Open Science to Foster Progress in Automatic ECG Analysis: Status and Future Directions

Nils Strodthoff[1]

[1] AI4Health Division, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

Abstract

Recently, there has been a significant increase in the availability of publicly accessible ECG datasets. These datasets have been used in thousands of research works, showcasing their profound impact on the research landscape. However, despite their widespread use, many currently available public datasets lack certain crucial components. These include the absence of high-quality expert annotations based on standardized ontologies or mapping schemes to facilitate comparisons across datasets. Additionally, there is a need for enriched metadata, such as ECG features or automatic diagnostic statements, to improve their usability and interoperability. Nevertheless, the most glaring shortfall of current publicly available datasets is the absence of clinical ground truth. This deficiency prevents the broader research community from replicating and expanding upon literature works based on closed in-hospital datasets, which impedes scientific progress. In this paper, we trace the recent history on publicly available ECG datasets and discuss possible directions for the enrichment of existing datasets. We close with a possible path towards ECG datasets with clinical ground truth based on recent work in the context of the MIMIC-IV-ECG dataset.

1. A short history of publicly available 12-lead ECG datasets

In the following, we will briefly describe the history of publicly available ECG datasets with a particular focus on 12-lead datasets suitable for supervised learning.

Pre-2020 era One of the most widely used 12 (or rather 15)-lead ECG dataset of the pre-2020 era was the PTB Diagnostic ECG dataset [1] with roughly 550 samples. Due to a focus of the dataset on myocardial infarction, it was most commonly used to develop algorithms for the detection of the latter. In 2018, it was supplemented by the ICBEB challenge dataset covering roughly 7000 ECGs, which were supposed to be classified into 8 diagnostic classes. A more comprehensive list of publicly available datasets preceding PTB-XL is provided in [2, 3].

2020-2023 In 2020, almost concurrently, the PTB-XL [2] and Chapman [4] datasets were published marking a surge of large-scale 12-lead ECG datasets with comprehensive label sets. Both were later included in the Computing in Cardiology Challenge 2021 (CinC2021) [5, 6], which brought comprehensive diagnostic prediction on large-scale ECG datasets into the attention of a broader research community. In 2021, also a 15%-subset of the large CODE dataset was released comprising more than 345k samples, however, with a relatively sparse annotation of six annotated pathologies. The most recent addition was the SPH [7] dataset, a large-scale 12-lead dataset with annotations according to the AHA standard [8].

Recent developments The most recent addition to the list of publicly available ECG datasets represents the MIMIC-IV-ECG datasets comprising 800k samples from 160k patients, which can be be linked to comprehensive clinical metadata from MIMIC-IV [9]. This enables to tackle entirely new clinical questions, as discussed in detail below.

Comparison to in-house datasets Comparing the nominal size of publicly available datasets such as the combined CinC2021 dataset with about 88k samples they still fall short by an order of magnitude compared to comprehensively annotated, large-scale datasets as used by MayoCV [10] with on the order of 2.5M samples. This applies even more to in-hospital datasets used to investigate specific conditions such as [11–13], which typically cover a few ten thousand samples. Nevertheless, unannotated large-scale ECG datasets such as the recently released Harvard-Emory ECG database [14] with more than 10M samples, in combination with self-supervised learning methods provide a promising path to significantly improve the label-efficiency of training procedures. This could help the open source community to overcome the limitations of too few comprehensively annotated datasets.

Role for the community Open source ECG datasets play a pivotal role for the research community, using citation counts of more than 874 for PTB Diagnostic and 680 for PTB-XL and 340 for Chapman (since 2020 in the two latter cases) are a proxy for the community relevance

 ISSN: 2325-887X DOI: 10.22489/CinC.2024.057

of those datasets, both within the core biomedical signal analysis but have now also found their way into the core domain of machine learning to benchmark newly proposed algorithms for time series data.

Benefits of open data Open datasets and open source implementations are the cornerstones of reproducible progress. Large-scale datasets such as those included in CinC2021 enable the broader research community to take part in the revolution of AI-enhanced ECG analysis that is currently taking over the field of ECG analysis. The role of publicly available datasets is most easily demonstrated at the example of the large-scale ImageNet dataset, which was one of the driving forces behind the rapid developments in computer vision in the recent decade.

2. Future directions

Documentation and benchmarking procedures Even though the number of open source datasets is growing steadily, the standards according to which they are annotated vary widely ranging from comprehensive data descriptors to datasets released as part of challenge datasets without further documentation. More stringent documentation practices, documenting the context in which the dataset was collected, details of annotation procedure as well as basic patient characteristics, would significantly enhance the reusability of datasets. In addition instructions on the intended use of the dataset along with specifics such as predefined train-test-splits, label sets and/or target metrics, as exemplified for PTB-XL [15], are key components to enable measurable progress through like-to-like comparisons.

Recommendation: We encourage dataset providers to provide basic metadata on context and purpose of the data collection. Furthermore, it would be helpful to carefully think about possible benchmarking scenarios and to provide clear usage instructions, such as stratified train-test splits and clinically most relevant target metrics.

Enrichment of existing datasets through ECG features Classical electrocardiography is largely driven by clinical expert features such as most notably intervals or amplitudes, which form the basis of most clinical decision rules and are integrated in rule-based decision support integrated in current ECG devices. Unfortunately, high-quality ECG features are only available from proprietary software. Ideally, these should be available to supplement existing datasets to allow to incorporate ECG features in hybrid algorithms, to stratify datasets according to ECG features or to investigate alignment of model decision with expert rules/concept [16]. PTB-XL+[17] represents a recent effort in this direction, incorporating ECG features from two commercial providers as well as automatic diagnosis statements from one commercial provider for the entire PTB-XL dataset.

Recommendation: We encourage the community to enrich publicly available datasets through ECG features and automatic diagnosis statements.

Common ontologies/mappings Existing publicly available ECG datasets are annotated based on broad variety of different ontologies. Some of them, such the SCP-ECG standard underlying PTB-XL, date back to the 1970s and do not align with modern diagnostic criteria. Another complication arises from a partial incompatibility of different annotations standards across different datasets, which is a major hurdle for crucially important cross-domain evaluations to externally validate models on unseen datasets. For newly annotated or re-annotated datasets this can be realized by adopting a common annotation standard, such as the widely used AHA standard [8]. This aligns with ongoing efforts of a re-annotation of the PTB-XL dataset based on a modern ontology, which contains AHA as a proper subset.

However, this should ideally go hand-in-hand with a clear specification of the underlying diagnostic criteria used for annotation, e.g. in terms of threshold values for heart rate or specific interval lengths. Otherwise, the same diagnostic term might refer to two different underlying conditions, which will negatively influence the training process when training on combinations of datasets and obfuscate the true model performance when evaluating on test datasets with mismatching diagnostic definitions.

Recommendation: We encourage dataset providers to annotate datasets based on a commonly accepted ontology such as AHA [8] along with clearly specified underlying diagnostic criteria. A short-term solution to facilitate cross-dataset evaluation is to provide at least mappings to a commonly used ontology such as SNOMED, which should ideally be provided by the dataset providers themselves.

Inclusion of clinical metadata Most studies in the context of automatic ECG interpretation are set up in a rather artificial scenario were the algorithm is supposed to provide a prediction solely based on the raw ECG as input. While this is the simplest scenario for benchmarking purposes, it neglects the clinical reality where additional clinical metadata is available. Not surprisingly, already the inclusion of basic demographic data yielded statistically significant performance improvements [18]. This encourages to move towards clinically more meaningful benchmarking tasks incorporating at least basic clinical metadata. The next step will require to specify more specific clinical use cases. A first example in this direction is MDS-ED [19] on clinical decision support in the emergency department providing first evidence for a statistically significant improvement in performance through the inclusion of raw waveform data.

Recommendation: Dataset creators are encouraged to *provide as much clinical metadata as possible but at least basic demographic/biometric data. Benchmarking scenarios should move beyond signal-only scenarios.*

Clinical ground truth By now there is a large body of evidence for the far-reaching diagnostic capabilities of AI-enhanced ECG analysis in extracting patterns from hidden correlations in the signals that remain hidden to the eyes even of human experts, see [20] for a recent review. This includes left ventricular dysfunction [21] valve disorders [22], and atrial fibrillation from sinus rhythm [23], but most notably also non-cardiac conditions such as anemia [11], diabetes [12] or cirrhosis [13]. All of these results were established on closed datasets due to the fact that no large-scale datasets with clinical labels (in particular of non-cardiac nature) were available.

This changed with the publication of MIMIC-IV-ECG [24] dataset, where clinical metadata such as discharge diagnoses [25] can be used as proxy for the clinical ground truth. This was subsequently used to demonstrate the feasibility of predicting a broad range of cardiac as well as non-cardiac conditions from a single model [25]. A different line of recent research, which also builds on MIMIC-IV-ECG, addresses the prediction of laboratory values from the ECG [26].

Recommendation: We encourage dataset creators to supplement clinical ICU/ED datasets with raw waveform data to further explore the full diagnostic potential of the ECG. This would allow a broader research community to participate in exploring the prospects of AI-enhanced ECG beyond retrospective annotation of cardiac conditions and would eventually enable external validation of results achieved for example on MIMIC-IV entirely based on publicly available data.

Foundation models Foundation models are the backbone of future generalist medical AI [27] and in short-term one path to improve the label-efficiency. Such foundation models can be trained in different ways, through self-supervision from signals alone [28, 29], through weak supervision from combinations of image and text [30] or through text dialogues with embedded ECG signals as assistant models. All of them crucially rely on the quality and composition of training data. We envision it would be a worthwhile endeavor to join forces to train comprehensive foundation models based on publicly available resources and to make them publicly available.

Recommendation: We encourage everyone to contribute to the collection and organization of pretraining data to develop ECG-based foundation models.

3. Summary and Conclusion

Reproducible progress in AI is grounded on open datasets, open code and the possibility to validate model results on external datasets. Electrocardiography is in the fortunate position to be able to serve as a role model for other medical domains since the richness of the diagnostic capabilities of AI-enhanced ECG is already widely recognized and ECGs are widely used for routine diagnostics and therefore relatively abundant. In addition, signals are low-dimensional and therefore easy to handle and to share and last but not least ECG signals are relatively easy to pseudonymize. We hope the recommendations put out in the previous section on how to enhance the value of existing datasets and how to maximize the reusability of newly created datasets will help existing and future dataset providers. We firmly believe that a strong focus on open data and open code can move the whole field forward.

References

[1] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik 1995;40(s1):317–318.

[2] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset. Scientific Data May 2020; 7(1).

[3] Clifford G. Past, Present and Future Challenges in Sharing Science: From PhysioNet to Foundation Models. In 2024 Computing in Cardiology (CinC). 2024; 1–4.

[4] Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead electrocardiogram database for arrhythmia research covering more than 10, 000 patients. Scientific Data February 2020;7(1).

[5] Reyna MA, Sadr N, Alday EAP, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, et al. Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. In 2021 Computing in Cardiology (CinC), volume 48. IEEE, 2021; 1–4.

[6] Reyna MA, Sadr N, Alday EAP, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, et al. Issues in the automated classification of multilead ECGs using heterogeneous labels and populations. Physiological Measurement 2022;43(8):084001.

[7] Liu H, Chen D, Chen D, Zhang X, Li H, Bian L, Shu M, Wang Y. A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements. Scientific Data June 2022;9(1).

[8] Mason JW, Hancock EW, Gettes LS. Recommendations for the Standardization and Interpretation of the Electrocardiogram. Journal of the American College of Cardiology March 2007;49(10):1128–1135.

[9] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman LwH, Celi LA, Mark RG. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data January 2023;10(1).

[10] Kashou AH, Ko WY, Attia ZI, Cohen MS, Friedman PA, Noseworthy PA. A comprehensive artificial intelligence–enabled electrocardiogram interpretation program.

Cardiovascular Digital Health Journal September 2020; 1(2):62–70.
[11] Kwon Jm, Cho Y, Jeon KH, Cho S, Kim KH, Baek SD, Jeung S, Park J, Oh BH. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. The Lancet Digital Health 2020;2(7):e358–e367.
[12] Kulkarni AR, Patel AA, Pipal KV, Jaiswal SG, Jaisinghani MT, Thulkar V, Gajbhiye L, Gondane P, Patel AB, Mamtani M, et al. Machine-learning algorithm to non-invasively detect diabetes and pre-diabetes from electrocardiogram. BMJ Innovations 2023;9(1).
[13] Ahn JC, Attia ZI, Rattan P, Mullan AF, Buryska S, Allen AM, Kamath PS, Friedman PA, Shah VH, Noseworthy PA, Simonetto DA. Development of the AI-Cirrhosis-ECG Score: An Electrocardiogram-Based Deep Learning Model in Cirrhosis. American Journal of Gastroenterology December 2021;117(3):424–432.
[14] Moura Junior V, Reyna M, Hong S, Gupta A, Ghanta M, Sameni R, Rosand J, Aguirre A, Li Q, Clifford G, Westover MB. Harvard-Emory ECG Database, 2023. URL https://bdsp.io/content/heedb/1.0/.
[15] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. IEEE Journal of Biomedical and Health Informatics 2020;25(5):1519–1528.
[16] Wagner P, Mehari T, Haverkamp W, Strodthoff N. Explaining deep learning for ecg analysis: Building blocks for auditing and knowledge discovery. Computers in Biology and Medicine 2024;176:108525.
[17] Strodthoff N, Mehari T, Nagel C, Aston PJ, Sundar A, Graff C, Kanters JK, Haverkamp W, Dössel O, Loewe A, Bär M, Schaeffter T. PTB-XL+, a comprehensive electrocardiographic feature dataset. Scientific Data May 2023;10(1).
[18] Mehari T, Strodthoff N. Towards quantitative precision for ECG analysis: Leveraging state space models, self-supervision and patient metadata. IEEE Journal of Biomedical and Health Informatics 2023;.
[19] Alcaraz JML, Bouma H, Strodthoff N. MDS-ED: Multimodal Decision Support in the Emergency Department – a Benchmark Dataset for Diagnoses and Deterioration Prediction in Emergency Medicine. ArXiv Preprint 2024;URL https://arxiv.org/abs/2407.17856.
[20] Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nature Reviews Cardiology 2021;18(7):465–478.
[21] Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine 2019;25(1):70–74.
[22] Elias P, Poterucha TJ, Rajaram V, Moller LM, Rodriguez V, Bhave S, Hahn RT, Tison G, Abreau SA, Barrios J, et al. Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. Journal of the American College of Cardiology 2022;80(6):613–626.
[23] Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, 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 2019;394(10201):861–867.
[24] Gow B, Pollard T, Nathanson LA, Johnson A, Moody B, Fernandes C, Greenbaum N, Waks JW, Eslami P, Carbonati T, Chaudhari A, Herbst E, Moukheiber D, Berkowitz S, Mark R, Horng S. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. PhysioNet 2023;URL https://physionet.org/content/mimic-iv-ecg/1.0/.
[25] Strodthoff N, Lopez Alcaraz JM, Haverkamp W. Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care. European Heart Journal Digital Health 2024;ztae039.
[26] Alcaraz JML, Strodthoff N. Cardiolab: Laboratory values estimation from electrocardiogram features – an exploratory study. In 2024 Computing in Cardiology (CinC). 2024; 1–4. URL https://arxiv.org/abs/2407.18629.
[27] Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. Nature April 2023; 616(7956):259–265.
[28] Mehari T, Strodthoff N. Self-supervised representation learning from 12-lead ECG data. Computers in Biology and Medicine 2022;141:105114.
[29] McKeen K, Oliva L, Masood S, Toma A, Rubin B, Wang B. Ecg-fm: An open electrocardiogram foundation model. ArXiv Preprint 2024;URL https://arxiv.org/abs/2408.05178.
[30] Liu C, Wan Z, Ouyang C, Shah A, Bai W, Arcucci R. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. In Forty-first International Conference on Machine Learning. 2024; .

Address for correspondence:

Nils Strodthoff
Carl von Ossietzky Universität Oldenburg
School VI Medicine and Health Services
Department of Health Services Research
AI4Health Division
Ammerländer Heerstr. 114-118
26129 Oldenburg, Germany
nils.strodthoff@uol.de