JACC: ADVANCES VOL. 2, NO. 10, 2023

ORIGINAL RESEARCH

EMERGING TECHNOLOGIES AND INNOVATIONS

Clinical Applications, Methodology, and Scientific Reporting of Electrocardiogram Deep-Learning Models

A Systematic Review

Vennela Avula, MD,[a] Katherine C. Wu, MD,[b] Richard T. Carrick, MD, PhD[b]

ABSTRACT

BACKGROUND The electrocardiogram (ECG) is one of the most common diagnostic tools available to assess cardiovascular health. The advent of advanced computational techniques such as deep learning has dramatically expanded the breadth of clinical problems that can be addressed using ECG data, leading to increasing popularity of ECG deep-learning models aimed at predicting clinical endpoints.

OBJECTIVES The purpose of this study was to define the current landscape of clinically relevant ECG deep-learning models and examine practices in the scientific reporting of these studies.

METHODS We performed a systematic review of PubMed and EMBASE databases to identify clinically relevant ECG deep-learning models published through July 1, 2022.

RESULTS We identified 44 manuscripts including 53 unique, clinically relevant ECG deep-learning models. The rate of publication of ECG deep-learning models is increasing rapidly. The most common clinical applications of ECG deep learning were identification of cardiomyopathy (14/53 [26%]), followed by arrhythmia detection (9/53 [17%]). Methodologic reporting varied; while 33/44 (75%) publications included model architecture diagrams, complete information required to reproduce these models was provided in only 10/44 (23%). Saliency analysis was performed in 20/44 (46%) of publications. Only 18/53 (34%) models were tested within external validation cohorts. Model code or resources allowing for model implementation by external groups were available for only 5/44 (11%) publications.

CONCLUSIONS While ECG deep-learning models are increasingly clinically relevant, their reporting is highly variable, and few publications provide sufficient detail for methodologic reproduction or model validation by external groups. The field of ECG deep learning would benefit from adherence to a set of standardized scientific reporting guidelines. (JACC Adv 2023;2:100686) © 2023 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

From the [a]Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; and the [b]Division of Cardiology, Johns Hopkins University Department of Medicine, Baltimore, Maryland, USA.

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received April 5, 2023; revised manuscript received August 31, 2023, accepted September 19, 2023.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2023.100686

ABBREVIATIONS AND ACRONYMS

AUROC = area under the receiver operator curve

ECG = electrocardiogram

Electrocardiograms (ECGs) are a mainstay of medical practice due to their clinical relevance, low cost, and wide availability. However, while ECG data can be used to diagnose both cardiac and noncardiac disorders, their interpretation requires significant training and expertise. Most modern ECG systems offer rule-based automated analysis, but these approaches rely on obvious, easily quantified ECG parameters such as the duration of, amplitude of, or intervals between segments of the cardiac cycle. These algorithms may miss more subtle ECG changes including those not apparent to the human eye and have limited ability to provide insight into more complex diagnoses.

More recently, advanced computational techniques including artificial intelligence and machine learning have expanded the breadth of clinical problems potentially addressable by ECG data. Deep-learning models, the subset of machine-learning models that rely on neural networks, are particularly adept at handling complex, high-dimensionality data like ECG waveforms, offering the potential to improve cardiac diagnoses (eg, identifying cardiomyopathy[1,2] or valvular disease[3,4]) and make relevant clinical predictions (eg, future arrhythmia or mortality[5,6]). Early results from ECG deep-learning models have been impressive, and their popularity has increased dramatically over the past several years. However, there is concern that the rapid growth of this branch of research has outstripped its reliability. The extent to which publications provide sufficient information to allow external validation and replication is unclear, as is model reproducibility.

To address this issue, we performed a systematic review to identify clinically relevant 12-lead ECG deep-learning models, describe the specific computational techniques employed for their creation, evaluate both the quality and consistency of the scientific reporting related to these models, and assess their reproducibility.

METHODS

All data relevant for these analyses are available online as part of the Supplemental Appendix.

IDENTIFICATION OF CLINICALLY RELEVANT 12-LEAD ECG DEEP-LEARNING MODELS. We performed a systematic review of novel deep-learning models that made use of 12-lead, surface ECG data to address clinically relevant problems. This review adhered to guidelines set out by the Cochrane Collaboration and Institute of Medicine[7] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement[8]; abstraction flowchart is shown in **Figure 1**.

We queried 2 large, open-access medical research databases, PubMed and EMBASE, using a standardized set of medical subject heading search terms (Supplemental Methods 1) to inclusively identify ECG deep-learning models published through July 1st, 2022. Briefly, manuscripts with references to both electrocardiography and deep learning, artificial intelligence, convolutional networks, or neural networks were targeted. Additional search filters included requirements for the English language, original research manuscripts (not reviews), and full-text availability.

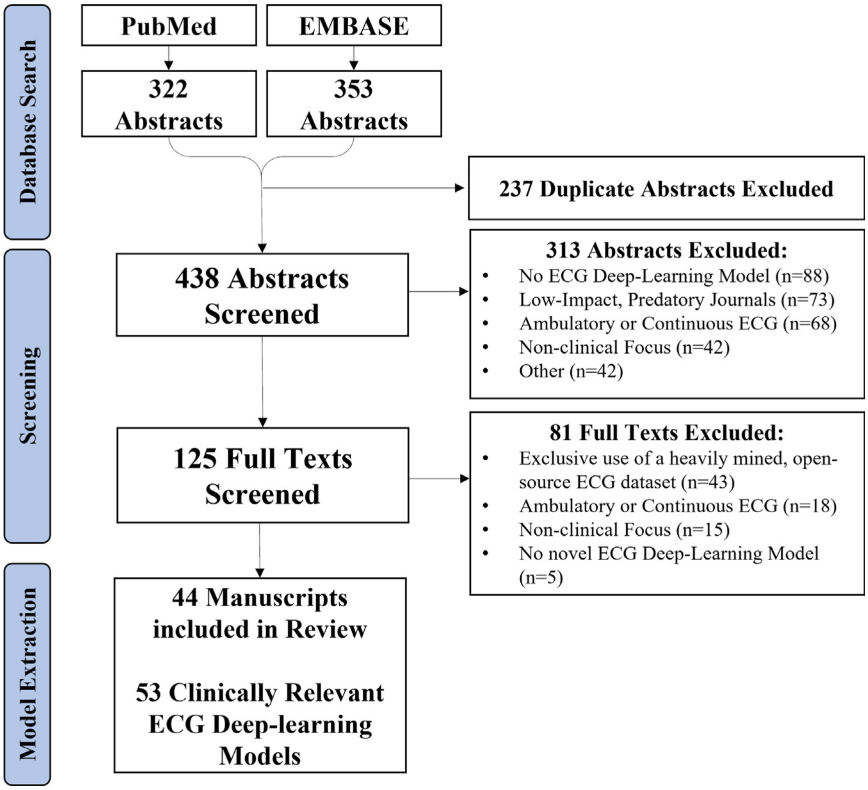
Two independent reviewers screened potential abstracts using *Rayyan*, a semiautomated online abstract screening and documentation program.[9] Discrepancies were discussed until consensus was achieved. We then selected abstracts for full-text review if they met the following inclusion criteria: 1) reference to the development of novel deep-learning models; 2) specified standard clinical 12-lead ECGs as a primary input into these models (rather than ambulatory electrocardiography or wearable devices); and 3) models were derived from a primarily adult population.

We then doubly screened full-text publications and included them for further analysis if they met the above inclusion criteria. At this stage, we identified a high volume of nonclinically focused ECG deep-learning models published in primarily low-impact journals.[10] To maintain the clinical focus of our review, we applied 3 additional exclusion criteria during review of full texts: 1) articles published in journals with overall H-index <40 as of July 1st, 2022; 2) articles that developed and tested their models exclusively using heavily mined, open-access ECG datasets such as the PhysioNet/Computing in Cardiology Challenge (CINC) ECG datasets[11]; and 3) obvious nonclinical focus (eg, use of ECG deep learning for the purposes of biometric identification or signal denoising). Here, newer journals published by reputable publishing groups that had not yet achieved our H-index criteria were included regardless.

The validity of our abstract screening was qualitatively verified by correct identification of several known "gold standard" ECG deep-learning manuscripts.[1,2,5,12,13]

DATA EXTRACTION. For those full-text manuscripts identified above, 2 independent reviewers extracted data on the studied population, the methods and results of the proposed ECG deep-learning models, as well as the rigor and approach to reporting of these

FIGURE 1 PRISMA Flowchart


Database Search
Screening
Model Extraction
PubMed
EMBASE
322 Abstracts
353 Abstracts
237 Duplicate Abstracts Excluded
438 Abstracts Screened
313 Abstracts Excluded:
• No ECG Deep-Learning Model (n=88)
• Low-Impact, Predatory Journals (n=73)
• Ambulatory or Continuous ECG (n=68)
• Non-clinical Focus (n=42)
• Other (n=42)
125 Full Texts Screened
81 Full Texts Excluded:
• Exclusive use of a heavily mined, open-source ECG dataset (n=43)
• Ambulatory or Continuous ECG (n=18)
• Non-clinical Focus (n=15)
• No novel ECG Deep-Learning Model (n=5)
44 Manuscripts included in Review
53 Clinically Relevant ECG Deep-learning Models

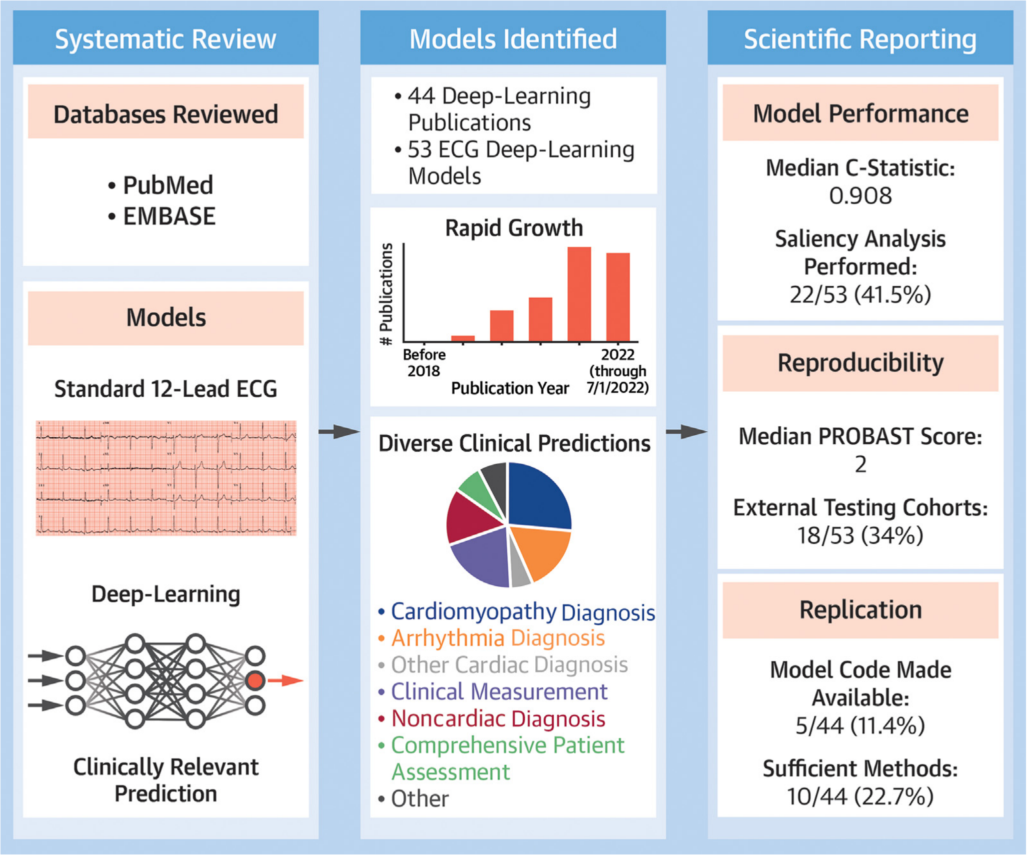
Flowchart showing abstract and manuscript screening, as well as ECG deep-learning model identification. ECG = electrocardiogram; PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

methods and results using a standard set of definitions and in accordance with the checklist for systematic reviews of prediction modeling studies.[14]

Collected fields included size and clinical characteristics of the cohorts, how the dataset was divided for the purposes of model development and training, details related to the design of the ECG deep-learning models, the outcomes predicted by the ECG deep-learning models, model performance reporting, and finally availability and reproducibility of the ECG deep-learning models. Models were considered reproducible if the publication included: 1) a diagram of model architecture (ie, a visual depiction of the connections between model layers); 2) specific definition of the kernel size and number of convolutional filters at each convolutional layer; and 3) specific definition of the techniques and hyperparameters used for model development including the type of activation function layers, the loss function, the learning rate, the training optimizer, and mini-batch size. A full list of extracted characteristics is presented in the Supplemental Methods. Risk of model bias was assessed using the previously validated short-form of the Prediction Model Risk of Bias Assessment Tool (PROBAST).[15] This tool, designed for critical appraisal of bias in prediction models, consists of 6 distinct fields reflecting methodologic approaches to: 1) outcome assessment; 2) events per predictor variables; 3) continuous predictors; 4) missing data; 5) univariable analysis; and 6) overfitting/optimism (Supplemental Methods).

In this systematic review, we defined the following types of datasets: 1) the training dataset, defined as the set of ECGs or patients that were used directly for training of model weights; 2) the validation dataset, defined as the set of ECGs or patients that were used as an internal cross-check during model training for the purposes of hyperparameter tuning, model selection, or to define scheduled changes to the learning rate or early stopping; 3) the development dataset, defined as the combination of training and validation datasets; 4) the testing dataset, defined as

CENTRAL ILLUSTRATION Systematic Review of ECG Deep-Learning Models Demonstrated Excellent Performances but Variable Approaches to Scientific Reporting



Systematic Review

Databases Reviewed

- PubMed
- EMBASE

Models

Standard 12-Lead ECG

Deep-Learning

Clinically Relevant Prediction

Models Identified

- 44 Deep-Learning Publications
- 53 ECG Deep-Learning Models

Rapid Growth

Diverse Clinical Predictions

- Cardiomyopathy Diagnosis
- Arrhythmia Diagnosis
- Other Cardiac Diagnosis
- Clinical Measurement
- Noncardiac Diagnosis
- Comprehensive Patient Assessment
- Other

Scientific Reporting

Model Performance

Median C-Statistic: 0.908

Saliency Analysis Performed: 22/53 (41.5%)

Reproducibility

Median PROBAST Score: 2

External Testing Cohorts: 18/53 (34%)

Replication

Model Code Made Available: 5/44 (11.4%)

Sufficient Methods: 10/44 (22.7%)

Avula V, et al. JACC Adv. 2023;2(10):100686.

ECG = electrocardiogram.

the dataset used to assess the performance of the model developed within the development dataset; and 5) the overall cohort, defined as the combination of development and testing datasets. An external testing dataset was defined as a dataset comprised of data from either a distinct location or from a temporally distinct time span.

STATISTICAL ANALYSIS. Model summary statistics were performed using Python (v3.9.13) and the open-source Pandas (v1.5.3) data analysis library.

RESULTS

IDENTIFICATION OF CLINICALLY RELEVANT 12-LEAD ECG DEEP-LEARNING MODELS. We identified 53 distinct, clinically relevant 12-lead ECG deep-learning models included in 44 published manuscripts (**Central Illustration**, **Table 1**); summary statistics of these models are presented in Supplemental Table 1. Manuscripts that were excluded during full review are shown in Supplemental Table 2. These manuscripts were published by groups throughout the world, the most prolific countries being the United States (19/44, 43%), China (8/44, 18%), Japan (5/44, 11%), and the Netherlands (5/44, 11%). The rate of publication of manuscripts related to ECG deep learning is escalating (**Figure 2**), with none published before 2018 and more than 30 published after January 1, 2022. The most common clinical applications of these models were identification of cardiomyopathies (14/53 [26%]), prediction of abnormal clinical, laboratory, or imaging findings (11/53 [21%]), and arrhythmia detection (9/53 [17%]). Four groups published comprehensive deep-learning models that

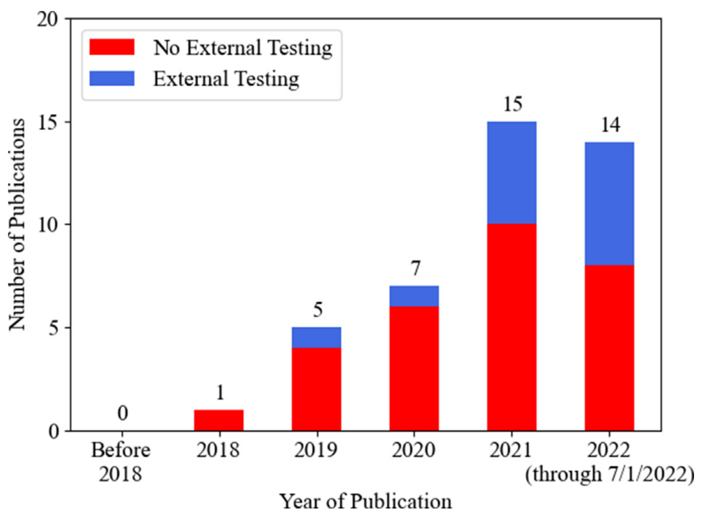
TABLE 1 Characteristics of ECG Deep-Learning Models

PMID	First Author	Publication Year	Outcome(s)	Overall Dataset Size	Total Number of Sequential Convolutional Layers	AUROC	External Testing Cohort?	Mean AUROC in External Cohort
30133452[16]	Attia	2018	Dofetilide toxicity	NR	7	NR	No	-
30617318[2]	Attia	2019	LVEF <35%	97,829	7	0.93	No	-
30942845[17]	Galloway	2019	Hyperkalemia	1,626,680	11	0.883	Yes	0.856
31378392[13]	Attia	2019	Silent atrial fibrillation	649,931	19	0.87	No	-
31450977[18]	Attia	2019	Patient age	774,783	9	NR	No	-
			Patient sex	774,783	9	0.973	No	-
None[19]	Li	2019	HF stages (From 2s ECG)	17,190	3	NR	No	-
			HF stages (From 5s ECG)	6,876	3	NR	No	-
32081280[12]	Ko	2020	Hypertrophic cardiomyopathy	67,001	NR	0.96	No	-
32393799[5]	Raghunath	2020	Mortality (at 1 y)	2,338,833	7	0.876	No	-
32406296[20]	van de Leur	2020	Multi-class triage category	337,819	35	0.93	No	-
33006947[21]	Gumpfer	2020	Myocardial Scar on CMR	114	5	NR	No	-
33328094[22]	Zhu	2020	21 distinct ECG rhythms	180,940	13	0.983	Yes	0.995
33392274[23]	Jiang	2020	Left atrial dilation	3,391	7	0.949	No	-
35265877[24]	Kashou	2020	66 distinct cardiology diagnoses	2,499,522	33	0.98	No	-
35265893[25]	Nakamura	2020	PVC origin	464	2	0.908	No	-
33401921[1]	van de Leur	2021	PLN mutation	1,806	14	0.95	No	-
33565217[26]	Sun	2021	LVEF <50%	26,792	NR	0.713	No	-
33566059[27]	Bos	2021	LQT syndrome	9,085	10	0.90	No	
			Specific LQT mutation	9,085	10	0.944	No	-
33588584[6]	Raghunath	2021	Atrial fibrillation (at 1 y)	1,151,037	6	0.85	No	-
			Atrial fibrillation (at 1 y)	564,573	6	0.83	Yes	0.85
33607378[28]	Lopes	2021	PLN mutation	13,622	9	0.90	No	-
33748852[4]	Cohen-Shelly	2021	Aortic stenosis	258,607	62	0.85	No	-
33850245[29]	Nishimori	2021	Accessory pathway location	NR	4	NR	Yes	NR
33917563[30]	Chang	2021	Digoxin toxicity	177,127	82	0.912	No	-
34126762[31]	Khurshid	2021	Left ventricular hypertrophy	37,142	10	0.653	Yes	0.621
34225095[32]	Jo	2021	9 distinct ECG rhythms	56,942	11	0.976	Yes	0.966
34308091[33]	Lin	2021	Thyrotoxic periodic paralysis	588	82	0.986	No	-
34347007[34]	Hughes	2021	38 distinct cardiac diagnoses	351,657	34	0.974	Yes	0.952
34468739[35]	Prifti	2021	Sotalol toxicity	10,292	22	0.948	Yes	0.92
34853226[36]	Katsushika	2021	LVEF <40%	37,103	7	0.945	No	-
34993487[37]	Akbilgic	2021	Heart failure (at 10 y)	14,613	NR	0.756	No	-
33930574[38]	Chen	2022	9 distinct cardiac diagnoses	26,130	8	NR	No	-
34544652[3]	Sawano	2022	Aortic regurgitation	29,859	7	0.802	No	-
34743566[39]	Khurshid	2022	Atrial fibrillation	NR	16	0.823	Yes	0.726
35029163[40]	Ahn	2022	Cirrhosis	25,940	9	0.908	No	-
35153641[41]	Zang	2022	Depression	5,060	2	NR	No	-
35332137[42]	Sangha	2022	6 distinct cardiac diagnoses	2,228,236	125	0.99	Yes	0.972
35360023[43]	Wu	2022	STEMI	793	3	0.999	Yes	1.00
			Culprit STEMI vessel	793	3	0.958	Yes	0.96
35387940[44]	Nakasone	2022	PVC origin	80	NR	NR	No	-
35463761[45]	Han	2022	CAC score >100	8,178	13	NP	Yes	0.718
			CAC score >400	8,178	13	NP	Yes	0.777
			CAC score >1,000	8,178	13	NP	Yes	0.803
35501785[46]	Aufiero	2022	LQT1 mutation	10,748	10	0.90	Yes	0.86
			LQT2 mutation	11,122	10	0.92	Yes	0.87
			LQT3 mutation	10,636	10	0.89	No	-
35533456[47]	Agrawal	2022	Post-COVID status	532	3	1.00	No	-
35629186[48]	Chang	2022	PVC origin	4,109	6	0.963	No	-
35707008[49]	Jiang	2022	Elevated CRP	12,315	10	0.85	No	-
36713005[50]	Siegersma	2022	Patient sex	287,547	NR	0.89	Yes	0.915
None[51]	Schlesinger	2022	Elevated PCWP	6,739	16	0.79	No	-

Full characteristics of the 44 publications containing 53 clinically relevant ECG deep-learning models.

AUROC = area under the receiver operator curve; ECG = electrocardiogram; NP = not performed; NR = not-reported.

FIGURE 2 Histogram of Increasing Electrocardiogram Deep-Learning Publication Rate



Blue = publications that included external testing datasets; red = publications with models tested in hold-out testing datasets only.

simultaneously predicted a large number of clinical diagnoses. The type of problems that models were designed to solve were largely classification (46/53, 86.8%), though a few models attempted to make predictions of future events (4/53, 9.4%), and 2 models generated regressions of a continuous variable.

ECG DEEP-LEARNING MODEL DESIGN AND IMPLEMENTATION. While, by definition, all models incorporated 12-lead surface ECGs as an input source, only 22 (41.5%) used the full 10-second (27/53, 61.4%) 12-lead (39/53, 88.6%) waveforms. Common alternative approaches were inclusion of 8 orthogonal ECG leads (10/53, 22.7%) or use of <2.5s worth of waveform data (22/53, 41.5%) such as specific segments containing median beats (2/53, 4.5%) or ectopic beats (3/53, 6.8%). The majority of ECG waveform data was sampled at 500 Hz frequency (34/53, 77.3%). In addition to ECG data, 9 of 53 models (17.0%) required additional clinical factors such as age and sex for generation of model predictions. Two models (4.5%) incorporated ECG images rather than raw voltage waveform data.

Diagrams showing model architecture were included in 33/44 (75%) of publications. The number of convolutional layers was highly variable across models, ranging from as low as 2 to as many as 125 layers. The organization of these convolutional layers also varied. While the most common organizational strategy used standard, sequential convolutional layers (23/53, 43.4%), residual connection blocks (17/53, 32.1%) including blocks of densely connected residual connections (4/53, 7.5%), and causal dilation blocks (3/48, 5.6%) were also commonly employed. Elements of recurrent neural networks (long short-term memory units) were occasionally added synergistically with convolutional layers (5/53, 9.4%).

ECG DEEP-LEARNING MODEL DEVELOPMENT AND PERFORMANCE ASSESSMENT. The majority of models were derived from retrospective cross-sectional datasets (37/53, 69.8%), while around one-third were derived from outcome-enriched case-control cohorts (16/53, 30.2%). Most models (51/53, 96%) included at least some description of the ECG dataset from which they were developed or tested, but descriptions were inconsistent. The most common strategies used for describing datasets were presenting: 1) only information related to the overall cohort (18/53, 34.0%); 2) information related to the development and testing datasets (14/53, 26.4%); or 3) complete information related to the training, validation, and testing datasets (17/53, 32.1%). For the 50 models reporting sufficient information for its estimation, overall dataset sizes ranged broadly between 80 ECGs and 2.5 million ECGs. Patient age (40/53, 75.5%) and sex (43/53, 81.1%) were largely reported for these datasets, but patient race (5/53, 9.4%) was rarely reported. Most, but not all, models included at least some information regarding the incidence of the outcome to be predicted (46/53, 86.8%).

The most common metric used for assessing the performance of ECG deep-learning models in testing datasets was the area under the receiver operator curve (AUROC), reported for 44/53 (83.0%) of the models. Median AUROCs by category of model prediction are presented in Supplemental Table 3. Other metrics reported with high frequency were sensitivity (35/53, 66.0%), specificity (30/53, 56.6%), accuracy (24/53, 45.3%), and F1 statistic (the harmonic mean of precision and recall) (18/53, 34.0%). Of note, specific criteria for selecting the operating threshold at which sensitivity, specificity, and accuracy were ascertained were reported in only 23/44 (52.3%) publications. Few publications assessed model calibration (6/44, 13.6%). Fewer than half of publications performed saliency analysis (20/44, 45.5%), with the most popular approach being gradient-based class activation mapping (Grad-CAM) (14/20, 70.0%).

MODEL REPRODUCIBILITY. Only 18 of the 53 ECG deep-learning models (34.0%) were tested in external testing cohorts. Among the few models reporting this metric for both hold-out testing and external testing, AUROC held up well, actually increasing by a median value of 0.022 [IQR: −0.001 to 0.032]. However, risk of bias, as assessed by short-form PROBAST, was elevated for all models with median score of 2 [IQR: 2-3].

Published details required for recreation of model training were frequently incomplete, with only 10/44 (22.7%) publications reporting the complete information required for reproduction of the described models. Only 5 publications (11.4%) included freely available code or online resources that would allow for model testing by an external group. While a few publications stated explicitly that model code would not be shared (4/44, 9.1%), the majority either provided nonspecific statements that code could potentially be shared upon request (16/44, 36.4%) or provided no statement regarding code availability (19/44, 43.2%).

DISCUSSION

In this systematic review, we identified over 40 unique publications including more than 50 distinct, clinically focused, ECG deep-learning models. These publications propose solutions to a wide variety of clinical problems, ranging from the highly specific identification of individual genetic mutations to the more comprehensive, automated prediction of a wide array of ECG abnormalities. The heterogeneity of these models is not limited to their applications; modeling techniques, the types of datasets used to develop and test these models, and the approaches used to assess model performances vary widely as well. While this variability demonstrates the strength and versatility of ECG deep learning, it also highlights the absence of a standardized approach for the scientific reporting of these models.

CLINICAL APPLICATIONS OF ECG DEEP-LEARNING. Deep learning has proven to be particularly well suited for extracting meaningful clinical information from high-dimensional, relational data such as ECG voltage waveforms. As testament to this, many of the ECG deep-learning models that we identified were able to make diagnoses that would not previously have been possible based on ECG alone. One of the most common and clinically relevant applications of these models is to identify patients within the general population who are likely to have cardiac conditions such as hypertrophic cardiomyopathy,[12] aortic stenosis,[4] or aortic regurgitation.[3] Models of this type have the potential to expand the role of ECG in screening for cardiac disease.

Another interesting clinical application illustrated by several models identified in this review involved the potential to predict abnormal downstream test results such as elevated pulmonary capillary wedge pressure during invasive right heart catheterization,[51] presence of scar on cardiac magnetic resonance imaging,[21] and reduced ejection fraction on echocardiography.[2,26,36] These types of ECG deep-learning models have the potential to help guide resource allocation. For example, focusing testing on those patients with a high likelihood of abnormal results may increase the rate of significant findings identified per test performed, in turn leading to decreased costs and reduced burden of unnecessary testing.

Finally, one of the most difficult (but potentially most clinically valuable) applications of ECG deep learning was for differentiation of alternative diagnoses with clinically similar ECG manifestations, such as discriminating between patients with drug-induced vs hereditary long QT syndromes,[35] identifying the location of accessory pathways in Wolff-Parkinson-White syndrome,[29] and differentiating left vs right-sided culprit vessel during inferior ST-segment elevation myocardial infarctions.[43] Models developed for these purposes are only relevant for those patients suffering from the index condition (eg, a patient must have Wolf-Parkinson-White to identify the site of an accessory pathway), but have the potential to help guide strategies during invasive procedures or clarify otherwise difficult diagnostic dilemmas.

Models designed for each of the 3 classes of clinical application described above demonstrated excellent discrimination, and the size of the ECG dataset used for model development seemed to have little impact on performances. Model performances were much more dependent on the specific outcomes being predicted. Outcomes with more obviously manifested ECG abnormalities resulted in excellent results (eg, the presence or absence of atrial fibrillation), while those with subtler ECG changes (eg, aortic stenosis) were more difficult to predict. As the clearest example of this, Han et al[45] presented 3 alternative models predicting coronary artery calcium scores on computerized tomography. While high-burdens of coronary calcium might be associated with other structural and electrophysiologic changes that manifest on ECG, a low-calcium burden may be found in patients with otherwise normal ECGs. This was clearly reflected by the performances of their models: the model predicting calcium score >1,000 had a good AUROC of 0.803 while the model predicting calcium score >100 had AUROC of only 0.718. These difficulties with detection of subtle ECG changes may in part be overcome through the use of very large datasets. The Mayo and Geisinger groups in particular have demonstrated the possibility of predicting future events such as mortality[5] and incident atrial fibrillation[6] using deep-learning models trained on millions of ECGs.

As the popularity of these ECG deep-learning models has grown, the barriers for entry into this field of research have contemporaneously decreased. The costs of computer hardware such as graphical processing units continue to drop, and a number of freely available, relatively easy to use deep-learning code libraries are now available. Further, researchers interested in ECG deep-learning models now have access to several large, open-source, online ECG databases. Of particular importance, PhysioNet has published data from nearly 90,000 ECGs as part of its annual CINC competition.[11] This important step forward in the democratization of ECG deep learning has proven to be a double-edged sword, however. Since its publication, many research teams have made use of the PhysioNet/CINC dataset as a 'toy' problem for demonstrating the value of technical innovations or novel deep-learning techniques. This has resulted in an overabundance of ECG deep-learning models demonstrating incremental improvements in the diagnosis of more readily identified clinical entities such as atrial fibrillation or myocardial infarction. While these computer-science and engineering-focused publications are essential for advancing the field of deep learning as a whole, we made the purposeful decision to exclude these ECG deep-learning models and instead focus our systematic review on those models with more direct clinical relevance.

LACK OF STANDARDIZED SCIENTIFIC REPORTING/REDUCED OPTIONS FOR EXTERNAL REPRODUCIBILITY. Because of the inherent complexity of deep-learning models and the innumerable possible variations in their development and design, an accurate, detailed description of methods is critical. However, our systematic review identified both a lack of standardization across the field as well as variable detail of methodologic reporting among individual publications. Even basic definitions varied substantially between manuscripts. For example, a cohort comprised of patients that did not contribute data during model development and which is subsequently used for assessing model performance might be described as a "hold-out testing" cohort (as we have defined for this review) in 1 manuscript or as a "validation" cohort in another. This lack of shared language can make it difficult to discern which models have undergone true external testing and which may be subject to bias or overfitting. Further confounding the evaluation of bias in these models, the method by which operating thresholds for evaluation of sensitivity, specificity, and accuracy were selected was defined for only around half of models.

Unlike traditional clinical predictive models, where publishing prognostic formulas is required for their clinical deployment,[52] the reproducibility and possibility for external testing of ECG deep-learning models are much more limited. As few as 1 in 5 ECG deep-learning models were described in sufficient detail to enable an experienced external research group to recreate their development and design. Further compounding this issue, only around 10% of manuscripts published code or provided online resources facilitating external model testing. While a larger contingent of publications (~35%) did declare that data or models could theoretically be shared upon specific request to the corresponding author, such statements place the burden of obtaining code on external groups, and there is no mechanism for their enforcement. Concerns regarding protection of intellectual property and/or the potential for future commercialization may contribute to decreased public availability.

These findings emphasize the need for a standardized set of guidelines for the scientific reporting of deep-learning models that includes descriptions of both the details required for model recreation (eg, architecture diagrams as well as those hyperparameters included in Supplemental Methods 2) and characteristics of the cohorts in which those models were developed and tested (eg, cohort size, age, sex, race, and event rates). Unfortunately, previously published guidelines designed for traditional clinical predictive models are ill-equipped for this purpose. The TRIPOD checklist for transparent reporting of traditional clinical predictive models[53] does not provide adequate standards for descriptions of deep-learning model design, standardized definitions for the ways in which model validation and testing should be performed, or expectations regarding the public availability of published models. Bias assessment using PROBAST[15] systematically overestimates deep-learning model bias due to the extremely low ratio of the numbers of outcome events to model parameters (eg, model weights) inherent to these models, and does not take cohort size, cohort composition (eg, case/control vs cross-sectional cohorts), or methods for selecting operating thresholds into consideration. These limitations to existing standards have been recognized, and updated versions of both TRIPOD and PROBAST specifically for use with machine learning are currently under development.[54] Of note, while standardized reporting of model details will allow for enhanced model interpretation and reproducibility, the specific values of those reported model parameters require problem-specific optimization and may therefore be different for different use cases. Finally, while explorations of model explainability are an important part of assessing the mechanisms by which models make their predictions, current techniques for saliency analysis have demonstrated poor performance on clinical tasks.[55] Thus, their role in standardized reporting of deep-learning models remains unclear.

STUDY LIMITATIONS. Although we applied a systematic approach to identify novel, clinically relevant ECG deep-learning models, our search was limited to the PubMed and EMBASE research databases. It is possible that there are models and/or validation studies published in alternative databases that we failed to include. Likewise, while our decision to exclude manuscripts published in journals with lower H-index was purposeful to maintain the clinical focus of our review,[10] it is possible that models with true clinical relevance could have been accidentally excluded; these criteria may also increase bias resulting from higher-performing models being selected for publication.

CONCLUSIONS

ECG deep-learning models are increasingly directed at clinically relevant endpoints and have demonstrated excellent performance over a wide range of diagnostic and predictive purposes. Their reporting is highly variable, however, and few publications provide the means for methodologic reproduction or model testing by external groups. The field of ECG deep learning would benefit from adherence to a standardized set of scientific reporting guidelines.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

Dr Carrick is funded by an NIH T32 Grant (T32HL007227) and the NIH Loan-Repayment Program (L30HL165535). All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Richard T. Carrick, Johns Hopkins Outpatient Center, 601 N. Caroline Street, 7th Floor, Baltimore, Maryland 21287, USA. E-mail: Rcarric5@jhmi.edu.

PERSPECTIVES

COMPETENCY IN MEDICAL KNOWLEDGE: ECG deep-learning models are increasingly relevant for the practice of clinical cardiology and medical research.

TRANSLATIONAL OUTLOOK: Development of a standardized set of guidelines for the scientific reporting of deep-learning models is critical.

REFERENCES

1. van de Leur RR, Taha K, Bos MN, et al. Discovering and visualizing disease-specific electrocardiogram features using deep learning: proof-of-concept in phospholamban gene mutation carriers. *Circ Arrhythm Electrophysiol*. 2021;14:e009056.

2. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25:70-74.

3. Sawano S, Kodera S, Katsushika S, et al. Deep learning model to detect significant aortic regurgitation using electrocardiography. *J Cardiol*. 2022;79:334-341.

4. Cohen-Shelly M, Attia ZI, Friedman PA, et al. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. *Eur Heart J*. 2021;42:2885-2896.

5. Raghunath S, Ulloa Cerna AE, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med*. 2020;26:886-891.

6. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, et al. Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation-related Stroke. *Circulation*. 2021;143:1287-1298.

7. Cumpston M, Li T, Page MJ, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. *Cochrane Database Syst Rev*. 2019;10:Ed000142.

8. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.

9. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210.

10. Grudniewicz A, Moher D, Cobey KD, et al. Predatory journals: no definition, no defence. *Nature*. 2019;576:210-212.

11. Reyna MA, Sadr N, Alday EAP, et al. Will two do? varying dimensions in electrocardiography: the PhysioNet/computing in cardiology challenge 2021. In: *2021 Computing in Cardiology (CinC)*. 48. IEEE; 2021:1-4.

12. Ko WY, Siontis KC, Attia ZI, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J Am Coll Cardiol*. 2020;75:722-733.

13. Attia ZI, Noseworthy PA, Lopez-Jimenez F, 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. *Lancet*. 2019;394:861-867.

14. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11:e1001744.

15. Venema E, Wessler BS, Paulus JK, et al. Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. *J Clin Epidemiol*. 2021;138:32-39.

16. Attia ZI, Sugrue A, Asirvatham SJ, et al. Noninvasive assessment of dofetilide plasma concentration using a deep learning (neural network) analysis of the surface electrocardiogram: a proof of concept study. *PLoS One*. 2018;13:e0201059.

17. Galloway CD, Valys AV, Shreibati JB, et al. Development and validation of a deep-learning model to screen for Hyperkalemia from the electrocardiogram. *JAMA Cardiol*. 2019;4:428-436.

18. Attia ZI, Friedman PA, Noseworthy PA, et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythm Electrophysiol*. 2019;12:e007284.

19. Li D, Li X, Zhao J, Bai X. Automatic staging model of heart failure based on deep learning. *Biomed Signal Process Control*. 2019;52:77-83.

20. van de Leur RR, Blom LJ, Gavves E, et al. Automatic triage of 12-lead ECGs using deep convolutional neural networks. *J Am Heart Assoc*. 2020;9:e015138.

21. Gumpfer N, Grün D, Hannig J, Keller T, Guckert M. Detecting myocardial scar using electrocardiogram data and deep neural networks. *Biol Chem*. 2021;402:911-923.

22. Zhu H, Cheng C, Yin H, et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *Lancet Digit Health*. 2020;2:e348-e357.

23. Jiang J, Deng H, Xue Y, Liao H, Wu S. Detection of left atrial enlargement using a convolutional neural network-enabled electrocardiogram. *Front Cardiovasc Med*. 2020;7:609976.

24. Kashou AH, Ko WY, Attia ZI, Cohen MS, Friedman PA, Noseworthy PA. A comprehensive artificial intelligence-enabled electrocardiogram interpretation program. *Cardiovasc Digit Health J*. 2020;1:62-70.

25. Nakamura T, Nagata Y, Nitta G, et al. Prediction of premature ventricular complex origins using artificial intelligence-enabled algorithms. *Cardiovasc Digit Health J*. 2021;2:76-83.

26. Sun JY, Qiu Y, Guo HC, et al. A method to screen left ventricular dysfunction through ECG based on convolutional neural network. *J Cardiovasc Electrophysiol*. 2021;32:1095-1102.

27. Bos JM, Attia ZI, Albert DE, Noseworthy PA, Friedman PA, Ackerman MJ. Use of artificial intelligence and deep neural networks in evaluation of patients with electrocardiographically concealed long QT syndrome from the surface 12-lead electrocardiogram. *JAMA Cardiol*. 2021;6:532-538.

28. Lopes RR, Bleijendaal H, Ramos LA, et al. Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning: an application to phospholamban p.Arg14del mutation carriers. *Comput Biol Med*. 2021;131:104262.

29. Nishimori M, Kiuchi K, Nishimura K, et al. Accessory pathway analysis using a multimodal deep learning model. *Sci Rep*. 2021;11:8045.

30. Chang DW, Lin CS, Tsao TP, et al. Detecting digoxin toxicity by artificial intelligence-assisted electrocardiography. *Int J Environ Res Publ Health*. 2021;18:3839.

31. Khurshid S, Friedman S, Pirruccello JP, et al. Deep learning to predict cardiac magnetic resonance-derived left ventricular mass and hypertrophy from 12-lead ECGs. *Circ Cardiovasc Imaging*. 2021;14:E012281.

32. Jo YY, Kwon JM, Jeon KH, et al. Detection and classification of arrhythmia using an explainable deep learning model. *J Electrocardiol*. 2021;67:124-132.

33. Lin C, Lin CS, Lee DJ, et al. Artificial intelligence-assisted electrocardiography for early diagnosis of thyrotoxic periodic paralysis. *J Endocr Soc*. 2021;5:bvab120.

34. Hughes JW, Olgin JE, Avram R, et al. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA Cardiol*. 2021;6:1285-1295.

35. Prifti E, Fall A, Davogustto G, et al. Deep learning analysis of electrocardiogram for risk prediction of drug-induced arrhythmias and diagnosis of long QT syndrome. *Eur Heart J*. 2021;42:3948-3961.

36. Katsushika S, Kodera S, Nakamoto M, et al. The effectiveness of a deep learning model to detect left ventricular systolic dysfunction from electrocardiograms. *Int Heart J*. 2021;62:1332-1341.

37. Akbilgic O, Butler L, Karabayir I, et al. ECG-AI: electrocardiographic artificial intelligence model for prediction of heart failure. *Eur Heart J Digit Health*. 2021;2:626-634.

38. Chen CY, Lin YT, Lee SJ, et al. Automated ECG classification based on 1D deep learning network. *Methods*. 2022;202:127-135.

39. Khurshid S, Friedman S, Reeder C, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*. 2022;145:122-133.

40. Ahn JC, Attia ZI, Rattan P, et al. Deep learning based AI-ECG-cirrhosis (ACE) score accurately predicts cirrhosis and gauges its severity. *Gastroenterology*. 2021;160:S776-S777.

41. Zang X, Li B, Zhao L, Yan D, Yang L. End-to-end depression recognition based on a one-dimensional convolution neural network model using two-lead ECG signal. *J Med Biol Eng*. 2022;42:225-233.

42. Sangha V, Mortazavi BJ, Haimovich AD, et al. Automated multilabel diagnosis on electrocardiographic images and signals. *Nat Commun*. 2022;13:1583.

43. Wu L, Huang G, Yu X, et al. Deep learning networks accurately detect ST-Segment elevation myocardial Infarction and culprit vessel. *Front Cardiovasc Med*. 2022;9:797207.

44. Nakasone K, Nishimori M, Kiuchi K, et al. Development of a visualization deep learning model for classifying origins of ventricular arrhythmias. *Circ J*. 2022;86(8):1273-1280.

45. Han C, Kang KW, Kim TY, et al. Artificial intelligence-enabled ECG algorithm for the prediction of coronary artery calcification. *Front Cardiovasc Med*. 2022;9:849223.

46. Aufiero S, Bleijendaal H, Robyns T, et al. A deep learning approach identifies new ECG features in congenital long QT syndrome. *BMC Med*. 2022;20:162.

47. Agrawal A, Chauhan A, Shetty MK, Girish MP, Gupta MD, Gupta A. ECG-iCOVIDNet: Interpretable AI model to identify changes in the ECG signals of post-COVID subjects. *Comput Biol Med*. 2022;146:105540.

48. Chang TY, Chen KW, Liu CM, et al. A high-precision deep learning algorithm to localize idiopathic ventricular arrhythmias. *J Pers Med*. 2022;12:764.

49. Jiang J, Deng H, Liao H, et al. Development and validation of a deep-learning model to detect CRP level from the electrocardiogram. *Front Physiol*. 2022;13:864747.

50. Siegersma KR, van de Leur RR, Onland-Moret NC, et al. Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk. *Eur Heart J Digit Health*. 2022;3:245-254.

51. Schlesinger Daphne E, Diamant N, Raghu A, et al. A deep learning model for Inferring elevated pulmonary capillary wedge pressures from the 12-lead electrocardiogram. *JACC Adv*. 2022;1:1-11.

52. Carrick RT, Park JG, McGinnes HL, et al. Clinical predictive models of Sudden cardiac arrest: a survey of the current science and analysis of model performances. *J Am Heart Assoc*. 2020;9:e017625.

53. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.

54. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11:e048008.

55. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol Artif Intell*. 2021;3:e200267.

KEY WORDS artificial intelligence, deep-learning, electrocardiography, risk-prediction

APPENDIX For supplemental methods and figures, please see the online version of this paper.