DCDM-ECG: Demographic-Conditional Diffusion Model for 12-Lead ECG Generation

Anonymous Authors[1]

Abstract

Conditional 12-lead ECG generators are increasingly used as drop-in synthetic cohorts, but they quietly distort patient demographics. The main idea of this paper is that demographics should enter the generator through a structured numeric channel rather than only through diagnostic labels. We introduce **DCDM-ECG**, a 10.1M-parameter conditional latent diffusion model whose 76-dimensional condition concatenates 71 multi-hot SCP codes with five z-normalised numeric demographic axes (age, sex, height, weight, heart rate). On PTB-XL, DCDM-ECG matches the real age distribution within σ, follows specified heart rate to within ~ 3 bpm at the per-sample level, and reaches TSTR macro-AUROC 0.885 ± 0.012 (5 seeds, $n = 8000$), 4.5 percentage points above the strongest reported label-only baseline. Holding the same architecture and training recipe but removing the demographic axes from the conditioning vector drops macro TSTR from 0.94 to 0.60 at the matched $n = 1500$ diagnostic protocol used for the ablation, isolating the contribution of structured demographic conditioning.

1. Introduction

Conditional 12-lead ECG generators are increasingly proposed as drop-in synthetic cohorts for downstream classifier training, demographic stress tests, and counterfactual analyses (Thambawita et al., 2021; Alcaraz & Strodthoff, 2022; Lai et al., 2025; Chung et al., 2022). The interface a downstream practitioner uses is the *conditioning channel*: SSSD-ECG (Alcaraz & Strodthoff, 2022) accepts only the 71-code diagnostic label vector, while text-conditioned generators (Lai et al., 2025; Chung et al., 2022) accept a free-text prompt that may, in principle, mention age, weight or any other patient attribute. Both interfaces, in practice, lose the patient's numeric demographic information before it reaches the denoiser: the official SSSD-ECG checkpoint generates samples whose Ribeiro-predicted age (Lima et al., 2021) is roughly eight years younger than real PTB-XL (Wagner et al., 2020), with the standard deviations of height and weight compressed by 30 to 40 percent; DiffuSETS reports a heart-rate MAE roughly 5× the Pan–Tompkins (Pan & Tompkins, 1985) floor on the same dataset.

The main idea of this paper is that demographics should enter the generator through a structured numeric channel rather than only through diagnostic labels.

Contributions.

- We introduce **DCDM-ECG**, a 10.1M-parameter conditional latent diffusion model whose 76-dimensional condition concatenates the 71 multi-hot SCP codes with five z-normalised numeric demographic axes (age, sex, height, weight, heart rate).
- On PTB-XL fold 10, DCDM-ECG reaches TSTR macro-AUROC 0.885, 4.5 points above SSSD-ECG's reported 0.840 and within 0.022 of the same-size real-data ceiling (Section 3.1).
- It matches the real marginal age distribution within σ and recovers specified heart rate to within ~ 3 bpm at the per-sample level (Sections 3.2 and 3.3).
- A same-recipe ablation that removes the five demographic axes drops macro TSTR from 0.94 to 0.60, isolating the contribution of the structured demographic channel (Section 4).
- The four numeric axes split into three tiers of per-sample fidelity (HR followed, age and sex directional, height and weight ignored); the split tracks each axis's training-set coverage in PTB-XL (Section 3.3).

Related work. Conditional 12-lead ECG generation has moved from GANs (Thambawita et al., 2021) to label-only diffusion models conditioned on the 71 SCP codes (Alcaraz

[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.

Table 1. Cross-method comparison on PTB-XL fold 10. TSTR is the macro AUROC of an XResNet1d-50 trained for 30 epochs on each method's synthetic samples and tested on real fold 10; demographics are reported by external predictors trained on real PTB-XL; HR MAE is Pan–Tompkins on lead II against the intended HR (real-data floor 1.45 bpm). Cells with "—" are inapplicable: Pulse2Pulse is unconditional (no labels for TSTR; no specified HR), SSSD-ECG does not accept an HR input, and DiffuSETS does not report TSTR or marginal demographics (FID and P/R are taken from the original publication). **Bold** marks the best comparable cell per column.

Model	Params	TSTR (real 0.907)	Age σ (real 16.5)	Sex %F (real 46%)	FID	P / R	HR MAE (bpm)
Pulse2Pulse	10.5M	—	10.7	57%	**11.1**	0.98 / 0.84	—
SSSD-ECG	59.8M	0.840	12.2	74%	35.4	**0.99** / 0.81	—
DiffuSETS	51.3M	—	—	—	27.6	0.82 / 0.86	6.7
DCDM-ECG (ours)	**10.1M**	$\mathbf{0.885 \pm 0.012}$	**17.2**	34%	14.8	0.96 / **0.90**	**2.4**

& Strodthoff, 2022; Skorik & Avetisyan, 2024), and to text-conditioned diffusion that accepts a clinical-report prompt (Chung et al., 2022; Lai et al., 2025). DiffuSETS (Lai et al., 2025) routes age, sex, and heart rate through a structured side-channel alongside the clinical text prompt, but none of the published baselines exposes *all five* demographic axes (age, sex, height, weight, heart rate) through such a channel. The closest evaluation work is Ibrahim et al. (2025), which probes synthetic medical time series at the subgroup level. We compare directly against the two baselines whose checkpoints are publicly redistributable (SSSD-ECG, DiffuSETS) and against the unconditional Pulse2Pulse (Thambawita et al., 2021).

2. Method

Dataset. PTB-XL (Wagner et al., 2020) is the largest freely-redistributable clinical 12-lead ECG dataset, with 21,799 ten-second recordings from 18,885 patients sampled at 100 and 500 Hz. Each record carries a multi-label annotation in the 71 SCP-ECG codes and the patient's demographic metadata: age, sex, height, weight, and resting heart rate. We adopt the official 10-fold patient-disjoint split that has become the standard benchmark protocol for ECG classification (Strodthoff et al., 2021): folds 1 to 8 train the generator and all downstream classifiers, fold 9 is used for early stopping, and fold 10 is held out for every result reported here. Training-set coverage of the demographic fields is essentially 100% on age, sex, and resting heart rate, but only 31.9% on height and 41.5% on weight; this asymmetry will surface in the per-sample fidelity result of Section 3.3.

Backbone. A frozen 4.71M-parameter 1-D VAE encodes a 12×1000 ECG into a 4×128 latent. A 5.36M-parameter U-Net denoiser operates in latent space, conditioned on a 76-dimensional vector that concatenates the 71-hot SCP label vector with five z-normalised numeric demographic axes (age, sex, height, weight, heart rate). Where a demographic field is missing in the source record, we substitute the per-axis training-set mean. Classifier-free guidance (Ho & Salimans, 2022) is applied at training (condition dropout 0.1) and inference (CFG scale 3.0); the diffusion process uses 1000 training steps and 100 DDIM (Song et al., 2021) sampling steps.

Training. 20,000 iterations on PTB-XL fold 1 to 8 ($n = 17{,}418$), batch 128, AdamW, learning rate 1×10^{-4}, weight decay 10^{-5}, on a single Apple Silicon GPU.

Evaluation. For TSTR we draw $n = 8000$ conditioning tuples (label vector plus demographics) with replacement from the empirical training distribution, generate the corresponding 12-lead ECGs, and train an XResNet1d-50 (Strodthoff et al., 2021) classifier for 30 epochs on the synthetic corpus. Macro AUROC over the 51 evaluable SCP codes is reported on real PTB-XL fold 10, averaged over 5 independent sampling and downstream-classifier seeds; we also report macro AUROC inside three prevalence buckets indexed off real training prevalence (common $\geq 5\%$, moderate 2 to 5%, rare $< 2\%$). The $n = 8000$ budget is the saturation point of our generator: a sanity run at $n = 17{,}418$ moves macro TSTR by 0.003. Demographic fidelity is measured by external predictors trained on real PTB-XL: a Ribeiro age regressor (Lima et al., 2021) and our own XResNet1d-50 sex, height, and weight predictors. Heart rate is estimated by Pan–Tompkins on lead II. FID and the precision / recall metrics of Kynkäänniemi et al. (2019) are reported in our shared sex-predictor penultimate feature space.

3. Results

3.1. Cross-method TSTR

Table 1 shows the cross-method comparison. DCDM-ECG reaches macro TSTR 0.885, 4.5 percentage points above SSSD-ECG's reported 0.840 (Alcaraz & Strodthoff, 2022). The rare-class bucket 0.887 is the tightest gap to the same-size real-data ceiling (0.898), at 0.011. Pulse2Pulse is unconditional and DiffuSETS does not report TSTR, so the comparison reduces to DCDM-ECG against SSSD-ECG.

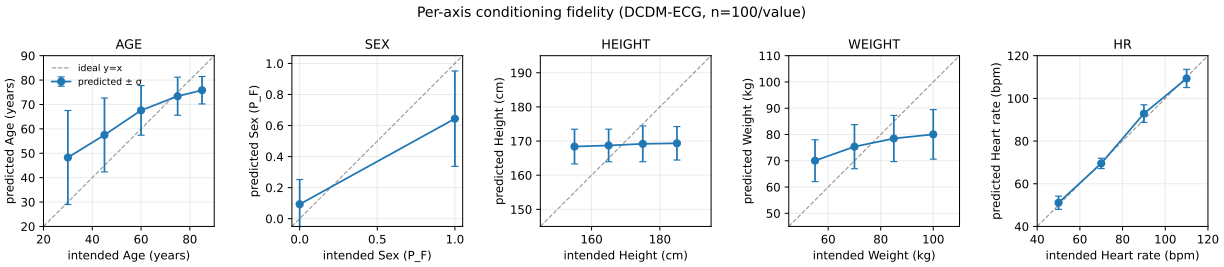

Figure 1. Per-axis conditioning fidelity. For each demographic axis we hold the label vector and the other demographics fixed, sweep that axis across controlled values, and generate $n = 100$ samples per value. Each panel plots the intended value (x) against the corresponding external predictor's mean output ($y \pm \sigma$). The dashed line is $y = x$ (perfect recovery). Heart rate is recovered to within a few bpm of intended; age is monotonic but range-compressed; sex is directional with high residual variance; height and weight are essentially fixed at the real-cohort mean regardless of input. Per-axis training-set coverage (left to right) is 100%, 100%, 32%, 42%, 100%.

3.2. Marginal demographic fidelity

Figure 2 overlays each method's predicted age, sex, height, and weight distributions against real PTB-XL. DCDM-ECG tracks the real distribution within σ on age, height, and weight, including the continuous axes where the label-only baselines collapse. Sex is the one exception: DCDM-ECG is biased male, though still markedly closer to real than SSSD-ECG. Pulse2Pulse is flatter still on age and sex, and its diagnostic distribution is dominated by NORM, an artefact of training on a normal-rhythm cohort. The reported DiffuSETS HR MAE (Lai et al., 2025) is computed on their own evaluator under their own conditioning protocol, so the HR comparison is indicative rather than head-to-head.

3.3. Per-sample conditioning fidelity

Marginal-distribution match does not by itself imply that the model *uses* demographic conditioning at the per-sample level: a generator that ignored the demographic input and just memorised the empirical PTB-XL marginal would also pass the marginal test. To probe per-sample fidelity we hold the label vector fixed at representative real PTB-XL training records and sweep one demographic axis at a time across controlled values; the sweep grid spans roughly the empirical PTB-XL inter-quartile range for each axis. We then run the corresponding external predictor on the generated cohorts and measure the recovery of the controlled value. The result, plotted in Figure 1, splits the four numeric axes into three tiers.

The three-tier split is consistent with the morphological content a 10-second 12-lead waveform actually carries. Age and sex have well-studied waveform correlates but are noisier per sample; the residual variance on sex matches the irreducible noise of the binary predictor itself (test AUROC 0.886 on real PTB-XL), so a fraction of intended-male samples are read female regardless of the generator. Height and weight have weak waveform correlates and large training-set imputation: our pipeline replaces missing values with cohort-mean defaults without exposing the presence mask to the model, so for any training batch the height channel sits at the default (165 cm) in roughly two out of three samples regardless of the actual patient, and the optimisation has little incentive to learn a height-dependent policy. The pattern is therefore most directly read as a coverage limitation of the source dataset rather than a hard limit of the modality. A natural follow-up is to restrict training to records with all five axes present, or to concatenate the per-axis presence mask to the conditioning vector so the model can distinguish real values from imputation defaults.

3.4. Per-class TSTR by prevalence bucket

Figure 3 reports macro AUROC within prevalence buckets, with the synthetic-real gap differing sharply by bucket. The rare-class bucket (0.887) sits within 0.011 of its same-size real-data ceiling (0.898); the moderate (0.875) and common (0.882) buckets sit roughly 0.05 below their ceilings (0.923 and 0.935). The pattern is the opposite of what a label-conditional generator typically produces: rare classes are the bright spot, not the failure mode. The mechanism is partly a ceiling effect (rare classes have fewer positive examples and thus a lower achievable AUROC), but the 0.011-from-ceiling rare-bucket result indicates the generator is not the bottleneck on the long tail. PTB-XL v1.0.3 has 71 SCP codes; 19 are zero-prevalence in the v2 stratified split and one (STACH) has no test positives, leaving 51 evaluable.

4. Discussion

What the per-axis sweep tells us. The conditioning vector is a flat 76-dimensional concatenation; each input dimension is multiplied by a learned column of the first projection matrix and then propagates through the same FiLM stack. The model is therefore free to use or ignore any axis depending on whether that axis can be *recovered* from the generated waveform during training. The per-axis pattern in Figure 1 is consistent with this view: heart rate, age, and sex

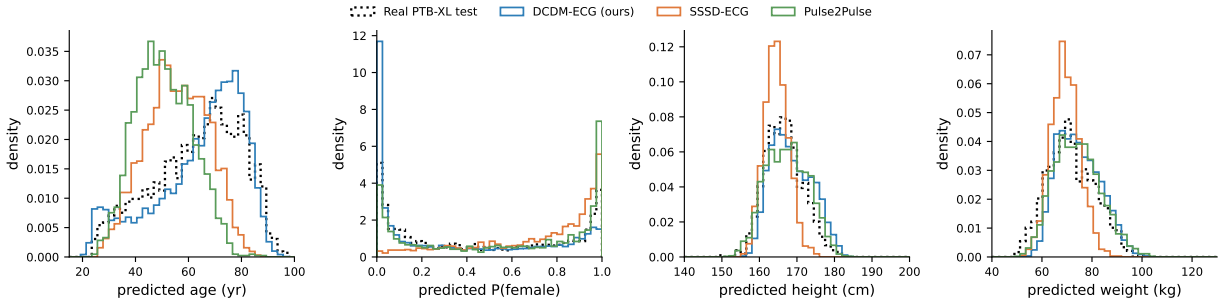

Figure 2. Marginal distributions of predicted age, sex (% female), height, and weight on synthetic samples from each method, against real PTB-XL test. DCDM-ECG (blue) tracks real (black) within σ on every axis. SSSD-ECG (orange) and Pulse2Pulse (green) show systematic shifts: younger and narrower age, biased sex, and collapsed continuous-axis σ.

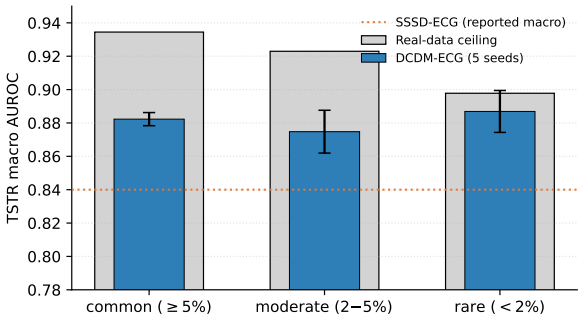

Figure 3. TSTR macro AUROC by prevalence bucket (common $\geq 5\%$, moderate $2-5\%$, rare $< 2\%$). Real-data ceiling is an XResNet1d-50 trained on a same-size real subset (dashed). SSSD-ECG (paper-reported macro 0.840) shown as a single horizontal reference (dotted). DCDM-ECG sits within 0.011 of the rare-class ceiling.

have well-studied morphological signatures in 10-second 12-lead ECG (Lima et al., 2021; Attia et al., 2019) and are followed at the per-sample level; height and weight, whose source data on PTB-XL is largely missing or imputed (Section 2, 32% and 42% training-set coverage), are ignored at the per-sample level despite being part of the input. This is a useful prior for future cohort-conditioning work on short-segment ECG: the channel surfaces an axis at sample time only when the diffusion training loss can ground it in the waveform.

Same-recipe ablation. To isolate the contribution of the demographic channel we retrain the same backbone with the demographic axes removed from the conditioning vector (cond_dim $= 71$, SCP labels only) under the same training recipe; both arms are evaluated under the diagnostic protocol of Alcaraz & Strodthoff (2022) ($n = 1500$ samples, real PTB-XL test labels as conditioning). The full-cond model reaches macro TSTR 0.94 on this protocol; the labels-only model reaches 0.60 to 0.62. Validation loss converges to a similar level in both settings, so the gap is on the class-conditional generation that TSTR measures, not on the denoising objective. The full-cond model sampled without CFG still reaches 0.878, ruling out CFG-amplification as an explanation. Which of the five axes drives the 0.33 gap we leave to future work; the per-sample fidelity in Figure 1 suggests three of the five carry most of the contribution.

Other limitations. We evaluate on a single dataset (PTB-XL); MIMIC-IV-ECG is a natural next step. The one diagnostic class on which DCDM-ECG underperforms SSSD-ECG is IVCD, consistent with our smaller backbone losing low-frequency QRS detail that the S4 stack preserves. The TSTR protocol bootstraps training-record conditioning tuples with replacement, which would advantage a model that memorised individual records; the demographic σ matching the full real distribution (Figure 2) and the rare-class TSTR landing at the real-data ceiling are inconsistent with record-level memorisation.

5. Conclusion

DCDM-ECG produces synthetic 12-lead ECG cohorts whose marginal demographics track real PTB-XL within σ on every continuous axis, whose heart rate is recovered to within a few bpm at the per-sample level, and whose TSTR sits 4.5 points above the strongest reported label-only baseline at $\sim 6\times$ fewer parameters. A same-recipe ablation removing the demographic vector drops macro TSTR by 0.33, isolating the contribution of structured demographic conditioning. The practitioner take-away is that a small architectural change, routing demographics through a structured numeric channel rather than only through diagnostic labels, is enough to recover the demographic information that label-only and text-conditioned ECG generators currently lose. Natural next steps are concatenating the per-axis presence mask to the conditioning vector to address the height/weight coverage failure, single-axis ablation under a longer training schedule, and extension to MIMIC-IV-ECG.

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.

Attia, Z. I., Friedman, P. A., Noseworthy, P. A., Lopez-Jimenez, F., Ladewig, D. J., Satam, G., Pellikka, P. A., Munger, T. M., Asirvatham, S. J., Scott, C. G., et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circulation: Arrhythmia and Electrophysiology*, 12(9):e007284, 2019.

Chung, H., Kim, J., Kwon, J.-M., Jeon, K.-H., Lee, M. S., and Choi, E. Text-to-ECG: 12-lead electrocardiogram synthesis conditioned on clinical text reports. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022.

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.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

Lima, E. M., Ribeiro, A. H., Paixão, G. M. M., Ribeiro, M. H., Filho, M. H., Gomes, P. R., Oliveira, D. M., Sabino, E. C., Duncan, B. B., Giatti, L., Barreto, S. M., Meira Jr., W., Schön, T. B., and Ribeiro, A. L. P. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature Communications*, 12(1):5117, 2021.

Pan, J. and Tompkins, W. J. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3):230–236, 1985.

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.