Empirical-Distribution Matching for Synthetic ECG Classification

Anonymous Authors[1]

Abstract

Conditional generative models are increasingly proposed as drop-in substitutes for restricted clinical datasets: train a generator once, release a synthetic cohort, and let practitioners reuse it without ever touching the source records. The practitioner's only design surface is then the sampling policy used to draw the synthetic training set. The main idea of this paper is that the standard rebalancing recipes from the CV and NLP imbalanced-learning literature do not transfer to this regime; the practitioner is better off matching the empirical training distribution as faithfully as possible at sample time. We test the idea by surveying thirteen sampling policies on a single 12-lead ECG latent diffusion model over PTB-XL using train-on-synthetic-test-on-real (TSTR) macro AUROC. No policy beats the naive bootstrap baseline at matched budget; `qstrat_matched` ties at one seed. The failures collapse onto three mechanisms: class-distribution distortion, within-class diversity collapse, and label-by-demographic joint decoupling.

1. Introduction

Synthetic-data pipelines for medical time series are increasingly deployed in a setting where the downstream practitioner never sees the underlying real records. A generator is trained once by a data custodian; the practitioner receives only synthetic samples and trains classifiers, fairness probes, or stress tests directly on those samples. The regime covers privacy-restricted clinical data (Thambawita et al., 2021) as well as differentially-private generators (Ghalebikesabi et al., 2023; Jordon et al., 2019), where the real records cannot be released even to downstream model trainers. In this regime, the practitioner's only design surface is the *sampling policy*: how to draw the conditioning vectors that the generator will expand into samples.

The default sampling policy in published synthetic-medical pipelines is to bootstrap real training-record conditioning tuples uniformly at random, matching the empirical training distribution by construction. This is unsatisfying to the practitioner who reads the imbalanced-learning literature: SMOTE-family variants (Chawla et al., 2002; Mariani et al., 2018; Dablain et al., 2022), synthetic minority generation (Ye-Bin et al., 2025), and recent surveys (Li et al., 2024; Liu et al., 2024) all converge on "rebalance toward the minorities". Two things differ in our setting: the practitioner cannot see the real records, and the "minority" samples come from a single shared generator rather than from a fixed real cohort. The main idea of this paper is that this regime change inverts the rebalancing recipe.

Contributions.

- We survey thirteen sampling policies for synthetic-only ECG classification on PTB-XL, all evaluated on a single shared latent diffusion generator at a matched downstream budget (Sections 3 and 4).
- Eleven of twelve alternatives underperform a naive bootstrap of real conditioning tuples at matched budget; the twelfth (`qstrat_matched`) ties within seed noise at a single seed and is the clearest follow-up rather than a positive finding (Section 5).
- The failures collapse onto three mechanisms that together cover the surface: class-distribution distortion, within-class diversity collapse, and label-by-demographic joint decoupling. A given policy can sit on more than one (Section 5).
- A $50/50$ real+synth blend matches the same-size real-data ceiling, confirming the corpus is additive; in the synthetic-only regime that motivates the paper, this additivity is out of reach (Section 6).

2. Related work

The practitioner's downstream-classifier metric, TSTR, was introduced for medical time-series GANs (Thambawita et al., 2021) and is now standard for ECG generators (Alcaraz & Strodthoff, 2022; Skorik & Avetisyan, 2024; Lai

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

et al., 2025). Most of these papers report a single TSTR number against a single sampling policy (a bootstrap of real conditioning tuples, in our taxonomy), without comparing alternative policies. Ibrahim et al. (2025) probe synthetic medical time series at the subgroup level and is the closest evaluation work; the focus there is per-subgroup fidelity rather than the sampling policy used to draw the synthetic training set.

3. Setup

Data. PTB-XL (Wagner et al., 2020) is the largest freely-redistributable 12-lead ECG dataset: 21,799 ten-second recordings from 18,885 patients, multi-label-annotated in the 71-code SCP-ECG ontology with patient demographics. We adopt the official 10-fold patient-disjoint split (Strodthoff et al., 2021) unchanged: folds 1–8 ($n_{\text{train}} = 17{,}418$) train the generator and the downstream classifiers, fold 9 is used only for early stopping, and fold 10 ($n_{\text{test}} = 2{,}198$) is held out for all reported results. The label distribution is heavily skewed: of the 71 SCP codes, 8 have prevalence $\geq 5\%$, 7 have 2–5%, and 37 have $< 2\%$.

Generator. A conditional latent diffusion model: a frozen 4.71M-parameter 1-D VAE encodes a 12×1000 ECG into a 4×128 latent, and a 5.36M-parameter U-Net denoiser operates in latent space, conditioned on a 76-dim vector that concatenates the 71-hot SCP label vector with five z-normalised demographic axes. Classifier-free guidance (Ho & Salimans, 2022) uses condition dropout 0.1 at training and CFG scale 3.0 at inference; sampling is 100-step DDIM (Song et al., 2021). The generator is held fixed; only the *sampling policy* that produces conditioning tuples changes across Table 1.

TSTR protocol. For each policy we draw $n = 8000$ conditioning tuples, generate ECGs, train an XResNet1d-50 (He et al., 2019) for 30 epochs from random init on the synthetic corpus, and report macro AUROC over the 51 evaluable SCP codes on real fold 10. We bucket results by real training prevalence (common $\geq 5\%$, moderate 2–5%, rare $< 2\%$) and average 2–5 seeds per policy; the seed varies both the synthetic draw and the classifier init. The downstream classifier always sees 8000 training samples; `qfilter_matched` and `diversity_oversample` generate a 16,000-sample pool and keep 8000, so they spend 2× the *generator* budget at matched downstream budget. The natural upper bound is `real_subset`: the same XResNet1d-50 trained on 8000 real records, which we use as a ceiling throughout.

4. Sampling policies

We organise the policies by which dimension of the empirical training distribution they deliberately deviate from.

Reference. `real_matched` re-uses an entire real training-record conditioning tuple drawn uniformly with replacement from folds 1–8: class prevalence, demographic marginals, and the label-by-demographic joint are all preserved by construction.

Class rebalancing. `uniform_class` equalises per-class prevalence by drawing one tuple from each non-empty SCP class in turn. `rare_boost` raises every class below 5% prevalence to that floor while leaving common classes alone. `rare_oversample_3x` multiplies every class's prevalence by 3 before normalisation.

Demographic. `demo_strat` grids each class by age tertile and sex into 6 cells, then sub-samples cells with at least 3 real records. `demo_balanced` bootstraps labels from real records but resamples demographics independently to enforce 50/50 sex and uniform age decile. `synth_cond` preserves real label co-occurrence but draws each demographic axis independently from its training-set marginal, so no real record's full conditioning tuple ever reappears.

Filtering and relabelling. A held-out classifier judges the generated samples. `qfilter_matched` oversamples to 16,000 then keeps the top-half by global confidence; `qstrat_matched` is the per-class-stratified version that removes the resulting prevalence drift. `iter_relabel` keeps all 8000 samples but replaces each intended label with the classifier's binarised prediction before downstream training, a self-distillation step.

Diversity and prototype. `diversity_oversample` generates a 16,000-sample pool and keeps the 8000 most diverse in the judge's 71-dim logit feature space, by greedy farthest-point selection. `prototype` replaces every training record's conditioning tuple with the per-class demographic centroid, generating from a class-typical prototype rather than from a real record.

5. Results

Table 1 reports each policy's TSTR macro AUROC with prevalence-bucket breakdowns; Figure 1 plots the same numbers with baseline and ceiling drawn for reference. The results sort cleanly by failure mode.

Class rebalancing. `uniform_class` loses macro, concentrated on the rare bucket the rebalancing was supposed to help. `rare_oversample_3x` repeats the shape: rare-

Table 1. Sampling policies for synthetic-only TSTR on PTB-XL. All rows generate $n_{\text{train}} = 8000$ synthetic 12-lead ECGs from a single shared generator, train an XResNet1d-50 classifier for 30 epochs on the synthetic corpus, and report macro AUROC on real PTB-XL fold-10 test ($n_{\text{test}} = 2198$), averaged across the indicated number of seeds. Buckets follow real training prevalence: common ($\geq 5\%$, 8 classes), moderate (2–5%, 7 classes), rare ($< 2\%$, 37 classes). `real_subset` is an XResNet1d-50 trained on a same-size random subset of the real training set and is the practical upper bound at $n = 8000$. `real_matched` is the naive bootstrap baseline. None of the synth-only policies recover `real_matched`'s macro; `qstrat_matched` ties within seed noise at one seed. **Bold** entries mark the strongest synth-only result in each column; the baseline row is bold by convention.

Group	Policy	Macro	Common	Moderate	Rare	seeds
Reference	`real_subset` (ceiling)	0.9070	0.9345	0.9230	0.8978	1
Reference	`real_matched` (baseline)	**0.8845±0.012**	**0.8823**	**0.8748**	**0.8869**	5
Class	`uniform_class`	0.8592±0.006	0.8766	0.8586	0.8555	2
Class	`rare_boost` (5% floor)	0.8816±0.003	0.8882	0.8748	0.8815	2
Class	`rare_oversample_3x`	0.8697±0.001	0.8848	0.8535	0.8695	3
Demographic	`demo_strat`	0.8676±0.005	0.8878	0.8647	0.8636	2
Demographic	`demo_balanced` (50/50 sex)	0.8712±0.005	0.8814	0.8693	0.8693	3
Demographic	`synth_cond` (decoupled demos)	0.8666±0.008	0.8767	0.8712	0.8634	3
Filter	`qfilter_matched` (global)	0.8376	0.8896	0.8655	0.8207	1
Filter	`qstrat_matched` (per-class)	0.8868	0.8979	0.8717	0.8873	1
Filter	`iter_relabel`	0.7907±0.033	0.9027	0.8760	0.7492	3
Diversity	`diversity_oversample`	0.8814±0.009	0.8868	0.8662	0.8831	3
Diversity	`prototype` (class centroid)	0.7892±0.003	0.7603	0.8200	0.7896	2

bucket AUROC drops despite each rare class being trained on roughly three times as many synthetic samples. The intuition imported from a fixed dataset (oversample the minority, gain on the minority) does not survive the move to a generator: each new sample for a rare class is drawn from the same conditional distribution as the last, so multiplying the rare cohort multiplies its *noise* along with its size. Only `rare_boost`, the mildest variant, ties the baseline within seed noise.

Filtering and relabelling. `qfilter_matched` (global top-50% by classifier confidence) is among the worst macro performers, and the loss is almost entirely on the rare bucket where the judge's probabilities are noisiest. The per-class-stratified `qstrat_matched` removes the prevalence drift and ends narrowly above the baseline at one seed, within seed noise. The clearest illustration is `iter_relabel`: at three seeds the common-class AUROC *rises* ($+0.020$) while rare-class AUROC *collapses* (-0.138) and seed variance triples. The mechanism is shared: any classifier-driven judge has well-calibrated probabilities on common classes and noisy ones on rare classes, so using it to filter, restratify, or relabel acts as a within-class diversity squeeze on the rare bucket specifically.

Demographic decoupling. `synth_cond` preserves both label prevalence and label co-occurrence; only the label-by-demographic joint is broken. It loses about as much macro as `uniform_class` despite changing nothing about class structure. `demo_balanced` (sex $50/50$, age uniform across deciles) and `demo_strat` (age-tertile $\times$ sex grid) land in the same band. Generator fidelity depends on conditioning-tuple combinations that actually occur in training; combinations the generator did not see are drawn from a weaker region of its conditional, even when the marginals look right.

Diversity and prototype. `prototype`, which replaces every record's conditioning tuple with the per-class demographic centroid, suffers the largest macro loss in the survey. The degradation is heaviest on the common bucket, the opposite of a pure within-class collapse, consistent with the class-mean centroid itself being off-distribution. `diversity_oversample`, which keeps the empirical class distribution but selects by farthest-point distance in classifier feature space, ties the baseline within seed noise on every bucket.

6. Synthetic + real reference

The policies above live in the synthetic-only setting motivated by the paper. As a sanity check, we also test the complementary regime in which the practitioner can mix real records into the downstream training set (Figure 2). A $50/50$ blend of 4000 real and 4000 synthetic records reaches macro TSTR 0.908 ± 0.011 at three seeds, statistically indistinguishable from the same-size pure-real ceiling. The synthetic corpus is therefore additive to real data rather than a noise floor. In the synthetic-only regime, however, this additivity is not something the practitioner can recover through a sampling policy.

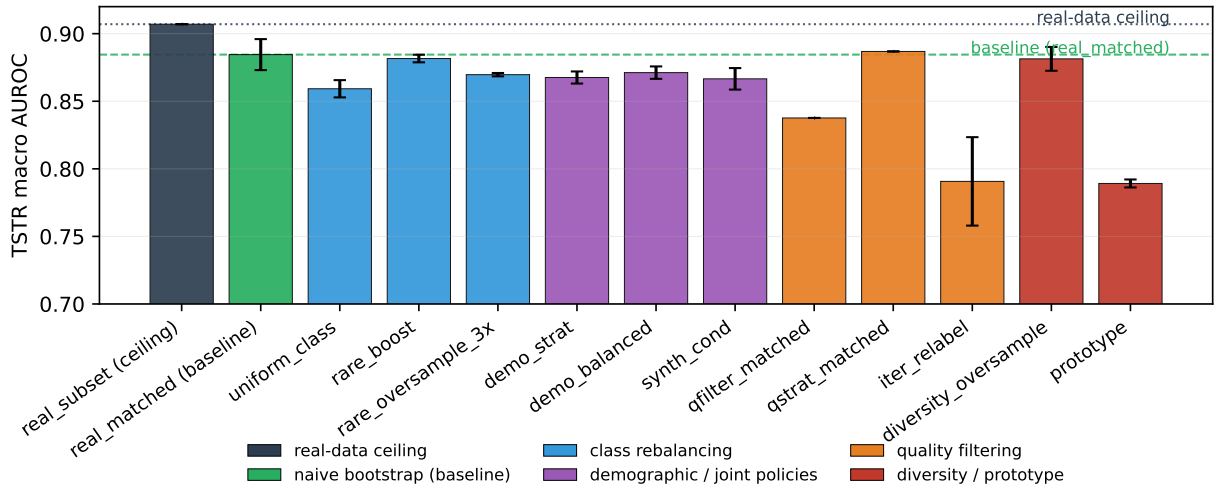

Figure 1. Macro TSTR by sampling policy, $n = 8000$ synthetic training samples, 30-epoch XResNet1d-50 downstream classifier. Color-coded by failure-mode group from Section 5. Dashed line: `real_matched` naive bootstrap baseline (0.8845 ± 0.012, 5 seeds). Dotted line: `real_subset` same-size real-data ceiling (0.907). Error bars are $1\,\sigma$ across seeds; the seed count varies per bar from 1 to 5 (see Table 1). `qfilter_matched` and `qstrat_matched` are one-seed bars and are shown without error bars; the bracket above `qstrat_matched` indicates a single-seed result that nominally exceeds the baseline by 0.0023 but has no measured variance.

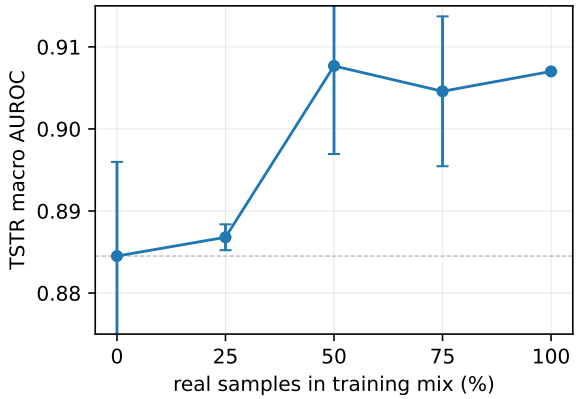

Figure 2. Macro TSTR vs. real fraction in a real+synth mix at fixed total $n = 8000$. Error bars are 1σ across 3 seeds (5 at the pure-synth 0% point). Dashed line: pure-synth baseline. The blend saturates the same-size real-data ceiling at 50/50.

7. Discussion

The headline result, that no synthetic-only policy beats the naive bootstrap, runs against the CV/NLP rebalancing literature, where downstream macro is routinely improved by oversampling minorities or generating synthetic minority neighbours (Chawla et al., 2002; Mariani et al., 2018; Dablain et al., 2022; Ye-Bin et al., 2025). There the underlying samples are real but minority-scarce, so resampling, interpolating, or topping up with extra generated examples increases the classifier's exposure to rare classes without disturbing where those samples sit in input space. In our setting the samples are *entirely* generated by a single model whose per-class fidelity is itself non-uniform: the rare classes the practitioner most wants to oversample are the classes where the generator is weakest, and oversampling them multiplies generator noise rather than recovering missing real support. The positive recipe is therefore the opposite of the CV/NLP one: track the empirical training distribution at sample time. Do not reweight per-class prevalence, do not filter by classifier signal, and do not synthesise conditioning tuples that did not appear in training. The naive bootstrap does all three by construction.

Limitations. We evaluate on a single dataset, generator family, and downstream architecture. The failure-mode taxonomy should generalise because each mechanism has a clean information-theoretic reading, but the deltas are dataset-specific. Where partial real access is available, the synth+real blend in Section 6 is the appropriate baseline.

8. Conclusion

On synthetic-only PTB-XL TSTR, eleven of twelve alternative sampling policies underperform the naive bootstrap at matched budget; the twelfth (`qstrat_matched`) ties within seed noise at a single seed and is the clearest follow-up. The failures collapse onto three mechanisms: class-distribution distortion, within-class diversity collapse, and label-by-demographic decoupling. A real+synth blend saturates the same-size real-data ceiling, but this additivity is out of reach for the synthetic-only practitioner. The recipe inverts the CV/NLP one: track the empirical training distribution at sample time, do not try to repair it.

References

Alcaraz, J. M. L. and Strodthoff, N. Diffusion-based conditional ECG generation with structured state space models. *Computers in Biology and Medicine*, 163:107115, 2022.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Dablain, D., Krawczyk, B., and Chawla, N. V. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):6390–6404, 2022.

Ghalebikesabi, S., Berrada, L., Gowal, S., Ktena, I., Stanforth, R., Hayes, J., De, S., Smith, S. L., Wiles, O., and Balle, B. Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*, 2023.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. *arXiv preprint arXiv:1812.01187*, 2019. We use the 1-D adaptation (XResNet1d-50) standard in PTB-XL benchmarks.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ibrahim, H. et al. Enabling granular subgroup-level model evaluations by generating synthetic medical time series. *arXiv preprint arXiv:2510.19728*, 2025.

Jordon, J., Yoon, J., and van der Schaar, M. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations (ICLR)*, 2019.

Lai, Y. et al. DiffuSETS: 12-lead ECG generation conditioned on clinical text reports and patient-specific information. *Patterns*, 2025.

Li, X., Liu, H., et al. A comprehensive survey on imbalanced data learning. *arXiv preprint arXiv:2502.08960*, 2024.

Liu, B. et al. A comprehensive survey of synthetic tabular data generation. *arXiv preprint arXiv:2504.16506*, 2024.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. BAGAN: Data augmentation with balancing GAN. In *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

Skorik, S. and Avetisyan, G. SSSD-ECG-nle: New label embeddings with structured state-space models for ECG generation. *arXiv preprint arXiv:2407.11108*, 2024.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Strodthoff, N., Wagner, P., Schaeffter, T., and Samek, W. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021.

Thambawita, V., Isaksen, J. L., Hicks, S. A., Ghouse, J., Ahlberg, G., Linneberg, A., Grarup, N., Ellervik, C., Olesen, M. S., Hansen, T., Graff, C., Holstein-Rathlou, N.-H., Strömberg, U., Andersen, S., Svensen, B., Maersk, M., Hansen, J., Kanters, J. K., Halvorsen, P., and Riegler, M. A. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Scientific Reports*, 11(1): 21896, 2021.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., and Schaeffter, T. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020.

Ye-Bin, M., Hyun, N., Ju, J., Kim, H. G., and Oh, T.-H. SYNAuG: Exploiting synthetic data for data imbalance problems. *Pattern Recognition Letters*, 193:115–122, 2025.