Initial Development of a Twelve-Lead Electrocardiogram Model for Digoxin Exposure Detection as a First Step Toward Toxicity Stratification

Jose Moon[1], Jong-Ho Kim[1], Hyung Joon Joo[2]

[1]Korea University Research Institute for Medical Bigdata Science, Korea University College of Medicine, Seoul, Republic of Korea
[2]Division of Cardiology, Department of Internal Medicine, Korea University Anam Hospital, Seoul, Republic of Korea

Abstract

Digoxin produces electrocardiographic effects, but disease-related abnormalities may obscure exposure- and concentration-related signals. We trained ECGResNet1D models to estimate a continuous digoxin target from twelve-lead electrocardiograms, assigning 0 to non-exposed controls and using measured serum concentrations for exposed ECGs. Patients from Hospitals A and B were pooled and assigned without patient overlap to training, validation, and internal-test sets; Hospital C was reserved for external testing. The cohort included 20,397 ECGs from 14,745 patients. One scalar output was evaluated post hoc for no recorded digoxin exposure versus exposure and, among exposed ECGs, high (>=2.0 ng/mL) versus non-high concentration. In 5,434 external ECGs, exposure AUROC was 0.989 for ECG only and 0.986 with metadata; AUPRC was 0.894 and 0.886. External RMSE was 0.270 and 0.293 ng/mL. High-versus-non-high discrimination was modest: AUROC was 0.682 and 0.630, and AUPRC was 0.235 and 0.198. Patient-cluster bootstrap 95% confidence intervals for high-level AUROC were 0.586-0.785 and 0.525-0.739. Mixed-site development therefore yielded robust external exposure detection, while high-concentration discrimination showed moderate performance and wider uncertainty.

1. Introduction

Digoxin remains an option for long-term ventricular rate control in atrial fibrillation, particularly when other agents are not preferred, tolerated, or suitable and in patients with heart failure symptoms [1]. Its narrow therapeutic index makes serum concentration monitoring clinically important. The 2023 ACC/AHA/ACCP/HRS guideline states that, when serum digoxin measurement is indicated in atrial fibrillation, targeting a level <1.2 ng/mL is reasonable [1]. Current U.S. prescribing information notes that toxicity becomes more frequent above 2.0 ng/mL but may also occur at lower concentrations, depending on the patient's clinical condition, renal function, electrolyte status, and concomitant treatment [2]. Therefore, serum concentration provides clinically relevant risk information but does not independently establish a diagnosis of digoxin toxicity.

Digoxin toxicity may combine increased automaticity with impaired conduction and can manifest as bradyarrhythmias, atrial tachycardia with block, junctional tachycardia, ventricular ectopy, or ventricular tachycardia; however, no single electrocardiographic abnormality is consistently present [3]. A previous single-center retrospective AI-assisted ECG study reported an internally validated AUROC of 0.912 for detecting digoxin toxicity defined by a serum concentration ≥2.0 ng/mL [4]. However, that study compared high-concentration cases with normal-ECG controls and did not include external validation. Consequently, its performance may have reflected both general digoxin-exposure effects and concentration-related changes, rather than information specifically associated with the high-concentration range. Performance also differed according to heart failure and atrial fibrillation status, indicating that underlying cardiac conditions may influence ECG-based predictions [4]. More broadly, deep neural networks can identify clinically relevant abnormalities directly from raw twelve-lead ECGs [5].

We therefore developed ECG-only and ECG-plus-metadata versions of a single ECGResNet1D regression model and evaluated the continuous output for two sequential endpoints: no recorded digoxin exposure versus exposure, followed by high versus non-high concentration among exposed ECGs. The 2.0-ng/mL threshold was used as an analytical reference because toxicity is reported more frequently above this level, not as a direct diagnostic definition of toxicity. Patients from two hospitals were pooled across patient-disjoint development partitions, while a third hospital was

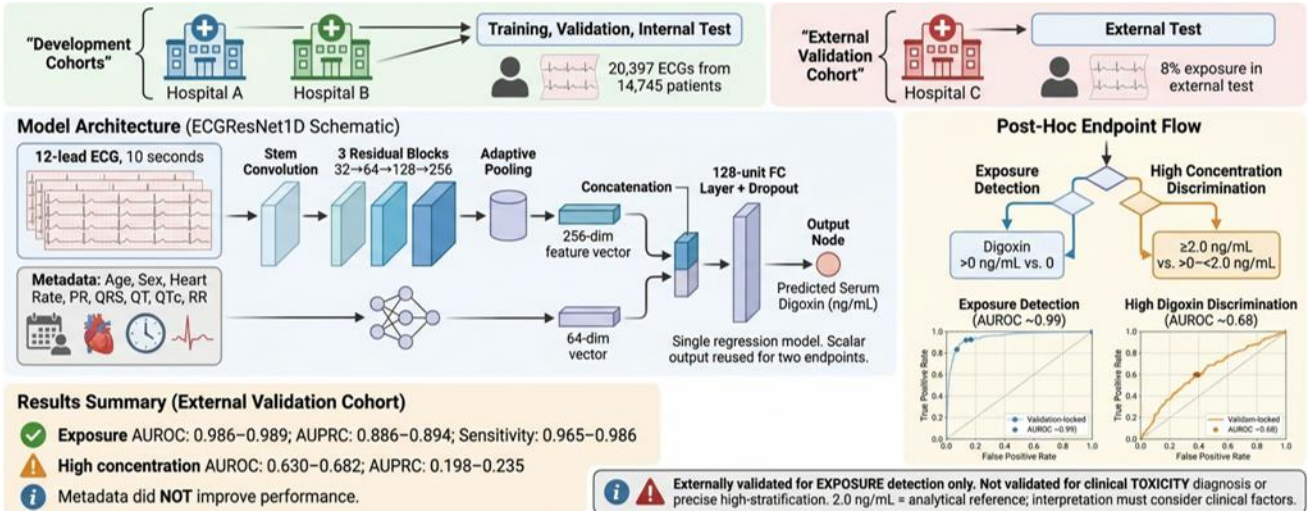

Figure 1. Multisite study design, ECGResNet1D architecture, and evaluation framework.

retained exclusively for external evaluation. This design allowed us to assess whether ECGs contain a transportable digoxin-exposure signal and whether the same output provides additional information related to the high-concentration range.

2. Methods

2.1. Study design and cohort

We retrospectively assembled digital twelve-lead, 10-s ECGs and associated serum digoxin values from three tertiary hospitals. Counts refer to ECGs, whereas cohort assignment and uncertainty estimation were performed at patient level. Concentrations are reported in ng/mL.

Non-exposed controls had no recorded digoxin exposure and were assigned a target value of 0; exposed ECGs had values >0. Patient identifiers were treated as hospital-independent. Fourteen identifiers appeared in both Hospital C and Hospitals A/B. To preserve Hospital C as an untouched external test, 411 corresponding A/B ECGs were excluded before splitting. The final cohort comprised 20,397 ECGs from 14,745 patients. Digital counts were converted to mV using the amplitude field, and nonstandard-length leads were spline-resampled to 5,000 samples.

2.2. One regression model and two evaluation endpoints

Each input configuration was trained only to regress the continuous digoxin target; non-exposed controls were assigned 0, and no separate binary classifier was trained. The same scalar prediction was evaluated post hoc for: (i) exposure, with non-exposed=0 and digoxin >0; and (ii) high concentration among exposed ECGs, with high defined as >=2.0 ng/mL and non-high as >0 to <2.0 ng/mL. The prespecified 2.0-ng/mL reference-label cut was motivated by prescribing information noting more frequent toxicity above this level, but it was neither the model operating cutoff nor a stand-alone toxicity diagnosis [1,2]. External analyses at 1.5 and 3.0 ng/mL tested sensitivity to this label definition.

2.3. ECGResNet1D architecture

Input tensors had 12 channels by 5,000 samples. A stem convolution (32 filters, kernel 7, stride 2) was followed by batch normalization, ReLU, and max pooling (kernel 3, stride 2). Three residual blocks increased channels from 32 to 64, 128, and 256 while downsampling with stride 2. Each block contained two kernel-7 convolutions with batch normalization and ReLU; a kernel-1 projection aligned the skip path when dimensions changed. Adaptive average pooling yielded a 256-dimensional ECG representation.

The ECG-plus-metadata model used age, sex, heart rate, PR interval, QRS duration, QT interval, corrected QT, and RR interval. Numeric missing values were median-imputed during frozen dataset construction and sex was numerically encoded. A linear layer projected the eight variables to 64 dimensions before concatenation with the ECG representation. The shared head comprised a 128-unit linear layer, ReLU, dropout 0.3, and one scalar output (Figure 2).

2.4. Training and site strategy

Patients from Hospitals A and B were jointly allocated to training, validation, and internal testing in prespecified proportions of 55.8%, 31.0%, and 13.3% (seed 42). Allocation was stratified by hospital and each patient's maximum concentration category (non-exposed, >0 to <2.0, or >=2.0 ng/mL); every ECG from a patient remained in one partition, and both hospitals were represented in each development partition. Hospital C was reserved for external testing. Models used mean

squared error, Adam (learning rate 0.001), batch size 64, and 40 epochs. To reproduce the prespecified protocol, final-epoch weights were evaluated; the minimum validation-RMSE epoch is reported descriptively, and no external-test information was used for model selection.

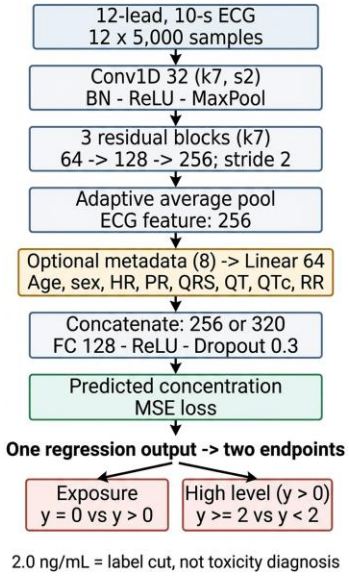

Figure 2. ECGResNet1D and endpoint flow. A single continuous regression output was reused for two post hoc evaluations. The 2.0-ng/mL reference-label cut is distinct from any model operating cutoff.

2.5. Evaluation

Continuous prediction was summarized by RMSE over all ECGs, including non-exposed controls assigned 0. AUROC and AUPRC were calculated for both post hoc endpoints. A prediction cutoff selected once in validation by Youden's index was applied unchanged to internal and external tests for sensitivity, specificity, and F1. Because high concentrations were uncommon, AUPRC was interpreted relative to prevalence. External 95% confidence intervals (CIs) used 2,000 percentile bootstrap replicates resampled by patient, thereby preserving repeated ECGs within clusters.

Table 1. Patient-disjoint mixed-site cohort; high denotes >=2.0 ng/mL.

Split	Site ECGs	Patients	ECGs	Exposed	High n (%)
Training	A 4,724; B 3,660	5,997	8,384	3,048	120 (3.9)
Validation	A 2,547; B 2,031	3,329	4,578	1,600	63 (3.9)
Internal	A 1,128; B 873	1,426	2,001	724	29 (4.0)
External	C 5,434	3,993	5,434	434	48 (11.1)

3. Results

3.1. Cohort

The mixed-site cohort contained 20,397 ECGs from 14,745 patients. Exposed ECGs comprised 36.4% of training, 35.0% of validation, 36.2% of internal testing, and 8.0% of external testing. Among exposed ECGs, 120/3,048 (3.9%), 63/1,600 (3.9%), 29/724 (4.0%), and 48/434 (11.1%), respectively, met the 2.0-ng/mL high-concentration definition (Table 1).

3.2. Continuous regression

Validation/internal/external RMSE was 0.338/0.373/0.270 ng/mL for ECG only and 0.354/0.375/0.293 ng/mL for ECG plus metadata (Table 2). Because non-exposed controls were assigned 0, these values reflect the combined exposure-plus-concentration regression task. External patient-bootstrap 95% CIs were 0.225-0.321 and 0.236-0.355 ng/mL, respectively; metadata did not improve overall error.

3.3. Post-hoc classification endpoints

Exposure AUROC/AUPRC remained high across validation/internal/external sets: 0.987/0.978, 0.989/0.980, and 0.989/0.894 for ECG only, and 0.982/0.973, 0.977/0.968, and 0.986/0.886 with metadata. With validation-locked cutoffs, external sensitivity/specificity was 0.965/0.921 and 0.986/0.892, respectively (Table 3; Figure 3A).

High-level discrimination was lower. Across validation/internal/external sets, ECG-only AUROC/AUPRC was 0.694/0.132, 0.628/0.083, and 0.682/0.235; ECG-plus-metadata values were 0.628/0.069, 0.714/0.126, and 0.630/0.198. At validation-locked cutoffs, external sensitivity/specificity was 0.479/0.780 and 0.708/0.505, respectively (Table 3; Figure 3B).

3.4. Sensitivity to the high-level definition

Among 434 exposed external ECGs, a 1.5-ng/mL cut yielded 95 high cases and AUROC/AUPRC of 0.690/0.353 for ECG only and 0.622/0.321 with metadata. At 3.0 ng/mL, only 18 high cases remained; AUROC/AUPRC was 0.600/0.130 and 0.468/0.065, respectively. At the 3.0-ng/mL threshold, estimates were imprecise because only 18 ECGs were available.

Table 2. Final-epoch continuous regression RMSE (ng/mL).

Model	Min-val epoch	Validation	Internal	External (95% CI)
ECG only	7	0.338	0.373	0.270 (0.225-0.321)
ECG+meta	7	0.354	0.375	0.293 (0.236-0.355)

Table 3. External performance with patient-bootstrap 95% CIs; cutoffs locked in validation.

Task / model	AUROC (95% CI)	AUPRC (95% CI)	Sens./Spec.
Exposure ECG	0.989 (0.985-0.992)	0.894 (0.852-0.929)	0.965/0.921
Exposure ECG+meta	0.986 (0.979-0.991)	0.886 (0.847-0.920)	0.986/0.892
High ECG	0.682 (0.586-0.785)	0.235 (0.146-0.398)	0.479/0.780
High ECG+meta	0.630 (0.525-0.739)	0.198 (0.120-0.345)	0.708/0.505

4. Discussion

Pooling patients from two hospitals across development partitions yielded reproducible digoxin-exposure detection, with Hospital C reserved for evaluation. Exposure AUROC remained 0.986-0.989 and AUPRC 0.886-0.894 despite exposure prevalence of 8.0%. Patient-level separation and cluster-based CIs reduced optimism from repeated ECGs, supporting cross-hospital transportability. High-concentration discrimination was lower: external AUROC was 0.682 for ECG only and 0.630 with metadata, with AUPRC values of 0.235 and 0.198 against a prevalence baseline of 0.111. Metadata did not improve external RMSE.

This difference may reflect the imperfect relationship between serum digoxin concentration and clinical toxicity. Toxicity is not determined by concentration alone and may vary with renal function, electrolytes, rhythm, and treatment. Nevertheless, prescribing information notes that toxicity is more frequent above 2.0 ng/mL [1,2]. We therefore used this threshold to examine whether ECGs contain signals associated with a high-concentration range. This was an initial concentration-stratification investigation rather than direct toxicity diagnosis. Variation at 1.5 and 3.0 ng/mL showed sensitivity to threshold definition and positive-case count.

Limitations include the retrospective design, few high-concentration ECGs, absence of exposed-only calibration, and lack of adjudicated toxicity outcomes. Overall RMSE included non-exposed controls assigned 0 and therefore does not represent concentration error among exposed patients alone. The external hospital also had lower exposure prevalence than the development cohorts.

Future work will investigate self-supervised pretraining on unlabeled ECG datasets, followed by task-specific fine-tuning in exposed patients. Class-balanced or ordinal objectives and integration of renal function, electrolytes, rhythm, and treatment information may improve concentration-sensitive prediction and toxicity-risk evaluation.

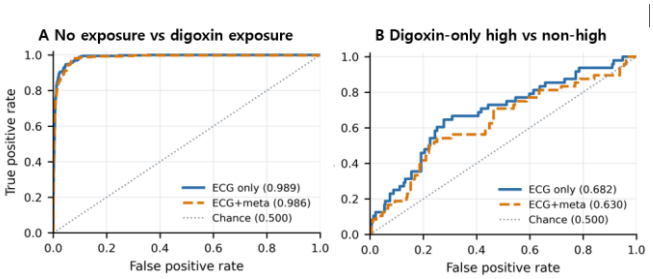

Figure 3. Hospital C external-test ROC curves after mixed-site patient-level development. (A) No recorded digoxin exposure versus digoxin exposure (N=5,434; exposed n=434). (B) Digoxin-only high versus non-high concentration at 2.0 ng/mL (n=434; high n=48).

5. Conclusion

After patient-level mixed-site development, ECGResNet1D achieved robust external detection of digoxin exposure. These findings establish a transportable ECG-based exposure signal across hospitals and provide a foundation for precise concentration and toxicity-risk modeling using self-supervised and clinically enriched approaches.

Acknowledgments

This work was supported by a Korea University grant (A920268078).

References

[1] Joglar JA, Chung MK, Armbruster AL, et al. 2023 ACC/AHA/ACCP/HRS guideline for the diagnosis and management of atrial fibrillation: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2024;149(1):e1-e156.

[2] Advanz Pharma (US) Corp. LANOXIN (digoxin) tablets prescribing information. Revised December 2024. Available from: https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=d91e3646-4c63-4512-ab22-db39c085c4dc. Accessed July 14, 2026.

[3] Ma G, Brady WJ, Pollack M, Chan TC. Electrocardiographic manifestations: digitalis toxicity. J Emerg Med. 2001;20(2):145-152. doi:10.1016/S0736-4679(00)00312-7.

[4] Chang DW, Lin CS, Tsao TP, et al. Detecting digoxin toxicity by artificial intelligence-assisted electrocardiography. Int J Environ Res Public Health. 2021;18(7):3839.

[5] Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun. 2020;11:1760.

Address for correspondence

Hyung Joon Joo
Division of Cardiology, Department of Internal Medicine, Korea University Anam Hospital, 73 Goryeodae-ro, Seongbuk-gu, Seoul 02841, Republic of Korea.
drjoohj@korea.ac.kr