# Prediction of Systolic Heart Failure from Real-World ECG Data

Lennart Graf[1], Gabriel Riedemann[2], Nicolai Spicher[3], Dagmar Krefting[1], Philip Hempel[1]

[1] Department of Medical Informatics, University Medical Center Göttingen, Germany
[2] Department of Cardiology and Pneumology, University Medical Center Göttingen, Germany
[3] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

## Abstract

***Aim:*** *Early identification of systolic heart failure (HF) from the electrocardiogram (ECG) could support scalable screening, but prior ECG models rely on labels derived from echocardiography (echo) that may be difficult to obtain at large scale. In this case study, we investigated whether systolic HF can be predicted from ECGs using weak labels derived from routine diagnosis codes.*

***Methods:*** *We used the Harvard-Emory ECG Database (HEEDB) for training and derived a patient-level binary target from routinely collected ICD diagnosis codes, yielding a development cohort of 1,625,032 ECGs. A supervised 12-lead ECG classification model based on a 1D ResNet architecture was externally validated on 5,442 patients from EchoNext using echo-derived labels.*

***Results:*** *Internal validation achieved an AUC of 0.962 (CI: 0.961–0.963) while external validation yielded an AUC of 0.797. Additionally, predicted probabilities showed an inverse relationship with continuous LVEF (Pearson's $r = -0.467$, $p < 0.001$).*

***Conclusions:*** *These findings suggest that ECG models for systolic HF can be developed using ICD-based labels from routine clinical data, with model output associated with echo-derived LVEF. However, the performance decrease in external evaluation underscores the importance of careful interpretation and validation against imaging-derived endpoints.*

## 1. Introduction

Heart failure (HF) is a global health problem, affecting 64 million people worldwide [1]. HF is associated with high mortality, with reported survival rates of 81.3% at 1 year, 51.5% at 5 years, and 29.5% at 10 years after diagnosis [2]. The electrocardiogram (ECG) is a widely available, low-cost diagnostic tool, making it attractive for large-scale screening. Although echocardiography (echo) remains the reference standard for assessing cardiac structure and function, its use at scale is constrained by the fact that image acquisition and interpretation are operator-dependent, require substantial expertise, and are time- and resource-intensive [3]. Established biomarkers such as NT-proBNP are clinically relevant but depend on thresholds and context, with reduced performance in early disease stages [4]. This highlights the need for scalable screening methods.

Left ventricular ejection fraction (LVEF) is a standard measure of left ventricular systolic function and reflects how effectively the left ventricle ejects blood with each heartbeat. Reduced LVEF is commonly used as an indicator of systolic dysfunction and is therefore clinically relevant for identifying patients with suspected systolic HF.

In this context, the ECG is well suited as an initial screening tool for further cardiac evaluation. Recent studies have shown that deep learning models can detect systolic dysfunction from ECGs when trained on datasets with explicit targets such as echo-derived LVEF [5, 6]. However, such approaches depend on high-quality labels, which are difficult to obtain at scale.

At the same time, the growing availability of large-scale clinical ECG repositories provides an important opportunity to develop and evaluate data-driven approaches in real-world settings [7]. One example is HEEDB, a large real-world ECG dataset that contains routinely collected diagnosis information rather than task-specific gold-standard labels [7,8]. While such diagnosis codes are readily available, they may not precisely capture phenotype severity or timing and can therefore introduce misclassification into the labeling process [9,10]. It therefore remains unclear whether ECG-based prediction of systolic HF can be achieved when training labels are derived from routine clinical coding rather than echo-based measurements.

In this case study, we evaluate whether a systematic ICD-based labeling strategy applied to HEEDB can support prediction of systolic HF and generalize to an external dataset with echo-derived LVEF labels.

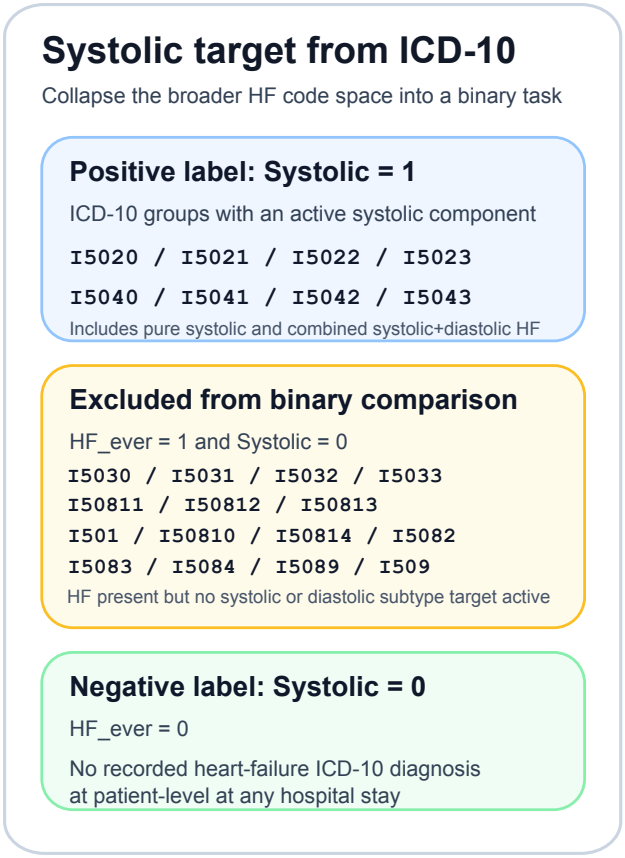


Figure 1. Overview: routinely collected HF diagnosis codes were mapped to a binary task: patients with codes indicating a systolic component were assigned a positive label, patients without any HF-related code were assigned a negative label, and patients with other HF codes but no systolic component were excluded from the binary comparison.

# 2. Methods

## 2.1. Training dataset and target definition

We used HEEDB as the training dataset [8, 11] with 11,440,211 12-lead ECGs and accompanying patient metadata and ICD diagnosis information from 2,167,794 patients. We restricted the cohort to adults aged 18–90 years and excluded patients with missing age or ICD information, resulting in 1,652,288 eligible patients.

The prediction target was defined at the patient level from routinely collected HF diagnoses as depicted in Fig 1. To account for legacy coding, ICD-9 heart failure codes (428.x) were used only to identify patients with any prior HF diagnosis and to exclude such patients from the negative cohort, whereas assignment of the systolic target was based on ICD-10 codes with explicit systolic subtypes. For patients in the positive cohort, the selected ECG was the first recording obtained after the initial systolic HF diagnosis. For patients in the negative cohort, one random ECG per patient was retained from the available recordings. This procedure yielded a final training set of 1,625,032 ECGs, including 22,423 positive and 1,602,609 negative cases.

## 2.2. External dataset

External validation was performed on the EchoNext dataset [6, 12], which contains paired 12-lead ECG recordings and echo-derived structural labels. To avoid repeated measurements and maintain a single-record design, only the most recent ECG per patient was included.

The external reference target for systolic dysfunction was defined from echo-derived LVEF:

- Positive: LVEF $\leq 45\%$
- Negative: LVEF $> 45\%$

An LVEF threshold of $\leq 45\%$ was used as an exploratory intermediate cutoff between the conventional HFrEF threshold ($\leq 40\%$) and preserved EF ($\geq 50\%$). After filtering, the external cohort comprised 5,442 ECGs from 5,442 patients.

All ECG signals in the external dataset were preprocessed consistently across leads using median filtering, clipping at the 0.1th and 99.9th percentiles, and normalization with dataset-wide mean and standard deviation.

## 2.3. Model development and performance evaluation

We trained a supervised 12-lead 1D ResNet using the architecture and training pipeline described previously [13], without task-specific architectural modifications. The model received 12-lead ECGs as channel-first input with 4,096 samples per lead at a sampling frequency of 400 Hz. The network comprised five sequential residual stages with 64, 128, 196, 256, and 320 filters, respectively, using a kernel size of 17 and progressively reducing the temporal resolution. Dropout with a rate of 0.8 was used for regularization. No additional ECG-quality-based exclusion or artifact rejection was applied. Internal model development and evaluation were performed on HEEDB using 5-fold cross-validation. For internal validation, predictions from the held-out fold were collected across all five folds and pooled to obtain out-of-fold predictions for the full internal cohort. External prediction was then performed using an ensemble of the five cross-validation models, by averaging their predicted probabilities.

Model discrimination was summarized using the area under the receiver operating characteristic curve (AUC). In addition, to reflect a screening-oriented operating point, we report accuracy, sensitivity, specificity, and F1-score at

Table 1. Model performance on internal (HEEDB) and external (EchoNext) validation at the operating point with target specificity of 95%. Values are reported with 95% confidence intervals.

| Metric | Internal | External |
|---|---|---|
| AUC | 0.962 (0.961–0.963) | 0.797 (0.780–0.813) |
| Accuracy | 0.948 (0.948–0.948) | 0.852 (0.843–0.861) |
| Sensitivity | 0.797 (0.792–0.803) | 0.390 (0.359–0.421) |
| Specificity | 0.950 (0.950–0.950) | 0.950 (0.945–0.957) |
| F1-score | 0.297 (0.294–0.301) | 0.482 (0.450–0.512) |

a decision threshold chosen to target 95% specificity. To assess the relationship between model output and echocardiographic systolic function, predicted probabilities in EchoNext were examined against continuous LVEF values.

# 3. Results

## 3.1. Internal validation

Across the 5-fold cross-validation on HEEDB, the model showed stable internal performance. The mean AUC was 0.962 (95% CI 0.961–0.963). At the operating point targeting 95% specificity, detailed performance metrics are reported in Table 1.

## 3.2. External validation

When transferred to EchoNext and evaluated against echo-derived LVEF, the model retained moderate discriminatory performance. The external AUC was 0.797 (95% CI 0.780–0.813). At the operating point targeting 95% specificity, detailed metrics are summarized in Table 1.

## 3.3. Prediction-LVEF relationship

Model output showed an inverse relationship with continuous LVEF values in the external cohort, with a moderate negative Pearson correlation ($r = -0.467$, $p < 0.001$) and a negative Spearman rank correlation ($\rho = -0.346$, $p < 0.001$). Higher predicted probabilities of systolic HF were observed predominantly at lower LVEF values, whereas lower model scores were more common among patients with higher LVEF.

# 4. Discussion

In this case study, we found that a systematic ICD-based labeling strategy applied to a large real-world ECG repository supported internal prediction of systolic HF and retained moderate external performance against echo-derived LVEF. These findings suggest that large-scale routine clinical data can support ECG model development even without gold-standard labels for the full training population.

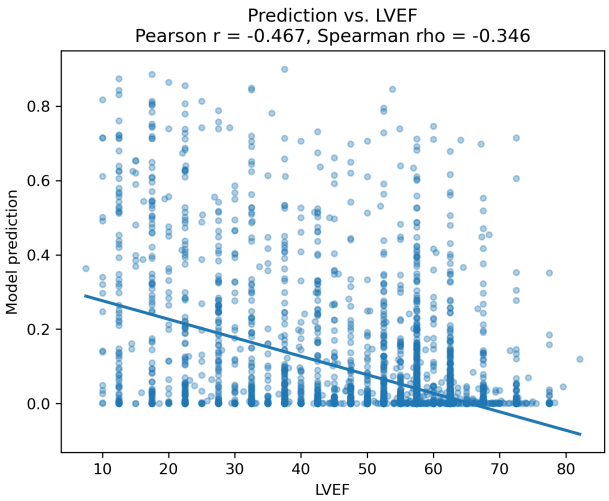


Figure 2. Model-predicted probability of systolic HF versus echo-derived left ventricular ejection fraction (LVEF) in the EchoNext cohort. Predicted probabilities increased as LVEF decreased, consistent with the expected relationship between impaired systolic function and reduced ejection fraction.

A central strength of this approach is its scalability. Echo-derived labels are clinically informative but difficult to obtain at the scale required for broad ECG model development. In contrast, diagnosis codes are widely available and permit construction of large training cohorts. Our results indicate that restricting codes to clear systolic labels and excluding ambiguous HF phenotypes still enables transfer to an external echocardiographic definition of systolic dysfunction.

The decrease in performance from internal to external validation is expected and likely reflects mismatches between development and evaluation. First, the training target was defined from routine ICD coding, whereas the external endpoint was defined from echo-derived LVEF. Second, differences in cohort composition, ECG preprocessing, and acquisition may have introduced domain shift. Third, systolic HF and reduced LVEF are related but not identical constructs; some patients with clinically coded systolic HF may not map cleanly to the chosen LVEF threshold, and vice versa.

A brief comparison with EchoNext provides context but should be interpreted cautiously. The published EchoNext model was trained on paired ECG–echo data for a broader task and reported stronger performance for the low-LVEF component (AUROC ∼0.90) [6]. By contrast, our model used weaker ICD-based labels and was evaluated against

LVEF in external validation. The lower external performance is therefore unsurprising and reflects the trade-off between label fidelity and training scale rather than a direct model comparison.

These findings suggest a practical development strategy for future work. Large routine datasets such as HEEDB may be valuable for pretraining or weakly supervised training, after which smaller datasets with echo-derived labels could be used for task-specific refinement [14]. This two-stage approach may combine the scale advantages of administrative labels with the clinical precision of imaging-based endpoints.

This study has several limitations. The training labels were derived from diagnosis codes and are therefore susceptible to misclassification, incomplete coding, and temporal imprecision. The external endpoint was defined using a pragmatic LVEF threshold of 45%, which is clinically useful but does not fully capture HF. Performance was not evaluated separately across demographic or clinical subgroups, and potential heterogeneity by age, sex, comorbidity, or cardiac rhythm therefore remains to be investigated. In addition, external validation was performed on a dataset with its own preprocessing and sampling framework, which may influence transportability.

Nevertheless, ECG-based prediction of systolic HF appears feasible using imperfect real-world labels, provided that target definitions are carefully constructed and externally validated against imaging-derived endpoints.

## Acknowledgments

The authors gratefully acknowledge the computing time granted by the Resource Allocation Board and provided on the supercomputer Emmy/Grete at NHR-Nord@Göttingen as part of the NHR infrastructure. The calculations for this research were conducted with computing resources under the project nhr_ni_starter_27179. This project was partially funded by the German Federal Ministry of Research, Technology and Space as part of the ACRIBiS project of the Medical Informatics Initiative (FK: 01ZZ2317B).

## References

[1] James SL, Abate D, Abate KH, Abay SM, Abbafati C, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. The Lancet November 2018;392(10159):1789–1858.

[2] Taylor CJ, Ryan R, Nichols L, Gale N, Hobbs FR, Marshall T. Survival following a diagnosis of heart failure in primary care. Family Practice January 2017;cmw145.

[3] Varudo R, Gonzalez FA, Leote J, Martins C, Bacariza J, Fernandes A, Michard F. Machine learning for the real-time assessment of left ventricular ejection fraction in critically ill patients: a bedside evaluation by novices and experts in echocardiography. Critical Care December 2022;26(1):386.

[4] Remmelzwaal S, van Ballegooijen AJ, Schoonmade LJ, Dal Canto E, Handoko ML, Henkens MTHM, van Empel V, Heymans SRB, Beulens JWJ. Natriuretic peptides for the detection of diastolic dysfunction and heart failure with preserved ejection fraction—a systematic review and meta-analysis. BMC Medicine October 2020;18(1). ISSN 1741-7015.

[5] McKeen K, Masood S, Toma A, Rubin B, Wang B. Ecg-fm: an open electrocardiogram foundation model. JAMIA Open September 2025;8(5).

[6] Poterucha TJ, Jing L, Ricart RP, Adjei-Mosi M, Finer J, et al. Detecting structural heart disease from electrocardiograms using ai. Nature July 2025;644(8075):221–230.

[7] Koscova Z, Li Q, Robichaux C, Moura Junior V, Ghanta M, Gupta A, Rosand J, Aguirre A, Hong S, Albert DE, Xue J, Parekh A, Sameni R, Reyna MA, Westover MB, Clifford GD. The harvard-emory ecg database. Scientific Data 2026; 13:538.

[8] Koscova Z, Moura Junior V, Reyna MA, Hong S, Gupta A, Ghanta M, Sameni R, Aguirre A, Li Q, Zafar S, Clifford GD, Westover MB. Harvard-emory ecg database, 2026.

[9] McCormick N, Lacaille D, Bhole V, Avina-Zubieta JA. Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis. PLOS ONE August 2014;9(8):e104519.

[10] Wei Y, Deng Y, Sun C, Lin M, Jiang H, Peng Y. Deep learning with noisy labels in medical prediction problems: a scoping review. Journal of the American Medical Informatics Association July 2024;31(7):1596–1607.

[11] Koscova Z, Li Q, Robichaux C, Moura Junior V, Ghanta M, Gupta A, Rosand J, Aguirre A, Hong S, Albert DE, Xue J, Parekh A, Sameni R, Reyna MA, Westover MB, Clifford GD. The harvard-emory ecg database. medRxiv 2024;.

[12] Elias P, Finer J. EchoNext: A Dataset for Detecting Echocardiogram-Confirmed Structural Heart Disease from ECGs. PhysioNet 2025;Version 1.1.0.

[13] Hempel P, Steinbrinker T, Graf L, Sievers S, Riedemann G, Machreich T, von Haehling S, Krefting D, Spicher N. Adapting 12-lead ecg ai model to 1-lead smart watches for diagnosis in clinical heart failure patients. In Proceedings of the 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2025; 1–5.

[14] Graf L, Hempel P, Riedemann G, Steinbrinker T, Abdullayev A, Becker C, von Haehling S, Krefting D, Spicher N. Finetuning enables Atrial Fibrillation detection in 12-lead electrocardiograms with active pacing, 2026. Research Square preprint.

Address for correspondence:

Lennart Graf
Robert-Koch-Straße 40,
37075, Göttingen, Germany
frederiklennart.graf@med.uni-goettingen.de