipeline matters because it lets us claim the result without generalizing anyone else’s specific design.

C. Post-hoc explanation and formal certificates

Interpretability for clinical models divides into post-hoc attribution and formal guarantees. Post-hoc methods such as SHAP [44] and LIME [45] assign importance to features, but their reliability is contested. Saliency maps can fail basic sanity checks [47], and post-hoc explanations in health care have been criticized as a false hope because they do not certify what the model actually used [40]. These methods estimate importance; they do not prove sufficiency, which is a poor fit for a decision that a clinician must be able to defend patient by patient.

A sharper object is a sufficient reason, a subset of feature values that provably fixes the model’s decision regardless of the others [51], [52]. Abduction-based explanation frameworks compute such reasons for machine-learning models [48], and a probabilistic, delta-sufficient variant relaxes the guarantee to a stated coverage level [49], which is the form we adopt. Formal XAI argues that this kind of guarantee, rather than a heuristic attribution, is what trustworthy clinical deployment requires [41], [42], [50]. These methods are established, and our use of them is an application to a clinical mortality model. We credit the informal precursor of rule-based sufficient conditions, Anchors [46], as the heuristic version of the guarantee we compute exactly, so the novelty we claim is the deployment context.

D. Leakage, contamination, and shortcut learning

Leakage inflates machine-learning benchmarks across fields, and training on data that overlaps the evaluation cohort is one of its most common forms [53]. The problem is quantified in medical machine learning. Connectome-based models inflate when subjects leak across the split [56], and leaked studies bias pooled meta-analytic estimates upward [57]. The membership-inference literature developed to detect evaluation data inside a model’s pretraining corpus [54], [55] is the direct analog of the concern for ECG-FMs, whose pretraining cohorts sometimes coincide with standard evaluation sets. These tools establish both the mechanism and its magnitude, which is why we treat leakage as a first-class variable rather than a footnote.

A related failure is shortcut learning, where models exploit spurious correlates that ace a benchmark but fail to transfer [58]. The pattern recurs in clinical imaging [59], [60] and across thirteen datasets that include ECG, where shortcut exploitation inflated performance by roughly twenty percent [61]. Benchmark protocols acknowledge the risk, for instance by excluding certain cohorts to preserve fairness [7], but the leaked-versus-clean effect has not been operationalized as a controlled dissociation across real ECG-FMs on a clinical outcome. We tag every comparison for pretraining overlap and treat leakage as a fairness control rather than a claimed mechanism, so that even an encoder evaluated on its own pretraining cohort is given that benefit and still measured against the named model. Crediting this literature is deliberate: we do not claim to discover leakage, only to hold it to account in an ECG-FM mortality comparison that the field has not run.

III. METHODS

A. Cohorts and outcome

We used two independent cohorts, each contributing one 12-lead ECG per patient. The internal cohort is MIMIC-IV-ECG [68], from which we drew $N=14,000$ recordings with a one-year all-cause mortality prevalence of 0.129 (approximately 1,806 deaths). The external cohort is CODE-15 [69], the public 15 percent release of the CODE study [15], from which we drew $N=20,000$ recordings with a prevalence of 0.141 (approximately 2,820 deaths). The outcome is death from any cause within one year of the ECG, matching the endpoint of the landmark deep-learning claim we re-examine [20]. For

CODE-15, patients who died within one year were labeled events, those known alive beyond one year were non-events, and patients censored before one year without a recorded death were excluded; for MIMIC-IV, mortality was defined as a linked date of death within 365 days of the ECG, and patients without a recorded death in that window were treated as survivors. Using two cohorts from different health systems lets us report an internal and an external number for every model and lets distribution shift, rather than a single site’s idiosyncrasies, be part of the comparison.

Within each cohort we drew a single ECG per patient to remove within-patient correlation, split patients into disjoint training, validation, and held-out sets so that no patient appeared in more than one split, and fit every model on the training patients with hyperparameters selected on the validation patients. All reported AUROC values and all certificate statistics are computed on the held-out patients. The same split protocol was applied to the named model and to every FM probe, so that a reported difference reflects the model and not an accident of which patients landed in which fold. Prevalence was left at its natural cohort value rather than rebalanced, because the endpoint is a screening-grade risk and calibration to the true base rate is part of what a deployed model must get right.

CODE-15 has a second role in this study beyond serving as the external cohort: it is the site of most pretraining overlap. CODE-15 is a publicly released ECG corpus that appears in the pretraining data of four of the six encoders we evaluate (HuBERT-ECG, xECG, ECG-JEPA, and ST-MEM), which makes it where a clean-versus-leaked contrast can be drawn. We label every encoder, per cohort, according to whether that cohort overlapped its pretraining corpus, so that each head-to-head number can be read alongside its overlap status. This tag is a fairness control: it gives an overlapping encoder the benefit of its own pretraining cohort and asks whether the named model still matches or beats it, rather than asserting that overlap explains any FM result.

B. Foundation-model encoders and probes

We evaluated six published ECG-FMs as frozen feature extractors: ECGFounder [1], ECG-FM [2], HuBERT-ECG [3], ECG-JEPA [4], xECG [7], and ST-MEM [5]. The roster spans the wav2vec, masked-unit, joint-embedding, masked-autoencoder, and benchmark-baseline paradigms and their self-supervised origins in speech and vision [10]–[13], so the comparison is not tied to one pretraining recipe. For each encoder we extracted the frozen embedding of every ECG in each cohort using the published weights and the authors’ preprocessing, and we did not fine-tune any encoder.

The pretraining-overlap tag on each encoder is taken from its published pretraining description. ECGFounder was pre-trained on the Harvard-Emory ECG database and is clean on both cohorts [1]. ECG-FM was pretrained on MIMIC-IV-ECG and PhysioNet-2021, so it is leaked on MIMIC and clean on CODE-15 [2]. HuBERT-ECG was pretrained on a multi-corpus set that includes MIMIC-IV-ECG and CODE/TNMG, so it is leaked on both cohorts [3]. ECG-JEPA was pretrained on

Shaoxing and CODE-15, so it is leaked on CODE-15 and clean on MIMIC [4]. ST-MEM was pretrained on Chapman, Ningbo, and CODE-15, so it is leaked on CODE-15 and clean on MIMIC [5]. xECG is the BenchECG baseline encoder rather than a large foundation model, and with CODE-15 in its pretraining it is leaked on CODE-15 and clean on MIMIC [7].

On top of each frozen embedding we trained two probes for the mortality label: a linear probe and a gradient-boosted probe. Reporting only a linear probe would understate an encoder whose useful structure is nonlinear, and reporting only a boosted probe would understate one whose embedding is already linearly separable, so we report each FM’s best probe, the higher-scoring of the two on held-out data. This best-probe rule follows established probing practice for ECG-FMs [7] and ensures that every encoder is represented by its strongest frozen-embedding result. It also removes an obvious objection in advance: if the named model wins, it does not win because the FM was handicapped by a poor probe choice, because the FM was given its best of two heads.

Both probe families used the same training and selection discipline as the named model. The gradient-boosted probe on an FM embedding was configured to match the head applied to the named features, so that the equal-modeling comparison in Section IV-D varies only the input representation. Probe hyperparameters were selected on the validation split by the same search for the FM embeddings and the named features, and neither the encoder weights nor the named-feature definitions were tuned to the mortality label, since the encoders were frozen and the feature bank was fixed in advance from the prognostic literature. The comparison is therefore symmetric: the FM cannot benefit from label-driven representation tuning that the named model is denied, and the named model cannot benefit from a probe search the FM is denied.

C. Named feature bank

The named model draws on a superset of 891 named, clinically interpretable features computed from a standard delineation of each ECG. We used NeuroKit to detect fiducial points and derive intervals, amplitudes, and rhythm statistics, then selected a compact bank of 231 features (235 on the internal cohort, reflecting cohort-specific availability of a few measurements). The compact bank was fixed before any mortality fitting by two criteria applied to the 891-feature superset: each retained feature belongs to a family with an established prognostic literature (the heart-rate-variability, interval, and rate measures cited below), and the compact-versus-full robustness check in Section IV-I confirmed that this literature-grounded subset holds AUROC within 0.006 of the full superset, so the reduction sacrifices no discrimination. Every feature is a quantity a clinician can name, and each family is grounded in prior mortality evidence. Heart-rate-variability indices are grounded by the Framingham and Zutphen cohorts, which linked reduced variability to all-cause mortality [27], [28], and are computed following the measurement standards of the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology [70].

Because those standards were set on longer recordings and our ECGs are approximately ten seconds, we restrict the heart-rate-variability features to time-domain measures defensible at that length. Resting heart rate and the RR interval are grounded by a meta-analysis in which a 10-bpm increase carried a relative risk of 1.09 [29]. The QT interval [30], QRS duration [31], and PR interval [32] each carry an independent prognostic literature, and patient age [19] is included as the named form of the covariate that age-estimation networks recover from the waveform.

Both the full 891-feature superset and the compact 231-feature bank were carried through the entire pipeline, so that compactness could be measured against the superset rather than assumed. Keeping the compact bank small is deliberate. A model a clinician can enumerate in full is a precondition for the per-patient certificate to be meaningful, because a certificate that names one feature out of two hundred is auditable in a way that a certificate naming one component out of a learned embedding is not.

D. Transparent model and a deliberately weaker linear control

The named model of record is a gradient-boosted tree ensemble [33]–[35] trained on the named features with standard scikit-learn tooling [36]. Gradient boosting is the appropriate transparent model here for a specific reason: tree ensembles remain the strongest general learners on tabular data [37], [38], so a boosted head gives the named features their best honest chance and matches the nonlinear probes applied to the FM embeddings. Choosing a weaker linear model of record would have stacked the comparison in the FMs’ favor and would have confounded model class with feature set.

To separate the effect of model class from the effect of features, we also trained a deliberately linear named-plus-age model as a labelled control alongside the boosted model of record. The linear control has one job: to reveal where an FM’s apparent edge comes from. If a clean FM beats the linear named model but loses to the boosted named model, the edge is attributable to probe nonlinearity rather than to un-nameable physiology. The linear model is labelled and reported as a control; the strong boosted named model is the reference in the primary table.

E. Per-patient sufficient-reason certificate

For each named prediction we computed a per-patient sufficient-reason certificate, applying the framework of abduction-based and probabilistic explanations [48], [49], [51], [52]. A certificate is a minimal subset of the patient’s named feature values that fixes the model’s above-threshold decision at a stated coverage level. We used a delta-sufficient relaxation with $\delta = 0.9$, so the guarantee holds for at least 90 percent of completions of the unfixed features [49], which trades an exact but often large certificate for a short one that still carries a quantified guarantee. The output is a provable statement of the form “these named values are enough to explain this prediction.”

Concretely, for a given patient the procedure searches for a small set of that patient’s named values whose fixing keeps the model above threshold across a sampled set of completions of the remaining features, and returns a minimal such set at the target coverage. Completions are drawn from the empirical joint distribution by row-bootstrapping the training set, not from independent per-feature marginals, so the sufficiency guarantee holds against clinically plausible patients rather than physiologically impossible feature combinations. We report the length of that set, its coverage measured empirically on held-out completions, and which features appear in it, aggregated across the held-out cohort. Because the search operates on the compact named model directly, the certificate refers to quantities a clinician already reads, and its length is a direct measure of how concentrated the model’s reasoning is for that case.

This is stronger than a post-hoc importance score [44], [46] because it certifies sufficiency rather than estimating attribution, and it is computed on the same compact model that produced the prediction, so there is no surrogate to mistrust. The certificate machinery is established in prior work on formal explanations, and our contribution is its application to a clinical mortality model at the exact point of FM comparison. We apply the framework of [48], [49], [51], [52], and do not propose a new certificate method.

F. Statistical comparison

All head-to-head comparisons used a paired patient-level bootstrap with 95 percent confidence intervals on the AUROC difference. The pairing matters because the named model and each FM probe are evaluated on the same patients, and a paired resample removes the between-patient variance that would otherwise widen the interval. We chose a bootstrap on the difference, rather than a comparison of two marginal intervals, because the AUROC itself is unstable across populations [62] and the quantity of interest is the within-cohort difference. A positive delta favors the named model.

Two rules kept the comparison fair. Every FM comparison carries its per-cohort clean or leaked tag [53], so the clean-versus-leaked contrast is legible in the primary table rather than argued only in prose. The equal-modeling rule is applied throughout: when we compare model classes, the FM embedding receives the same boosted head as the named features, so a reported win reflects the features and not an asymmetry in the probe. Together these rules make each cell of the results table interpretable on its own, without a reader having to reconstruct which model class or which cohort exposure produced it.

Two further pre-specified rules govern inference across the family of comparisons. We declared a two-sided equivalence margin of $M = 0.01$ AUROC and tested equivalence by two one-sided tests (TOST) using the 90 percent confidence interval on the difference. Because the head-to-head family contains $K = 12$ comparisons, we applied a Bonferroni correction to a per-test alpha of 0.0042 and report which wins survive family-wise correction. Sub-0.02 AUROC differences are statistical rather than clinical effects, a distinction we keep explicit throughout [71].

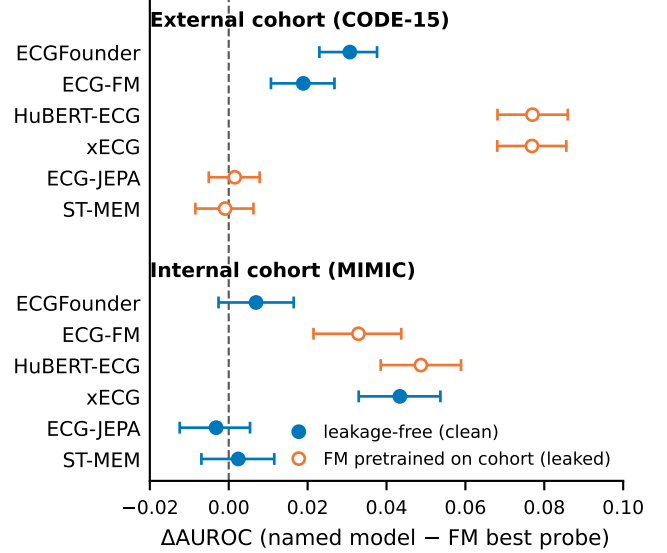


Fig. 1. Head-to-head AUROC deltas between the transparent gradient-boosted named-feature model and each foundation model’s best probe, by cohort. Positive Δ AUROC favors the named model; markers show the point estimate and horizontal bars the 95% confidence interval. Filled markers denote leakage-free (clean) comparisons; open markers denote comparisons in which the foundation model was pretrained on the evaluation cohort (leaked). The twelve points cover six encoders on two cohorts; the named model wins 7, ties 5 (ECG-JEPA and ST-MEM in both cohorts, ECGFounder on MIMIC), and loses none.

G. Reproducibility

A single command, `run_all.sh`, rebuilds every number, table, and figure in this paper from the raw ECG archives in dependency order. The pipeline regenerates the cohorts and splits, the delineated feature caches, the frozen FM embeddings, the trained probes, the certificate tables, and the final figures, with no step assuming a pre-existing intermediate. All paths resolve from the repository root and environment variables, with no hardcoded user or machine paths, so another user on another account obtains identical behavior.

The external prerequisites that cannot be redistributed are checked explicitly at the top of the run. The credentialed cohorts [68], [69] and the published encoder weights must be obtained by the user, and the script fails loudly with the exact acquisition command if any is absent, rather than failing midway with a missing-cache error. Obtaining those inputs is the only manual step, and the reproduction otherwise runs unattended from raw data to reported artifact.

IV. RESULTS

A. Main comparison: no foundation model out-discriminates the named model

The headline result is a positive for the named baseline: no FM out-discriminated the best named model on either cohort. The transparent gradient-boosted named model reached AUROC 0.811 on the external CODE-15 cohort (231 features) and 0.773 on the internal MIMIC cohort (235 features). Compared against the best linear-or-boosted probe of each of

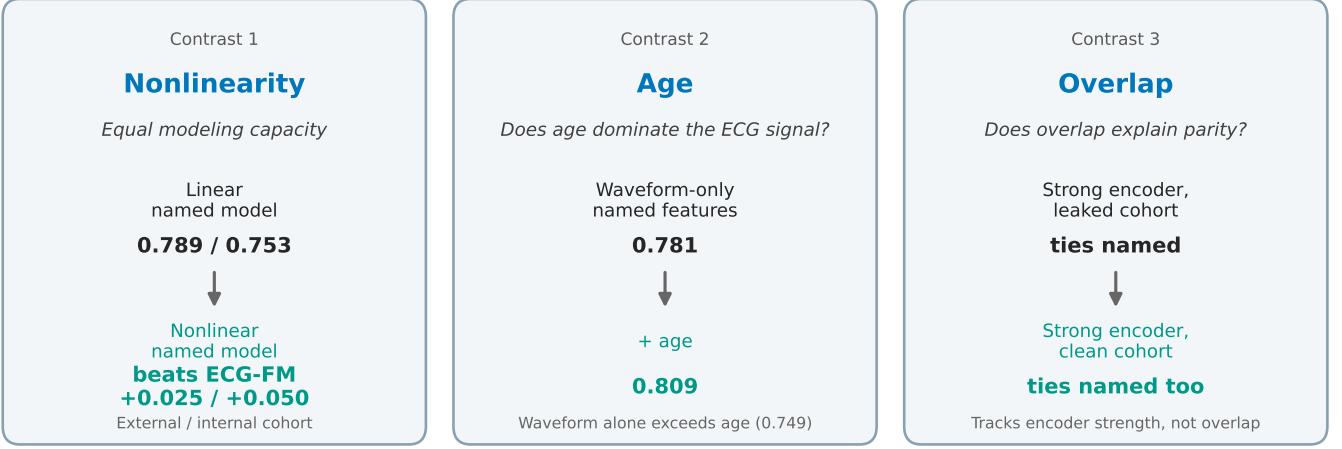


Fig. 2. Three controlled contrasts that scope the apparent foundation-model edge. Nonlinearity: under matched (nonlinear) modeling capacity, the named model exceeds ECG-FM by +0.025 (external) and +0.050 (internal), whereas a linear named model scores 0.789/0.753. Age: the ECG waveform features alone score 0.781 externally, above age alone at 0.749, and adding age raises this to 0.809; age is itself a nameable feature. Overlap: the strong encoders reach parity on both their own pretraining cohort and a cohort they never saw (ECG-JEPA and ST-MEM tie on clean MIMIC as well as on leaked CODE-15), so pretraining overlap does not explain the parity and the pattern tracks encoder strength. The three factors are shown as independent contrasts and are not claimed to sum to a single gap.

the six encoders, the named model won 7 of the 12 head-to-head comparisons and tied 5, with zero losses (Table I); a forest plot of the twelve deltas with 95% CIs is shown in Fig. 1, and the three scoping contrasts are summarized in Fig. 2. The five ties were against ECG-JEPA and ST-MEM, the two strongest encoders in the roster, on both cohorts, and ECGFounder on the internal cohort, and every other FM lost by a margin whose 95 percent confidence interval excluded zero. All seven wins survive Bonferroni family-wise correction across the twelve comparisons. The two CODE-15 ties are formally equivalent within the pre-specified $M = 0.01$ margin by TOST (ECG-JEPA and ST-MEM), and the three MIMIC ties are TOST-inconclusive.

The head-to-head pattern tracks encoder strength rather than pretraining overlap. On the external cohort the named model led ECGFounder by +0.031, ECG-FM by +0.019, HuBERT-ECG by +0.077, and xECG by +0.077, and tied ECG-JEPA at +0.002 and ST-MEM at -0.001. On the internal cohort it led ECG-FM by +0.033, HuBERT-ECG by +0.049, and xECG by +0.043, and tied ECGFounder at +0.007, ECG-JEPA at -0.003, and ST-MEM at +0.002. The two encoders that reach the named model, ECG-JEPA and ST-MEM, do so on both a cohort they were pretrained on (CODE-15, leaked) and a cohort they never saw (MIMIC, clean), so overlap does not manufacture the parity. The weaker overlapping encoders, HuBERT-ECG and xECG, lose by 0.04 to 0.08 on the very cohort they were pretrained on. The clean-cohort tie is the load-bearing observation: it shows that the encoders reaching the named model are genuinely strong rather than memorizing an evaluation set. The pretraining-overlap tag therefore functions as a fairness control, giving a leaked encoder the benefit of its own cohort, and the named model still matches or beats it.

B. Cross-cohort transportability

The strongest test of a prognostic model is whether it transfers to a cohort it was not trained on. We trained each model on one cohort and tested it on the other, using a common name-aligned schema of 227 features (age, sex, rich-waveform, and morphology descriptors; intervals were excluded because they are not comparable across cohorts). The transparent named model is gradient-boosted, and each FM was given its best transfer probe. This arm is the constructive core of the study: the transparent named model beats a leakage-clean ECGFounder in both transfer directions, the positive-transfer condition, and no FM beats the named model in both directions (Table II).

Transferring from MIMIC to CODE-15, the named model reached AUROC 0.762, and every FM scored below it, from ECGFounder at 0.714 (-0.048 [-0.058, -0.039], clean) down to xECG at 0.610 (-0.152 [-0.163, -0.142], leaked); the named model also beat age-only (0.743) and age+sex (0.748). Transferring from CODE-15 to MIMIC, the named model reached 0.738 and again beat age-only (0.672). The named model beats a leakage-clean ECGFounder in both directions, and both confidence intervals exclude zero (-0.048 [-0.058, -0.039] and -0.013 [-0.025, -0.002]), so this is a significant transfer win rather than a point estimate alone. Two strong encoders edge the named model in the CODE-15 to MIMIC direction: ECG-JEPA by +0.013 [+0.002, +0.025], whose interval excludes zero and is therefore a significant one-directional edge, and ST-MEM by +0.010 [-0.001, +0.021], whose interval includes zero and is not significant; the named model beats both in the other direction. ECG-FM, competitive within a cohort, is the encoder that collapses under transfer, to 0.685 and 0.615, losing to the named model by 0.077 and 0.123. The defensible claim is therefore symmetric: no FM beats the named model in both directions, while the two strong encoders, ECG-JEPA and ST-MEM, transport comparably to

TABLE I

TRANSPARENT GRADIENT-BOOSTED NAMED MODEL VERSUS EACH FOUNDATION MODEL’S BEST FROZEN-EMBEDDING PROBE (BEST OF A LINEAR AND A BOOSTED PROBE). PAIRED PATIENT-LEVEL BOOTSTRAP, 95% CI; POSITIVE DELTA FAVORS THE NAMED MODEL. CLEAN = THE FM WAS NOT PRETRAINED ON THAT COHORT, LEAKED = IT WAS.

Cohort	Named AUROC	Foundation model	Overlap	Best-probe AUROC	Δ (95% CI)	Outcome
CODE-15 (external)	0.811	ECGFounder	clean	0.781	+0.031 [+0.023, +0.038]	named wins
CODE-15 (external)	0.811	ECG-FM	clean	0.792	+0.019 [+0.011, +0.027]	named wins
CODE-15 (external)	0.811	HuBERT-ECG	leaked	0.734	+0.077 [+0.068, +0.086]	named wins
CODE-15 (external)	0.811	xECG	leaked	0.735	+0.077 [+0.067, +0.086]	named wins
CODE-15 (external)	0.811	ECG-JEPA	leaked	0.810	+0.002 [−0.005, +0.008]	tie
CODE-15 (external)	0.811	ST-MEM	leaked	0.812	−0.001 [−0.008, +0.006]	tie
MIMIC (internal)	0.773	ECGFounder	clean	0.766	+0.007 [−0.003, +0.016]	tie
MIMIC (internal)	0.773	ECG-FM	leaked	0.740	+0.033 [+0.021, +0.044]	named wins
MIMIC (internal)	0.773	HuBERT-ECG	leaked	0.725	+0.049 [+0.039, +0.059]	named wins
MIMIC (internal)	0.773	xECG	clean	0.730	+0.043 [+0.033, +0.054]	named wins
MIMIC (internal)	0.773	ECG-JEPA	clean	0.776	−0.003 [−0.012, +0.005]	tie
MIMIC (internal)	0.773	ST-MEM	clean	0.771	+0.002 [−0.007, +0.012]	tie

the named model and even edge it in one direction.

C. Calibration and precision

Discrimination is not the only property a screening model must have, so we report calibration and precision for the named model on held-out patients [71], [73]. Calibration was near-ideal on both cohorts: the calibration slope was 1.02 externally and 1.01 internally, with intercepts of +0.09 and +0.08 and Brier scores of 0.100 and 0.099. Precision stayed well above the base rate, with an area under the precision-recall curve of 0.402 on CODE-15 (prevalence 0.141) and 0.330 on MIMIC (prevalence 0.129), roughly 2.5 to 3 times prevalence. The screening-grade claim therefore rests on calibrated risk and precision, and a decision-curve net-benefit analysis [72] is a natural next step that we leave to future work.

D. Contrast 1: the apparent edge is probe nonlinearity

The first contrast locates an FM’s apparent edge in probe nonlinearity rather than in un-nameable physiology. The deliberately linear named-plus-age model scored 0.789 on the external cohort and 0.753 on the internal cohort. That linear control is itself competitive: externally it sits above ECG-Founder’s best probe at 0.781 though below ECG-FM’s at 0.792, so even a linear named model holds its own against real encoders. Read against that linear control alone, a clean FM can still look competitive or better on some comparisons (ECG-FM externally, several encoders internally), which is how a benchmark that omits a strong named baseline could conclude that the FM adds value.

That comparison is misleading because it varies two things at once, the feature set and the model class. When the ECG-FM embedding is given the same boosted head as the named features, matching the model class exactly, the named model leads by +0.025 (95% CI [+0.017, +0.033]) on the external cohort and by +0.050 (95% CI [+0.038, +0.061]) on the internal cohort. The apparent FM advantage is therefore attributable to probe nonlinearity, consistent with the general finding that tree ensembles outperform deep learning

on tabular inputs [37], [38]. This is a contrast against a labelled weaker baseline rather than a decomposition of the named score into a linear term and a nonlinear term.

E. Contrast 2: the named edge over ECG-only features tracks age

The second contrast weighs age against the rest of the named signal. Named features computed from the ECG waveform alone, with age and intervals withheld, scored 0.781 on the external cohort and 0.741 on the internal cohort; adding age raised these to 0.809 and 0.771. Age was the single strongest predictor by permutation importance and by certificate rank, followed by the RR interval, the named form of heart rate, yet the waveform features collectively exceeded age alone (0.749 external, 0.668 internal), so the aggregate signal is genuinely in the ECG.

Two demographic baselines bound the ECG-specific contribution from the other side. An age-only gradient-boosted model reached AUROC 0.749 externally and 0.668 internally, and adding sex changed little (0.755 and 0.666), so sex carries essentially no independent prognostic weight here. The full named model added a measured ECG-specific increment over age alone of +0.062 [+0.056, +0.069] on CODE-15 and +0.105 [+0.093, +0.117] on MIMIC. The named model thus adds six to ten AUROC points over age, a substantial and quantified ECG-specific effect that replaces any vague appeal to performance above chance. Age remains a strong single covariate, one that deep networks are known to estimate from the ECG [19] and that predicts mortality through network-estimated ECG age [22], consistent with the wider evidence that age informs short-horizon mortality models [26]. Adding age to the waveform features nonetheless raised AUROC only modestly, from 0.781 to 0.809 externally, so age is the single strongest feature without dominating the aggregate ECG signal.

The numbers these first two contrasts turn on are separate controlled contrasts, not sub-rungs of the full 231-feature gradient-boosted model. The linear named-plus-age model at 0.789 (Section IV-D) restricts the model class to isolate probe nonlinearity, and the waveform-only subset at 0.781 rising to

TABLE II

CROSS-COHORT TRANSFER. EACH MODEL IS TRAINED ON ONE COHORT AND TESTED ON THE OTHER OVER A COMMON 227-FEATURE NAME-ALIGNED SCHEMA (AGE, SEX, RICH-WAVEFORM, MORPHOLOGY; INTERVALS EXCLUDED). THE TRANSPARENT NAMED MODEL IS GRADIENT-BOOSTED; EACH FM USES ITS BEST TRANSFER PROBE. Δ IS AUROC RELATIVE TO THE TRANSPARENT NAMED MODEL ON THE SAME TEST COHORT, WITH THE PAIRED PATIENT-LEVEL BOOTSTRAP 95% CI; OVERLAP IS TAGGED WITH RESPECT TO THE TEST COHORT.

Model	Train MIMIC, test CODE-15			Train CODE-15, test MIMIC		
	AUROC	Δ (95% CI)	Overlap	AUROC	Δ (95% CI)	Overlap
Transparent named	0.762	—	—	0.738	—	—
Age only	0.743	−0.020 [−0.027, −0.012]	—	0.672	−0.066 [−0.077, −0.056]	—
Age + sex	0.748	−0.014 [−0.021, −0.007]	—	0.675	−0.063 [−0.074, −0.053]	—
ECGFounder	0.714	−0.048 [−0.058, −0.039]	clean	0.725	−0.013 [−0.025, −0.002]	clean
HuBERT-ECG	0.644	−0.118 [−0.129, −0.108]	leaked	0.704	−0.035 [−0.047, −0.021]	leaked
ECG-JEPA	0.732	−0.031 [−0.040, −0.021]	leaked	0.752	+0.013 [+0.002, +0.025]	clean
xECG	0.610	−0.152 [−0.163, −0.142]	leaked	0.688	−0.051 [−0.064, −0.038]	clean
ECG-FM	0.685	−0.077 [−0.089, −0.066]	clean	0.615	−0.123 [−0.139, −0.106]	leaked
ST-MEM	0.737	−0.026 [−0.034, −0.017]	leaked	0.748	+0.010 [−0.001, +0.021]	clean

0.809 with age restricts the feature set to isolate age; each deliberately weakens one axis to expose one factor. The full 231-feature model at 0.811 restricts neither axis and stays the reference throughout.

F. Contrast 3: pretraining overlap is a fairness control

The third contrast confirms that pretraining overlap does not manufacture parity. The two encoders that reach the named model, ECG-JEPA and ST-MEM, tie it on their leaked pre-training cohort (CODE-15) and on the clean cohort they never saw (MIMIC), so the parity survives removing the overlap. Meanwhile xECG, also pretrained on CODE-15, lost to the named model by +0.077 on that same cohort because its probe is weaker, and HuBERT-ECG, leaked on both cohorts, lost on both. Overlap is therefore neither necessary nor sufficient for parity here: the pattern follows encoder strength, and the overlap tag serves only to give a leaked encoder the benefit of its own cohort [53], [56]. The clean-cohort ties are the evidence that the strong encoders are genuinely strong [1], [2], [69].

G. Complementarity: equal discrimination is not identical information

Equal AUROC does not mean the named model and an FM use the same information, so we measured, for each encoder, the conditional increment from adding its embedding to the named model and the Spearman rank-correlation of the per-patient risk scores. A named-plus-FM ensemble, boosting on the named features together with a clean FM embedding, exceeded the named model alone by +0.013 (95% CI [+0.009, +0.017]) on CODE-15 and by +0.002 on MIMIC. Per encoder, the strong encoders added a small increment whose CI excluded zero: on CODE-15, ECG-JEPA +0.018, ST-MEM +0.020, ECG-FM +0.013, and ECGFounder +0.006; on MIMIC, ST-MEM +0.020, ECG-JEPA +0.016, and ECGFounder +0.014. The weak encoders added nothing (HuBERT-ECG −0.004 and −0.001, xECG −0.002 and −0.000). Rank agreement was high but imperfect, with Spearman ρ between 0.61 and 0.81 across encoders and cohorts. The named model

TABLE III

PER-PATIENT SUFFICIENT-REASON CERTIFICATE ON THE COMPACT GRADIENT-BOOSTED NAMED MODEL, COMPUTED ON THE HELD-OUT PATIENTS.

Certificate statistic	Value
Sufficiency level (delta)	0.9
Median certificate length	1 feature
Mean certificate length	1.5 features
90th-percentile length	1 feature
Empirical coverage	0.96
Held-out AUROC (certified)	0.775
Held-out AUROC (age-excluded, certified)	0.741
Certificates containing age	100%

and the strong FMs therefore use largely the same information for this endpoint, though not identically, which scopes the claim to head-to-head discrimination in the frozen-probe regime: no FM out-discriminates the named model, while a leakage-clean FM contributes a small disclosed increment.

H. Per-patient certificates are short and well-covered

Every named prediction carried a short, provable certificate. On the held-out set, the delta = 0.9 sufficient-reason certificate had a median length of 1 feature, a mean of 1.5, and a 90th percentile of 1, with an empirical coverage of 0.96 against the stated 0.90 guarantee (Table III). The certified compact model retained an AUROC of 0.775 on the held-out cohort, matching the 0.773 of the uncertified named model, so certification carried no accuracy cost. Age appeared in 100 percent of certificates: for the median patient, the sufficient reason is age.

Because age dominates the certificate, we also certified a model constrained not to use age. That age-excluded model certified at a held-out AUROC of 0.741 with a median certificate of one feature that names an inspectable ECG measurement, most often heart rate (the RR interval), and otherwise the aVR and lead-II QRS and T-wave amplitudes, the P-wave axis, aVR band-power, or V1 kurtosis. The take-home is that the certificate reduces to age for the typical patient, and where it exceeds age it names a single inspectable ECG measurement.

A median certificate of one feature means that, for a typical patient, a single named value provably suffices to explain the model’s above-threshold prediction at the stated coverage, a per-patient auditable statement that no post-hoc importance score provides [44], [46]. The empirical coverage of 0.96 exceeding the 0.90 target indicates that the delta-sufficient relaxation was conservative on these cohorts. The certificate is an application of established abductive-explanation machinery [48], [49], [51], [52], reported here to show that a competitive named model is also a certifiable one.

I. The compact bank matches the full superset

Compactness costs almost nothing. The 231-feature compact bank scored within 0.006 AUROC of the 891-feature superset on both cohorts, so the named model used for every comparison above is small enough to inspect in full. A clinician can enumerate the inputs, and the certificate then names the one or two that decided each case. This closes the loop from a competitive score to an auditable model without a separate distillation step, which is what makes the certificate in Section IV-H a statement about the model of record itself.

V. DISCUSSION

A. Principal findings

For one-year all-cause mortality from a single 12-lead ECG, a transparent named-feature model was competitive with the best frozen-embedding probe of six real ECG foundation models, and no FM out-discriminated it under equal modeling. The named model won 7 of 12 comparisons and tied 5, against ECG-JEPA and ST-MEM on both cohorts and ECGFounder internally, reaching AUROC 0.811 externally and 0.773 internally with near-ideal calibration. The head-to-head pattern tracks encoder strength: the two strongest encoders tie on cohorts they never saw, while weaker encoders lose on cohorts they were pretrained on, so pretraining overlap is a fairness control rather than the explanation.

The most transferable finding is cross-cohort. Trained on one cohort and tested on the other, the named model beat a leakage-clean ECGFounder in both directions and beat age-only in both, while ECG-FM’s within-cohort competitiveness collapsed under transfer. The two strong encoders, ECG-JEPA and ST-MEM, transported comparably to the named model and even edged it in one direction, so the collapse is specific to ECG-FM and the weaker xECG and HuBERT-ECG rather than a general property of FM embeddings; a physiology-grounded named model transports in both directions, which is the property a deployable screening model most needs.

The scope of the claim is bounded. Equal discrimination is not identical information: a leakage-clean FM ensembled with the named model adds a small complementary increment of +0.013 [+0.009, +0.017] on CODE-15 and +0.002 on MIMIC, and the strong encoders each add +0.006 to +0.020 with rank agreement of 0.61 to 0.81. Age is the single strongest individual feature, yet the ECG waveform features collectively exceed age alone and the ECG-specific increment over age is substantial, a measured +0.062 externally and +0.105 internally. Within these bounds, for this endpoint under equal

modeling, no foundation model out-discriminated a named, certifiable baseline.

B. Relation to the closest neighbor and to prior baselines

The result extends the closest neighbor by exactly the controls it lacked. Lukyanenko *et al.* [24] showed on these same two cohorts that a gradient-boosted demographic-and-feature baseline is competitive with deep survival models, but did not compare against foundation models, did not tag leakage, ran no cross-cohort transfer arm, and did not certify predictions. Our delta is that FM comparison across six encoders, the pretraining-overlap tagging, the cross-cohort transfer arm, the calibration and precision reporting, and the per-patient certificate, which together move the claim from “a boosted baseline is competitive with deep survival networks” to “no real ECG-FM out-discriminates a certified named baseline, and the result strengthens under cross-cohort transfer.”

The finding is consistent with a converging body of evidence. Compact models rival foundation models across ECG tasks in independent benchmarks [8], [9]. Age and demographics dominate externally validated one-to-three-year mortality models [26], which is exactly the pattern our age contrast recovers. Interpretable representations retain prognostic value on this endpoint [23], and interpretable models can match black boxes on high-stakes tasks when the features are well chosen [39]. Our contribution within that literature is the controlled demonstration of when an FM appears to win and why, on a mortality endpoint, with leakage held to account; the observation that a simple model competes precedes us.

C. The cost of explainability, bounded for this endpoint

The claim that explainable ECG models pay an accuracy penalty [43] does not hold for this endpoint under equal modeling. Patlatzoglou *et al.* [43] reported a cost of explainability across ECG tasks, and we credit that finding before qualifying it. Our result is a bounded counterexample, not a general overturn. For one-year mortality, a fully transparent named model with per-patient certificates matched or beat every FM probe we tested, so here the transparent model paid no accuracy penalty at all. We do not extend that conclusion to other tasks, where the trade-off Patlatzoglou *et al.* describe may well hold.

The certificate sharpens the interpretability beyond what post-hoc methods offer. Attribution methods [44], [45] and saliency maps [47] estimate importance without guaranteeing sufficiency, and post-hoc explanation in health care has been criticized as unreliable [40]. A delta-sufficient certificate [48], [49], [51], [52] instead states a provable per-patient reason, which is the form of guarantee that formal XAI argues clinical decisions require [41], [42], [50]. Reporting a certified model that is also competitive is the practical payoff: a clinician gets a defensible reason for each prediction without trading away discrimination, which is the combination the cost-of-explainability framing assumes is unavailable.

D. Pretraining overlap, shortcut learning, and generalization

The overlap analysis situates our result within a broader reproducibility concern. Training on a cohort that also serves as an evaluation set inflates benchmark performance across fields [53], is quantified in medical machine learning [56], and biases pooled estimates upward [57]. We tagged every comparison for overlap and found that it does not explain the FM results here: the encoders reaching the named model tie it on clean cohorts too, and a leaked but weaker encoder loses on its own pretraining cohort. Overlap is a control we account for, not a demonstrated cause of any FM edge.

The cross-cohort transfer arm speaks directly to distribution shift. AUROC is unstable across populations [62], cross-site collapse is well documented [63], [64], and most clinical prediction models carry a high risk of bias and overoptimism under external appraisal [65]. Predictors that stay stable across environments capture invariant rather than spurious structure [66], [67]. Our transfer results are the concrete instance: the named model transferred in both directions while ECG-FM, competitive within a cohort, collapsed across cohorts and the two strong encoders transported comparably, consistent with physiologically grounded named inputs being less exposed to cohort-specific shortcuts [58]–[61] than ECG-FM’s learned embedding.

E. Practical implications

Two practical implications follow for teams choosing a model for ECG-based risk screening. First, an FM gain reported on a cohort should be checked against a strong named baseline under matched modeling and against the encoder’s pretraining provenance before it is treated as real, because on our endpoint the apparent gains dissolved once those two checks were applied. A benchmark that omits the named baseline, or that scores an encoder on a cohort it was pretrained on, can report an advantage that would not survive deployment. Second, when a named model matches the encoder, its per-patient certificate offers something an embedding cannot: a defensible, case-specific reason of median length one that a clinician can inspect at the point of care. For a screening-grade endpoint where the model routes attention rather than dictates treatment, that auditability is a deployment advantage in its own right, independent of the small external increment a clean encoder can add on top.

F. Limitations

Six limitations bound the claim. First, age is the single strongest individual feature and appears in every certificate, so a large share of the discriminative signal is carried by one demographic covariate; the ECG waveform features nonetheless exceed age alone (0.781 versus 0.749 externally, 0.741 versus 0.668 internally) and the ECG-specific increment over age is substantial, at +0.062 externally and +0.105 internally [19], [22]. Second, the endpoint is single, one-year all-cause mortality from one ECG, so the result does not transfer automatically to other horizons, cause-specific outcomes, or serial recordings. Third, the named features depend on NeuroKit

delineation, which may behave asymmetrically across cohorts; the same pipeline was applied to both and the result holds on the cleaner internal cohort, but delineation quality remains a confounder we cannot fully rule out.

Fourth, the heart-rate-variability features are computed on approximately ten-second recordings, whereas the measurement standards were set on longer recordings [70], so we restrict those features to time-domain measures defensible at that length. Fifth, we probed frozen embeddings rather than fine-tuning the encoders, following standard protocol [7]; a fine-tuning budget could change the FM numbers, and our claim is scoped to the frozen-embedding regime. Sixth, a leakage-clean FM adds a measured complementary increment of +0.013 (95% CI [+0.009, +0.017]) on CODE-15 and +0.002 on MIMIC, with per-encoder increments of +0.006 to +0.020 for the strong encoders, so the claim is that no FM out-discriminates the named model, not that an FM captures nothing extra. The single within-cohort concern is addressed by the cross-cohort transfer arm, which tests generalization directly rather than relying on two within-cohort experiments alone.

VI. CONCLUSION

For one-year all-cause mortality from a single 12-lead ECG, no real foundation model out-discriminated a transparent named-feature model under equal modeling. Across two cohorts the named model won 7 of 12 head-to-head comparisons and tied 5, against ECG-JEPA and ST-MEM on both cohorts and ECGFounder internally, with zero losses, reaching AUROC 0.811 externally and 0.773 internally with near-ideal calibration. The head-to-head pattern tracks encoder strength, and the strongest, most transferable result is cross-cohort: trained on one cohort and tested on the other, the named model beat a leakage-clean encoder in both directions while a within-cohort-competitive FM collapsed under transfer.

Every named prediction carries a per-patient provable sufficient-reason certificate with a median length of one feature and coverage 0.96, applying established abductive-explanation machinery [48], [49] rather than a new method, and the whole comparison regenerates from raw data with one command. The claim is bounded: a leakage-clean FM adds a small complementary increment of +0.013, and although age is the single strongest feature the ECG waveform features collectively exceed age alone, so this is a positive competitiveness result about head-to-head discrimination.

What a convincing ECG-FM result for this endpoint would need is now clear. It would have to out-discriminate a strong named baseline under matched modeling, transfer across cohorts rather than win only within its pretraining distribution, and ideally carry a per-patient reason a clinician can audit. On these two cohorts no encoder met that bar, though ECG-JEPA and ST-MEM came closest, tying on both a leaked and a clean cohort. Future work should test whether a fine-tuning budget, other endpoints and horizons, or genuinely disjoint pretraining changes that picture, and should keep a named, certifiable baseline in the frame while doing so. Interpretable models remain the right default when they match the black

box [39], and the closest FM-free baseline [24], evaluated with overlap held to account [53], is the one to extend.

DECLARATIONS

Data availability. Both cohorts are public, credentialed resources. MIMIC-IV-ECG is available from PhysioNet [68] and CODE-15 from Zenodo [69]. Raw archives are not redistributed with the code, and the reproduction script documents the exact acquisition steps and checks for their presence before running.

Code availability. All code to reproduce every number, table, and figure in this paper from the raw data is released in a public repository, driven by a single command (`run_all.sh`) with external-prerequisite checks and no hardcoded paths. A permanent archive with a DOI will be deposited on acceptance.

Funding. None.

Conflicts of interest. The author declares no competing interests.

Ethics. This study is a secondary analysis of two de-identified, publicly available, credentialed ECG datasets, used under their respective data use agreements. No new human-subjects data were collected.

AI-use disclosure. Generative AI tools assisted with drafting and language editing of this manuscript. The author designed the study, implemented and ran the analysis, and verified all methods, results, and references; the author takes full responsibility for the accuracy and integrity of the work.

Author contributions (CRediT). A. Faes: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review and editing.

REFERENCES

- [1] J. Li, A. Aguirre, J. Moura, C. Liu, L. Zhong, C. Sun, G. Clifford, B. Westover, and S. Hong, “An Electrocardiogram Foundation Model Built on over 10 Million Recordings,” *NEJM AI*, vol. 2, no. 7, 2025, doi: 10.1056/Aloa2401033. (Preprint: arXiv:2410.04133, full title adds “with External Evaluation across Multiple Domains.”)
- [2] K. McKeen, L. Oliva, S. Masood, *et al.*, “ECG-FM: an open electrocardiogram foundation model,” *JAMIA Open*, vol. 8, no. 5, art. no. ooaf122, 2025, doi: 10.1093/jamiaopen/ooaf122. (Preprint: arXiv:2408.05178.)
- [3] E. Coppola, M. Savardi, A. Signoroni, *et al.*, “HuBERT-ECG: a self-supervised foundation model for broad and scalable cardiac applications,” *medRxiv* 2024.11.14.24317328, 2024, doi: 10.1101/2024.11.14.24317328.
- [4] S. Kim, “Learning General Representation of 12-Lead ECG With a Joint-Embedding Predictive Architecture,” *Neural Networks*, art. no. 109334, 2026, doi: 10.1016/j.neunet.2026.109334. (Preprint: arXiv:2410.08559.)
- [5] Y. Na, M. Park, Y. Tae, and S. Joo, “Guiding Masked Representation Learning to Capture Spatio-Temporal Relationship of Electrocardiogram (ST-MEM),” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024. (arXiv:2402.09450.)
- [6] C. Liu, Z. Wan, C. Ouyang, A. Shah, W. Bai, and R. Arcucci, “Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement (MERL),” in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR 235, 2024.
- [7] R. Lunelli, A. Nicolson, S. M. Proell, S. J. Reinstadler, A. Bauer, and C. Dlaska, “BenchECG and xECG: a benchmark and baseline for ECG foundation models,” arXiv:2509.10151, 2025.
- [8] M. A. Al-Masud, J. M. Lopez Alcaraz, and N. Strodthoff, “Benchmarking ECG Foundational Models: A Reality Check Across Clinical Tasks,” arXiv:2509.25095, 2025. (ICLR 2026.)
- [9] F. Filice, E. De Rose, S. Bartucci, F. Calimeri, and S. Perri, “Looking Beyond Accuracy: A Holistic Benchmark of ECG Foundation Models,” arXiv:2601.21830, 2026.
- [10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020. (arXiv:2006.11477.)
- [11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, 2021. (arXiv:2106.07447.)
- [12] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, “Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (I-JEPA),” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023. (arXiv:2301.08243.)
- [13] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, “Masked Autoencoders Are Scalable Vision Learners (MAE),” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022. (arXiv:2111.06377.)
- [14] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, “On the Opportunities and Risks of Foundation Models,” arXiv:2108.07258, 2021.
- [15] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixao, *et al.*, “Automatic diagnosis of the 12-lead ECG using a deep neural network,” *Nature Commun.*, vol. 11, art. no. 1760, 2020, doi: 10.1038/s41467-020-15432-4.
- [16] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory ECGs using a deep neural network,” *Nature Med.*, vol. 25, pp. 65-69, 2019, doi: 10.1038/s41591-018-0268-3.
- [17] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, *et al.*, “An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction,” *The Lancet*, vol. 394, no. 10201, pp. 861-867, 2019, doi: 10.1016/S0140-6736(19)31721-0.
- [18] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, *et al.*, “Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram,” *Nature Med.*, vol. 25, pp. 70-74, 2019, doi: 10.1038/s41591-018-0240-2.
- [19] Z. I. Attia, P. A. Friedman, P. A. Noseworthy, *et al.*, “Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs,” *Circulation: Arrhythmia Electrophysiol.*, vol. 12, no. 9, art. no. e007284, 2019, doi: 10.1161/CIRCEP.119.007284.
- [20] S. Raghunath, A. E. Ulloa Cerna, L. Jing, *et al.*, “Prediction of mortality from 12-lead ECG voltage data using a deep neural network,” *Nature Med.*, vol. 26, pp. 886-891, 2020, doi: 10.1038/s41591-020-0870-z.
- [21] S. Raghunath, J. M. Pfeifer, A. E. Ulloa-Cerna, *et al.*, “Deep Neural Networks Can Predict New-Onset Atrial Fibrillation From the 12-Lead ECG and Help Identify Those at Risk of Atrial Fibrillation-Related Stroke,” *Circulation*, vol. 143, pp. 1287-1298, 2021, doi: 10.1161/CIRCULATIONAHA.120.047829.
- [22] E. M. Lima, A. H. Ribeiro, G. M. M. Paixao, *et al.*, “Deep neural network-estimated electrocardiographic age as a mortality predictor,” *Nature Commun.*, vol. 12, art. no. 5117, 2021, doi: 10.1038/s41467-021-25351-7.
- [23] R. R. van de Leur, M. N. Bos, K. Taha, *et al.*, “Improving explainability of deep neural network-based ECG interpretation using variational autoencoders,” *Eur. Heart J. Digit. Health*, vol. 3, no. 3, pp. 390-404, 2022, doi: 10.1093/ehjdh/ztac038.
- [24] P. Lukyanenko, J. Mayourian, M. Liu, *et al.*, “Deep survival analysis from adult and pediatric electrocardiograms: a multi-center benchmark study,” *BioData Mining*, vol. 19, art. no. 6, 2026, doi: 10.1186/s13040-025-00510-4. (Preprint: arXiv:2406.17002.)
- [25] C. H. Kim, S. Al-Kindi, Y. Tarabichi, S. Gohel, R. Vyas, and S. Srinivasan, “Machine Learning to Predict 10-year Cardiovascular Mortality from the Electrocardiogram: Analysis of the Third National Health and Nutrition Examination Survey (NHANES III),” *medRxiv* 2021.09.09.21263327, 2021, doi: 10.1101/2021.09.09.21263327. (Preprint only; no peer-reviewed journal version located.)
- [26] L. Ho, C. Pugh, S. Seth, *et al.*, “Performance of models for predicting 1-year to 3-year mortality in older adults: a systematic review of externally validated models,” *Lancet Healthy Longevity*, vol. 5, no. 3, pp. e227-e235, 2024, doi: 10.1016/S2666-7568(23)00264-7.
- [27] H. Tsuji, F. J. Venditti Jr., E. S. Manders, *et al.*, “Reduced heart rate variability and mortality risk in an elderly cohort (Framingham Heart Study),” *Circulation*, vol. 90, no. 2, pp. 878-883, 1994, doi: 10.1161/01.CIR.90.2.878.
- [28] J. M. Dekker, E. G. Schouten, P. Klootwijk, *et al.*, “Heart rate variability from short electrocardiographic recordings predicts mortality from all causes in middle-aged and elderly men (The Zutphen Study),” *Amer. J.*

- Epidemiol.*, vol. 145, no. 10, pp. 899-908, 1997, doi: 10.1093/oxford-journals.aje.a009049.
- [29] D. Zhang, X. Shen, and X. Qi, "Resting heart rate and all-cause and cardiovascular mortality in the general population: a meta-analysis," *Can. Med. Assoc. J. (CMAJ)*, vol. 188, no. 3, pp. E53-E63, 2016, doi: 10.1503/cmaj.150535.
- [30] E. G. Schouten, J. M. Dekker, P. Meppelink, F. J. Kok, J. P. Vandenbroucke, and J. Pool, "QT interval prolongation predicts cardiovascular mortality in an apparently healthy population," *Circulation*, vol. 84, no. 4, pp. 1516-1523, 1991, doi: 10.1161/01.CIR.84.4.1516.
- [31] A. D. Desai, T. S. Yaw, T. Yamazaki, A. Kaykha, S. Chun, and V. F. Froelicher, "Prognostic Significance of Quantitative QRS Duration," *Amer. J. Med.*, vol. 119, no. 7, pp. 600-606, 2006, doi: 10.1016/j.amjmed.2005.08.028.
- [32] S. Cheng, M. J. Keyes, M. G. Larson, *et al.*, "Long-term Outcomes in Individuals With Prolonged PR Interval or First-Degree Atrioventricular Block (Framingham Heart Study)," *JAMA*, vol. 301, no. 24, pp. 2571-2577, 2009, doi: 10.1001/jama.2009.888.
- [33] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189-1232, 2001, doi: 10.1214/aos/1013203451.
- [34] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
- [35] G. Ke, Q. Meng, T. Finley, *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 3146-3154.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011.
- [37] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets and Benchmarks Track*, 2022, (arXiv:2207.08815).
- [38] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84-90, 2022, doi: 10.1016/j.inffus.2021.11.011.
- [39] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Mach. Intell.*, vol. 1, pp. 206-215, 2019, doi: 10.1038/s42256-019-0048-x.
- [40] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745-e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.
- [41] Z. C. Lipton, "The Mythos of Model Interpretability," *ACM Queue*, vol. 16, no. 3, pp. 31-57, 2018, doi: 10.1145/3236386.3241340.
- [42] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.
- [43] K. Patlatzoglou, L. Pastika, J. Barker, *et al.*, "The cost of explainability in artificial intelligence-enhanced electrocardiogram models," *npj Digit. Med.*, vol. 8, art. no. 747, 2025, doi: 10.1038/s41746-025-02122-y.
- [44] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions (SHAP)," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765-4774.
- [45] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier (LIME)," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135-1144, doi: 10.1145/2939672.2939778.
- [46] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1527-1535, doi: 10.1609/aaai.v32i1.11491.
- [47] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 9505-9515.
- [48] A. Ignatiev, N. Narodytska, and J. Marques-Silva, "Abduction-Based Explanations for Machine Learning Models," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1511-1519, doi: 10.1609/aaai.v33i01.33011511.
- [49] Y. Izza, X. Huang, A. Ignatiev, N. Narodytska, M. Cooper, and J. Marques-Silva, "On computing probabilistic abductive explanations," *Int. J. Approx. Reason.*, vol. 159, art. no. 108939, 2023, doi: 10.1016/j.ijar.2023.108939. (Preprint: arXiv:2212.05990.)
- [50] J. Marques-Silva and A. Ignatiev, "Delivering Trustworthy AI through Formal XAI," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 11, 2022, pp. 12342-12350, doi: 10.1609/aaai.v36i11.21499.
- [51] A. Shih, A. Choi, and A. Darwiche, "A Symbolic Approach to Explaining Bayesian Network Classifiers," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 5103-5111, doi: 10.24963/ijcai.2018/708.
- [52] A. Darwiche and A. Hirth, "On the Reasons Behind Decisions," in *Proc. 24th Eur. Conf. Artif. Intell. (ECAI)*, Frontiers Artif. Intell. Appl. 325, 2020, pp. 712-720. (arXiv:2002.09284.)
- [53] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
- [54] W. Shi, A. Ajith, M. Xia, *et al.*, "Detecting Pretraining Data from Large Language Models," arXiv:2310.16789, 2023. (ICLR 2024.)
- [55] O. Sainz, J. Campos, I. Garcia-Ferrero, *et al.*, "NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for each Benchmark," in *Findings Assoc. Comput. Linguist. (EMNLP)*, 2023. (arXiv:2310.18018.)
- [56] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nature Commun.*, vol. 15, art. no. 1829, 2024, doi: 10.1038/s41467-024-46150-w.
- [57] L. A. van de Mortel and G. A. van Wingen, "Data leakage in machine learning studies creep into meta-analytic estimates of predictive performance," *Molecular Psychiatry*, vol. 30, pp. 6070-6071, 2025, doi: 10.1038/s41380-025-03336-y.
- [58] R. Geirhos, J.-H. Jacobsen, C. Michaelis, *et al.*, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, pp. 665-673, 2020, doi: 10.1038/s42256-020-00257-z.
- [59] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Mach. Intell.*, vol. 3, pp. 610-619, 2021, doi: 10.1038/s42256-021-00338-7.
- [60] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re, "Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging," in *Proc. ACM Conf. Health, Inference, Learn. (CHIL)*, 2020, pp. 151-159, doi: 10.1145/3368555.3384468.
- [61] C. Ong Ly, B. Unnikrishnan, T. Tadic, *et al.*, "Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data," *npj Digit. Med.*, vol. 7, art. no. 124, 2024, doi: 10.1038/s41746-024-01118-4.
- [62] F. D. van Leeuwen, E. W. Steyerberg, D. van Klaveren, B. Wessler, D. M. Kent, and E. W. van Zwet, "Instability of the AUROC of Clinical Prediction Models," *Statist. Med.*, vol. 44, art. no. e70011, 2025, doi: 10.1002/sim.70011.
- [63] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Med.*, vol. 15, no. 11, art. no. e1002683, 2018, doi: 10.1371/journal.pmed.1002683.
- [64] S. G. Finlayson, A. Subbaswamy, K. Singh, *et al.*, "The Clinician and Dataset Shift in Artificial Intelligence," *New Engl. J. Med.*, vol. 385, no. 3, pp. 283-286, 2021, doi: 10.1056/NEJMc2104626.
- [65] L. Wynants, B. Van Calster, G. S. Collins, *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, art. no. m1328, 2020, doi: 10.1136/bmj.m1328.
- [66] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant Risk Minimization," arXiv:1907.02893, 2019.
- [67] J. Peters, P. Buhlmann, and N. Meinshausen, "Causal inference by using invariant prediction: identification and confidence intervals," *J. Roy. Statist. Soc. B*, vol. 78, no. 5, pp. 947-1012, 2016, doi: 10.1111/rssb.12167.
- [68] B. Gow, T. Pollard, L. A. Nathanson, A. Johnson, B. Moody, *et al.*, "MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset (version 1.0)," *PhysioNet*, 2023, doi: 10.13026/4nqg-sb35.
- [69] A. H. Ribeiro, G. M. M. Paixao, E. M. Lima, *et al.*, "CODE-15%: a large scale annotated dataset of 12-lead ECGs," *Zenodo*, 2021, doi: 10.5281/zenodo.4916206.
- [70] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043-1065, 1996, doi: 10.1161/01.CIR.93.5.1043.
- [71] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement," *Ann. Intern. Med.*, vol. 162, no. 1, pp. 55-63, 2015, doi: 10.7326/M14-0697.
- [72] A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," *Med. Decis. Making*, vol. 26, no. 6, pp. 565-574, 2006, doi: 10.1177/0272989X06295361.
- [73] B. Van Calster, D. J. McLernon, M. van Smeden, E. W. Steyerberg, *et al.*, "Calibration: the Achilles heel of predictive analytics," *BMC Med.*, vol. 17, art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.