Automated Diagnostic Reports from Images of Electrocardiograms at the Point-of-Care

Akshay Khunte[1*], Veer Sangha BS[2,3*], Evangelos K Oikonomou MD DPhil[2], Lovedeep S Dhingra MBBS[2], Arya Aminorroaya MD MPH[2], Andreas Coppi PhD,[2,4] Sumukh Vasisht Shankar[2], Bobak J Mortazavi PhD[4,5], Deepak L Bhatt MD MPH[6], Harlan M Krumholz MD SM[2,5,7], Girish N Nadkarni MD MPH[8,9#], Akhil Vaid MD[8,9#], Rohan Khera MD MS[2,4,10#]

[1]Department of Computer Science, Yale University, New Haven, CT
[2]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT
[3]Department of Engineering Science, Oxford University, Oxford, UK
[4]Department of Computer Science & Engineering, Texas A&M University, College Station, TX
[5]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT
[6]Mount Sinai Fuster Heart Hospital, Icahn School of Medicine at Mount Sinai, New York, NY
[7]Department of Health Policy and Management, Yale School of Public Health, New Haven, CT
[8]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[9]The Division of Data Driven and Digital Medicine, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[10]Section of Health Informatics, Department of Biostatistics, Yale School of Public Health, New Haven, CT

*Contributed equally as co-first authors
#Contributed equally as co-senior authors

Correspondence to: Rohan Khera, MD, MS
195 Church St, 6th Floor, New Haven, CT 06510,
203-764-5885; rohan.khera@yale.edu; @rohan_khera

ABSTRACT

Timely and accurate assessment of electrocardiograms (ECGs) is crucial for diagnosing, triaging, and clinically managing patients. Current workflows rely on a computerized ECG interpretation using rule-based tools built into the ECG signal acquisition systems with limited accuracy and flexibility. In low-resource settings, specialists must review every single ECG for such decisions, as these computerized interpretations are not available. Additionally, high-quality interpretations are even more essential in such low-resource settings as there is a higher burden of accuracy for automated reads when access to experts is limited. Artificial Intelligence (AI)-based systems have the prospect of greater accuracy yet are frequently limited to a narrow range of conditions and do not replicate the full diagnostic range. Moreover, these models often require raw signal data, which are unavailable to physicians and necessitate costly technical integrations that are currently limited. To overcome these challenges, we developed and validated a format-independent vision encoder-decoder model – ECG-GPT – that can generate free-text, expert-level diagnosis statements directly from ECG images. The model shows robust performance, validated on 2.6 million ECGs across 6 geographically distinct health settings: (1) 2 large and diverse US health systems- Yale-New Haven and Mount Sinai Health Systems, (2) a consecutive ECG dataset from a central ECG repository from Minas Gerais, Brazil, (3) the prospective cohort study, UK Biobank, (4) a Germany-based, publicly available repository, PTB-XL, and (5) a community hospital in Missouri. The model demonstrated consistently high performance (AUROC≥0.81) across a wide range of rhythm and conduction disorders. This can be easily accessed via a web-based application capable of receiving ECG images and represents a scalable and accessible strategy for generating accurate, expert-level reports from images of ECGs, enabling accurate triage of patients globally, especially in low-resource settings.

MAIN

Electrocardiography (ECG) is a widely available, first-line, noninvasive tool for diagnosing, triaging, and managing cardiovascular disease.[1] Traditional workflows often rely on computerized ECG interpretation algorithms to generate preliminary reads, which, despite limited accuracy,[2,3] provide diagnostic support and enable faster triage for high-risk conditions.[4] Such algorithms, however, are often proprietary and require raw signal data. This makes computerized pre-reads inaccessible to clinicians in rural and low-resource settings, a disparity further exacerbated by the lower availability of expert-level readers.[5,6] This lack of automated system-generated ECG reports in many low-resource settings globally highlights the need for an accurate, easily accessible ECG interpretation tool.

Though recent advances in deep learning enable accurate classification of specific ECG abnormalities,[7–10] they are generally limited to a select number of commonly encountered abnormalities and do not address less common conduction and rhythm disorders or variations within common rhythms. Additionally, while our prior work has demonstrated that diagnostic models can directly identify information from ECG images,[10,11] most traditional signal-based models, like computerized interpretation algorithms, predominantly rely on raw signal data, limiting their scalability to the point-of-care and across low-resource settings. The development of models exclusively for images is challenged by variations in the layout of the leads, labeling of the leads, structure and design of the graphed background, and the quality of the image acquisition.

In this study, we report the development of ECG-GPT (**Figure 1**), a novel vision-text transformer model capable of generating diagnostic reports from ECG images regardless of the layout, trained against the full breadth of expert-verified ECG interpretations across 1.2 million

ECGs collected over 21 years in a large US-based hospital system. ECG-GPT can be accessed as a web-based application that can receive ECG images across formats and layouts as the only input and generate diagnostic reports (demonstration hosted at https://www.cards-lab.org/ecg-gpt).

It demonstrates consistently high discriminative performance across the range of conduction and rhythm disorders, validated in 2.6 million ECGs across temporally and geographically distinct datasets, including a separate US-based major referral hospital system, a Brazil-based large telehealth network, a UK-based prospective cohort study, a publicly available ECG dataset from Germany, and a rural US-based community hospital. By enabling direct inference from ECG images, ECG-GPT can be directly integrated into existing clinical workflows across low-resource settings to help triage care and provide diagnostic support without requiring significant system-wide investments.

RESULTS

Designing ECG-GPT: a vision-text transformer-based architecture

We replicated the traditional ECG interpretation by expert clinicians who rely on visual evaluation of ECG images in their practice, pursuing a systematic approach to define a series of distinct rhythm, rate, conduction, ischemic, and other abnormalities summarized as an unstructured diagnostic impression.[12] To reflect this pattern, we built a generative vision-text transformer model, ECG-GPT, to enable complete, free-text diagnosis statement generation from ECG images in any format and lead layout. This algorithm relied on a custom vision encoder-decoder model architecture built on the backbone of a BEiT vision transformer encoder and a GPT2 decoder.[13,14] In designing ECG-GPT, we chose BEiT for its ability to capture complex

visual features in ECG images, akin to expert clinicians' visual assessments, and chose GPT-2 for its lightweight and efficient nature for text generation given the narrower vocabulary diagnostic text. This ensured a balance between computational depth and operational efficiency essential for deployment.

For the model, a linear sequence of 16x16 pixel patches extracted from each ECG image input, described in further detail below, represented the input for the BEiT transformer model, which produces a lower-dimensional representation of the image that the pre-trained GPT2 decoder uses to generate the diagnosis statement.

Training ECG-GPT

Data source and population: We developed ECG-GPT in a set of 1,162,727 12-lead ECG recordings with accompanying cardiologist-confirmed diagnosis statements performed on 328,891 unique patients at the Yale-New Haven Health System (YNHHS) between 2000 and 2021. These ECGs reflected a wide and balanced distribution of demographics, with a mean age of 67.6 (SD 17.2) years at the time of the ECG. In the development population, 556,565 (47.9%) of the ECGs were obtained among women, 823,421 (70.8%) ECGs were from non-Hispanic White, 151,387 (13.0%) non-Hispanic Black, 129,346 (11.1%) Hispanic, 16,298 (1.4%) non-Hispanic Asian, and 42,275 (3.6%) from patients from other racial backgrounds (**Table S1**), reflecting broad and diverse representation.

To enable format-independent inference that would generalize to real-world images, we took raw ECG signal data from these recordings and plotted ECG images in several distinct lead configurations and layouts. These plotting schemes, which we have developed and extensively validated in previous studies,[10,11] included standard, two-rhythm, shuffled, and alternative

formats, reflecting common and uncommon variations in ECG image printouts (**Figure 1**, see **Methods** for details). The model randomly selected a format each time an ECG was loaded during training to reduce overfitting. We also applied random image rotations between -10 and +10 degrees before training to further mimic variations seen in real-world ECG images. The handling of additional acquisition artifacts was addressed during model deployment (details in **Methods**) to enable broad utility on real-world ECG images.

Diagnostic statement processing and label definitions: We developed a standardized, rule-based approach to process each cardiologist-confirmed diagnostic statement before training. This included removing all references to dates, prior ECGs, providers, and patients. The 100 most prevalent acronyms and misspellings were expanded and corrected, respectively, based on a review by expert cardiologists (**Table S2**). Lastly, a single designation was chosen for disorders with multiple synonymous designations (**Figure S1**). After processing, the diagnostic statements had a median length of 110 characters (IQR 72, 154). These processed diagnosis statements had a median length of 28 tokens (IQR 19, 38) after tokenization using the GPT2 tokenizer. To align the model's evaluation with its intended use, we demonstrate the performance of generated text on extracted labels for 20 distinct conditions extracted from the processed diagnosis statements using a rule-based approach (see **Methods** for details, **Table S3**). Two cardiologists in discussion selected the conditions. Of note, the diagnosis statements represent the full range of diagnoses, and these conditions were chosen strictly for model evaluation.

Using this approach, we report metrics for a few ECG abnormalities. In the development cohort, most ECGs reported normal sinus rhythm (788,999, 67.8%), followed by left ventricular hypertrophy (126,795, 10.9%), left atrial enlargement (121,654, 10.5%), and atrial fibrillation

(AF) (111,533, 9.6%). A total of 89,675 (7.7%) and 66,483 (5.7%) ECGs reported sinus tachycardia (ST) and sinus bradycardia (SB), respectively. Right bundle branch block (RBBB) and left bundle branch block (LBBB) were present in 83,810 (7.2%) and 40,790 (3.5%) ECGs, respectively, and 101,836 (8.76%) ECGs reported atrioventricular block (AVb). The proportion of ECGs with the 20 conditions, spanning rhythm and conduction disorders extracted from the diagnosis statements is listed in **Table S1**.

Evaluating Model Performance

Natural language generation metrics and structured label evaluation: As detailed below, we employed three distinct approaches to assess the performance of ECG-GPT for generating comprehensive, clinically accurate diagnosis statements. First, we used a fine-tuned DistilBERT model to quantify the semantic similarity between reference and model-generated diagnosis statements.[15–17] Next, we employed natural language generation (NLG) metrics conventionally used to evaluate models designed for image captioning or translation tasks to assess the syntactic similarity between statements. These included ROUGE, BLEU, and METEOR, ranging from 0 to 1, and CIDEr, ranging from 0 to 5, with higher values indicating better overlap between reference and generated text.[18–21] Third, we performed a clinical assessment to evaluate the diagnostic accuracy of the model-generated statements for 20 key conditions spanning a wide range of rhythm and conduction disorders. Each diagnostic label was extracted from the reference and model-generated diagnosis statements using a rule-based approach.

Internal testing – Semantic similarity: We fine-tuned a clinically pretrained DistilBERT language model in the same set of diagnosis statements used to train the vision-text transformer

model (see **Methods** for details). This language model was then used to generate embeddings for each reference and model-generated statements for each ECG in the 146,096 ECGs in the internal held-out test set, obtained from 40,827 patients who had not contributed any data to the training set (**Table S1**). The median cosine similarity between the embeddings for reference statements and their paired model-generated statements was 0.93 (IQR 0.83-1.00). This result was significantly higher than the median cosine similarity between 100,000 randomly selected combinations of reference and model-generated statements (0.69 (IQR 0.62-0.77, $p<0.001$).

We also performed a secondary analysis to assess the model's ability to capture individual conditions within their full clinical context. For each of the 20 diagnostic labels extracted from the reference statements, we generated a subset with all ECGs flagged as positive for that condition. For each subset, we computed the pairwise cosine similarity between embeddings for reference and model-generated statements and the baseline similarity between random combinations of reference and model-generated statements within the subset. Across all subsets, the median pairwise cosine similarity was significantly greater than the respective median random cosine similarity (**Table 1**). Across all 20 conditions, pairwise and baseline similarities ranged from 0.85-0.96 and 0.74-0.84, respectively. For key rhythm disorders – AF, ST, SB, premature atrial complexes (PACs), and premature ventricular complexes (PVCs) – pairwise and random similarities ranged from 0.87-0.96 and 0.76-0.84, respectively. For conduction abnormalities –LBBB, RBBB, AVb, left anterior fascicular block (LAFB), and left posterior fascicular block (LPFB) – pairwise and random similarities ranged from 0.87-0.96 and 0.76-0.84, respectively.

Internal testing – NLG agreement: For NLG metrics, ECG-GPT matches or outperforms most state-of-the-art medical image captioning models.[22–24] For ROUGE scores, which measure the overlap of word sequences between the generated and reference diagnosis statements, emphasizing recall, we report scores of 0.748 and 0.742 for ROUGE-1 and ROUGE-L, respectively. For BLEU scores, which focus on precision and assess the quality of model-generated statements, we report scores ranging from 0.619 for BLEU-1 to 0.472 for BLEU-4. We also report a METEOR score of 0.750, indicating substantive agreement in both the word usage and order of model-generated and reference diagnosis statements and a CIDEr score of 4.69, demonstrating that the model-generated statements closely matched the diversity of the language used in the reference statements (**Table S4**).

Internal testing – Structured label assessment: Model performance in the held-out test set for each of the 20 rhythm and conduction disorders, including accuracy, positive and negative predictive values, specificity, sensitivity, AUROC, AUPRC, and F1 scores, are recorded in **Table 2**. For AF, ST, SB, PACs, and PVCs, AUROCs and AUPRCs ranged from 0.87-0.97 and 0.58-0.87, respectively. For LBBB, RBBB, AVb, LAFB, and LPFB, the AUROCs and AUPRCs ranged from 0.88-0.96 and 0.30-0.85, respectively. Across all 20 conditions, diagnostic accuracy ranged between 0.95-0.99.

External Validation

Testing in an independent hospital-based system (Mount Sinai Health System): To assess the external validity of our algorithm and account for possible variations in the interpretation and recording of diagnostic statements, we deployed our computational platform to a library of

1,434,455 ECGs drawn from a geographically distinct large hospital system, Mount Sinai Health System (MSHS) in New York, using a federated validation approach (described in **Methods**). This dataset also consisted of 12-lead ECG data with paired free-text diagnosis statements, enabling the evaluation of our training and validation pipeline in a large population. Consistent with our internal test findings, ECG-GPT reported high semantic similarity between the model-generated statements and the MSHS reference statements and, in structured label assessment, maintained robust performance across the 20 rhythm and conduction disorders.

For semantic similarity, embeddings had a median pairwise similarity of 0.86 (IQR 0.78-0.94), significantly greater than the median baseline similarity of 0.73 among 2 random statements (IQR 0.66-0.79, $p<0.001$). This separation persisted across the 20 subsets corresponding to each extracted rhythm and conduction disorder (**Table 3**). For key rhythm disorders – AF, ST, SB, PACs, and PVCs – pairwise and baseline similarity ranged from 0.83-0.89 and 0.75-0.81, respectively. For key conduction abnormalities – LBBB, RBBB, AVb, LAFB, LFPB – pairwise and baseline similarity ranged from 0.87-0.93 and 0.78-0.82, respectively.

The model performed well in clinical assessment across the 20 extracted labels. For AF, ST, SB, PACs, and PVCs, AUROCs and AUPRCs ranged from 0.80-0.95 and 0.48-0.78, respectively. For LBBB, RBBB, AVb, LAFB, and LFPB, AUROCs, and AUPRCs ranged from 0.71-0.97 and 0.06-0.78, respectively. Model performance in the external MSHS validation dataset across all 20 conditions is reported in **Table 4**.

Testing across geographically distinct open-source datasets: Next, we assessed the performance of ECG-GPT across four publicly available ECG datasets, thus providing benchmarks for prior

and future models. These datasets had more limited coverage of diagnostic flags, with 6 conditions spanning rhythm and conduction disorders available across these datasets (**Table S5**). We generated full diagnosis reports for these ECGs and evaluated performance on the available AF, ST, SB, LBBB, RBBB, and AVb conditions. This was done by extracting these labels from the ECG-GPT-generated diagnosis statements using the rule-based approach described above. The model's diagnostic performance in each of these external validation sets is noted in **Table 5**.

First, in a randomly selected sample of 1,000,000 ECGs from the previously described CODE15 dataset collected by the Telehealth Network of Minas Gerais (TNMG), Brazil, between 2010 and 2017.[7,25] Here, AUROCs for rhythm disorders were 0.93, 0.94, and 0.93 for AF, ST, and SB, respectively, with similar performance for conduction abnormalities (0.91, 0.95, and 0.89, for LBBB, RBBB, AVb, respectively).

When deployed to a smaller, cardiologist-validated dataset collected by TNMG in Brazil between April and September 2018, consisting of 827 ECGs manually annotated by two cardiologists with disputes resolved by a third,[10] AUROCs were higher across nearly all diagnostic labels. The model reported AUROCs of 0.96, 0.96, and 0.91 for AF, ST, and SB, respectively. For LBBB, RBBB, and AVb, the model reported AUROCs of 0.97, 0.98, and 0.89, respectively.

The third of these four open-source datasets consisted of 45,389 ECGs obtained from patients enrolled in the UK Biobank, a prospective study with protocolized testing in the community outside the context of clinical evaluation. Here, the model had AUROCs ranging from 0.92 to 0.99 (**Table 5**), thus highlighting the reproducibility of our approach across clinical and non-clinical settings.

Lastly, to better understand the generalizability of our tools across distinct temporal and geographical settings reflecting different acquisition systems, we also evaluated ECG-GPT in the Germany-based PTB-XL dataset, which consists of 21,784 ECGs obtained between 1989 and 1996.[26] Here, ECG-GPT maintained robust performance, with AUROCs of 0.94, 0.96, and 0.86 for AF, ST, and SB, respectively, and 0.98, 0.99 and 0.85 for LBBB, RBBB, and AVb respectively.

External Validation Using Real-World ECG Images: Finally, to further illustrate the robustness of ECG-GPT against real-world images, we further report its performance in a real-world dataset of 64 ECG images collected at the Lake Regional Hospital (LRH) system in Missouri. In these ECGs, the model had AUROCs of 0.99, 0.90, and 0.79 for AF, ST, and SB, respectively. For LBBB, RBBB, and AVb, the model reported AUROCs of 1.00, 1.00, and 0.87, respectively.

ECG-GPT as a web-based tool: To demonstrate the potential utility of a platform capable of receiving ECG images and generating reports, ECG-GPT is publicly available through an online, interactive platform. This web-based application is for research use (https://www.cards-lab.org/ecg-gpt) - it demonstrates the potential use via incorporating quality control on uploaded ECGs (based on a prior approach, see **Methods** for details),[11] and generating full-text reports.

DISCUSSION

We describe the development and external validation of ECG-GPT, a first-of-its-kind AI pipeline that enables the direct generation of automated, complete diagnosis statements from images of ECGs in any format. The model performs well against clinician-certified reports across various

natural language generation metrics and diagnostic labels spanning a wide array of conduction, rhythm, and structural heart disorders. The model's scalability is further supported by its robust performance across a range of demographically, temporally, and geographically distinct cohorts, its online deployment, and the capacity to containerize and deploy the model without sharing data in a federated approach.

Our work on format-independent, image-based ECG captioning represents a novel development for vision-text machine learning models. Our model significantly outperforms prior implementations for generating free-text reports from ECG signals, with CIDEr and METEOR scores of 4.69 and 0.75, respectively, compared with scores of 2.55 and 0.27, respectively, reported in a previous study suggesting high textual consistency between original and generated reports.[24] Machine learning-based multilabel models have been previously developed to simultaneously diagnose large sets of conditions, but these approaches are inherently limited to those labels selected for training and do not capture the full diagnostic range of ECGs.[7,10,27] Moreover, the utility of such signal models is limited to healthcare systems with the resources to store signal data and incorporate models into the clinical workflow. We demonstrate consistent performance across a range of external validation sets for classifying key rhythm and conduction disorders. Of note, the performance for labels in external validation sets where specific labels were explicitly available matches the performance on those labels in prior published reports. The simplicity of this system, based on images, is that there is inherent interoperability and the absence of a requirement to integrate with ECG machines to extract signals. This approach is particularly advantageous in low-resource regions, where ECGs are currently not stored beyond printing ECG images at the time of acquisition.[28] This approach also adds convenience, and

provides access in any venue, including, for example, to emergency medical services providers or in remote locations.

ECG-GPT can be used directly by clinicians at the point of care by uploading ECG images from their phones or as scanned images to a web-based interface, with a demonstration accompanying this study.[29] This applies anywhere that end-users may still lack access to automated reads or require interpretation before a specialist's review. The image-based model can also be more easily integrated into repositories of scanned ECGs, the most prevalent and interoperable format for storing and sharing ECGs. Further, ECG-GPT has a unique combination of diagnostic accuracy and range, demonstrating expert-level performance for key conditions while also retaining the capability to generate statements for rare conditions frequently not captured by standard multi-label models. This feature could make ECG-GPT a tool for generating pre-reads and enabling more efficient triage globally in areas with insufficient access to specialists and computerized ECG interpretation algorithms.

Our study has several limitations. First, while the model accurately diagnosed the selected conditions, it is impossible to determine and thus evaluate the performance for the full extent of possible diagnoses the model could output for a given ECG image due to the size and variety of the corpus of diagnosis statements used for model development. However, we report model performance across various rhythm and conduction disorders of varying severity and prevalence, suggesting that ECG-GPT's performance generalizes to other conditions. Moreover, using the federated approach implemented for external validation, ECG-GPT could be continually fine-tuned to improve performance in individual healthcare systems. This would ensure a reliable pipeline with consistent performance for future ECGs within the specific patient populations in which the model is deployed. Second, while four different formats were used during model

development, we cannot ascertain whether the model generalizes equally well to every other novel ECG image format. However, the model's performance within the dataset from LRH, consisting of ECG images plotted in a distinct configuration from those used in model development, indicates the model can generate accurate diagnosis statements for ECGs in formats not seen during training. Finally, though strictly expert-validated diagnosis statements were used to develop the model, these statements are not always completely accurate, limiting the model's performance. As evidenced by the high diagnostic accuracy of the model in the external set of ECGs manually annotated by two cardiologists, the model may perform better if developed and evaluated in more rigorously validated diagnosis statements. Furthermore, the current practice of clinicians over-reading the computer-generated reads and correcting them without version control precludes the head-to-head assessment of ECG-GPT against computer-generated reads. However, the higher performance in labels assigned by more than one expert suggests that it likely performs at or above the performance of the current computerized reads at the US health systems, especially for the tested diagnoses.

Thus, we have developed and extensively validated a novel vision-text transformer capable of generating complete diagnostic statements from ECG images in any lead layout and configuration. Our approach represents a scalable and accessible strategy for generating accurate, expert-level reports from photos of ECGs, enabling accurate ECG interpretation anywhere that an ECG image, paper or digital, can be produced.

METHODS

The study was reviewed by the Yale Institutional Review Board, which approved the study protocol and waived the need for informed consent as the study represents secondary analysis of existing data. The development dataset from Yale and the validation datasets from Mount Sinai and Lake Regional Hospital are not publicly available, given the stipulations of the relevant Institutional Review Boards. The external datasets from Brazil, UK Biobank, and PTB-XL are available directly from the respective groups and are outside the purview of the authors.

Vision-Encoder Decoder Model Architecture

We built a custom Vision Encoder-Decoder model using the HuggingFace framework.[30] For the image encoder, we selected a BEiT transformer model, pretrained on the ImageNet dataset,[31] due to its robust performance on the ImageNet-1K benchmark, relatively few trainable parameters compared to other state-of-the-art models, and the 384x384 pixel input size.[13] We selected the base-size version of the model, with ~84 million trainable parameters, which takes a linearly embedded sequence of 16x16 pixel patches as the input.

We selected the Generative Pretrained Transfomer-2 (GPT-2) transformer for the text decoder, initially developed for prompt-based text generation. This architecture, which contains ~124 million parameters, decoded the lower-dimensional representations generated from images by the BEiT encoder into diagnosis statements with a maximum output token length of 115 tokens. In addition to the lightweight nature of the model relative to other large language models capable of text generation, the extensive pretraining of the GPT2 decoder enabled direct integration into the vision-text transformer architecture without further fine-tuning.

Overall, the composed Vision Encoder-Decoder model, consisting of the BEiT encoder and GPT-2 decoder, has over 239 million trainable parameters (**Figure 2**). The model was trained at a learning rate of 5 x 10^{-5} for 20 epochs. We used the Adam optimizer, a minibatch size of 14, and a cross-entropy loss function to minimize the error between the GPT2 output and the tokenized reference diagnostic statements to train the model.[32] Model development used the HuggingFace Transformers 4.28.1 framework with PyTorch 2.0.0 and Python 3.11.3 on four RTX 3090 graphics processing units.

Model Development

Data Source: Raw voltage data was collected for all 12-lead ECGs with corresponding diagnosis statements obtained at the YNHHS during 2000-2021. Each ECG was recorded with a sampling frequency of 500 Hz using Philips PageWriter and GE MAC machines. The subset of these ECGs with continuous recording across all 12 leads was then split at the patient level into training, validation, and test sets (85%, 5%, 10%). For each of these sets, we restricted to ECGs with no marked abnormalities and abnormal ECGs with confirmed reports certified by a cardiologist. An exception to this was ECGs with STEMI, since ECGs for STEMI are done in the emergency setting, interpreted at the point-of-care, and may not be listed as a confirmed read in the system. In the training set, 150,921 ECGs with no marked abnormalities were randomly removed to match the prevalence of such ECGs in the original cohort (13.6%) (**Figure S2**).

Signal Preprocessing: First, all ECGs that did not contain 10 seconds of continuous recording across all 12 leads were excluded. To enable the generation of ECG images like those used in clinical settings, we further preprocessed the signal before plotting. For this, we subtracted a one-

second median filter from the original raw voltage for each lead of 10-second ECGs to remove baseline wander, mirroring the approach undertaken by ECG machines before printing clinical ECGs available to and interpreted by physicians.

All ECGs obtained in the YNHHS were plotted at their original sampling rate of 500 Hz. ECGs used for external validation, which were recorded at sampling frequencies between 300-600 Hz, were down sampled to 300 Hz before plotting, as described previously.[10]

ECG Image Generation: The preprocessed ECG signals were transformed into ECG images at 100 dots per inch (DPI) using the python library ecg-plot.[33] We employed multiple strategies to ensure the robustness of the model.

First, we converted each signal waveform to multiple images using four different layouts of the leads to account for different schemes of real-world ECGs. To ensure the model is resilient to these different formats, it was trained using all these variations of ECG images, an approach we previously developed and validated.[10] The four formats used in model development (**Figure 1**) included (1) The standard printed ECG format in the US with lead I as the rhythm strip. This format consists of four 2.5-second sequential columns, each containing a 2.5-second strip from three leads. (2) The same as the standard US format that includes an additional rhythm strip from lead II. (3) An alternate format with no rhythm strip comprising two 5-second columns. The first column represents a simultaneous recording from limb leads, while the second column represents a simultaneous recording from precordial leads. (4) A shuffled format in which precordial leads are recorded in the first two columns and limb leads are represented in the last two columns.

Second, the conversion of ECG signals to images was done independently of the model development to ensure the model remains agnostic to preprocessing steps from ECG signals to

images. Third, all images were rotated by a random amount between -10 to 10 degrees prior to training. Finally, we used Python Image Library (PIL v9.2.0) to convert all ECG images to greyscale and down-sample them to the required size for input into the model regardless of their initial resolution.

Image Standardization for Model Inference: We have implemented a previously validated approach to standardize model inputs.[11] First, inputs are limited to 12-lead ECG tracings which are vertically oriented, minimally rotated, have a uniform background, and do not have peripheral annotations. Additionally, to mitigate the effects of noise, a two-step preprocessing approach is applied to each image: first, the image is straightened and cropped to correct for rotations and to remove any elements outside of the ECG tracing. Second, the algorithm scales the brightness and contrast of the ECGs to the mean values of the development population before generating model predictions. ECGs with deviations in brightness and contrast 50% greater or lower than those seen in the development set are flagged as requiring the image to be recaptured in better quality before inference.

Diagnostic Statement Preprocessing: Cardiologist-confirmed diagnosis statements were preprocessed to remove all identifying information using a rule-based approach. A string search was performed for all diagnostic statements to identify and remove names, references to previous ECGs, and all dates and times. The Python PyEnchant package was then used to generate a list of all abbreviations present in the processed diagnosis statements.[34] A pair of clinical experts manually generated a dictionary containing each abbreviation and its expanded form for the 100 most common abbreviations (**Table S2**). Each instance of these abbreviations in the diagnosis

statement was then replaced with its expansion. Additionally, the most common misspellings and all synonyms for a condition were replaced with a single term. This processed diagnosis statement was then used for model development and evaluation (**Figure S1**).

Data Sources for External Validation

Independent hospital-based system (Mount Sinai Health System): To further evaluate the ability to generalize to external data sources, five ECG signal datasets acquired outside the YNHHS were used to validate model performance. First, to externally assess ECG-GPT's performance in free-text diagnostic statements, we deployed the model in a set of 1,434,455 ECGs with corresponding diagnosis statements collected at the MSHS from 2013 to 2023.

A federated approach was implemented to enable external validation within the MSHS. The model was containerized using Docker and securely deployed within the hospital's infrastructure. It accepted file paths as input, ensuring that patient data remained within the hospital's system without the need for external data sharing. This approach facilitated accurate prediction generation while safeguarding patient privacy.

Geographically distinct open-source datasets: To further evaluate the model's performance on six key rhythm and conduction disorders, including AF, ST, SB, LBBB, RBBB, and AVb, we deployed the model in four distinct open-source ECG datasets. First, we obtained 45,389 ECGs from the UK Biobank, under research application #71033, to pursue external validation of our model. UK Biobank represents the largest population-based cohort of 502,468 people in the United Kingdom with protocolized imaging, laboratory testing, and linked electronic health records.

We also used a set of 1,000,000 ECGs randomly sampled from the CODE15 study dataset, a set of 2,322,513 ECG recordings previously used for both signal- and image-based multilabel ECG classification models.[7,10,25] As the primary CODE15 dataset consists of ECGs collected and annotated for six rhythm and conduction disorders by individual clinicians during routine care, we also deployed the model in a secondary cardiologist-validated dataset. This dataset contained 827 additional ECGs collected in the Telehealth Network of Minas Gerais between April and September 2018.[7,10] For each of these ECGs, annotations for the six rhythm and conduction disorders were made by two independent cardiologists following criteria from the American Heart Association,[35] with disagreements resolved by a third cardiologist.

Finally, the model was also validated on PTB-XL, a previously described dataset of ECGs.[26] This dataset contains 21,837 10-second, 12-lead recordings collected at 500 Hz from 18,885 patients in Germany between 1989 and 1996. The records for each ECG, including diagnostic, form, and rhythm statements, were used to extract labels for the same set of rhythm and conduction disorders.

Real-World ECG Images: We pursued external validation on a real-world ECG image dataset to evaluate the model's performance when applied to ECG images plotted independent of our ECG preprocessing and plotting pipeline. This dataset consisted of 64 ECG images obtained at the LRH System in Osage Beach, MO.[10] This dataset included 8-10 ECGs with each of the six rhythm and conduction disorders assessed with the other external validation datasets and ECGs labeled as normal.

Though the layout of these ECGs was similar to the standard layout used for model development, there were multiple key distinctions. First, the V1 lead, instead of lead I, was the

rhythm lead. Second, the signal was black, as opposed to blue, and vertical lines were separating the leads. Additionally, there were variations in background and grid color, as well as in the position and font of the lead label.

Model Evaluation

Label extraction: Clinical labels for 20 conditions, spanning key rhythm and conduction disorders selected by two cardiologists, were extracted from each reference and model-generated diagnosis statement using a standardized string search approach. For each condition, basic string search was performed to extract each label using a set of strings (**Table S3**). A condition was flagged as positive if the diagnosis statement contained a full match for any string in the set. ECGs were flagged as negative if there was no match or if the match was preceded by a negation, including “no” or “without”.

Semantic Similarity: To evaluate the semantic similarity between original and model-generated diagnosis statements, we fine-tuned a lightweight DistilBERT model,[15,36] pretrained on a large corpus of electronic health record notes,[16,17] in the same set of cardiologist-confirmed diagnosis statements used to train the vision-text transformer model. The training mirrored the standard approach for training masked language models, with a chunk size of 128 tokens and a masking probability of 0.15. The model was trained at a 2×10^{-5} learning rate until validation loss did not improve for three consecutive epochs. We used the Adam optimizer, a minibatch size of 16, and a cross-entropy loss function to minimize the error between the masked and generated tokens to train the model.[32] Model development used the HuggingFace Transformers 4.28.1 framework with Torch 2.0.0 and Python 3.11.3 on four RTX 3090 graphics processing units.

After fine-tuning, we deployed the masked language model in the held-out test set and the MSHS external validation set to generate 768-dimensional embeddings for each reference and model-generated diagnosis statement. We used an identical federated approach to the model-generated statements to generate embeddings for the diagnosis statements within the MSHS. Pairwise similarity was computed as the median cosine similarity between the embeddings for each reference diagnosis statement and its paired model-generated statement. Baseline similarity was computed as the median cosine similarity between the embeddings for 100,000 random pairings of reference and model-generated statements.

To assess the model's ability to diagnose specific conditions within their complete clinical context, we created subsets for each of the 20 diagnostic labels identified from the reference statements, consisting of all ECGs marked positive for that condition. Pairwise and baseline similarity were computed identically to the approach used for the complete datasets. For subsets too small to generate 100,000 random pairings, the baseline similarity was reported as the median cosine similarity between embeddings for all possible reference and model-generated statement pairings within the subset.

Syntactic Similarity: Four conventional NLG metrics assessed the syntactic similarity between the original diagnosis reports and generated text. ROUGE and BLEU, which range from 0 to 1, evaluate recall and precision, respectively, for the overlap of n-grams between generated and reference statements, providing insight into content overlap and coherence.[18,19] METEOR, which ranges from 0 to 1, incorporates both syntactic and semantic similarity by aligning word stems and synonyms, enabling an evaluation of content relevance in addition to word overlap.[20] Finally, CIDEr, which ranges from 0 to 5, measures consensus between generated text and

reference summaries through similarity to human consensus, enhancing evaluation robustness across various linguistic styles.[21] Collectively, these metrics offer a comprehensive assessment of the syntactic similarity between original diagnosis statements and model-generated statements. Each of these metrics was deployed in both the internal held-out test set using the HuggingFace Evaluate package for computing ROUGE, BLEU, and METEOR scores and the COCO Caption Evaluation package for computing CIDEr scores, respectively.[37,38]

Structured Label Assessment: We also implemented a secondary analysis of model performance using the 20 extracted labels. For each reference and model-generated statement, we computed agreement between the statements for each of the 20 rhythm and conduction disorders.
The open-source datasets, which each contained reference labels for AF, ST, SB, RBBB, LBBB, and AVb, were compared to labels extracted from the model-generated diagnosis statements for these ECGs.

We used the area under the receiver operating characteristic (AUROC) to measure model discrimination. 95% confidence intervals for AUROC were calculated using DeLong's algorithm.[39,40] We also assessed the area under precision-recall curve (AUPRC), accuracy, sensitivity, specificity, F1 score, positive predictive value (PPV), and negative predictive value (NPV). We employed the bootstrap resampling method to estimate confidence intervals for AUPRC.[41]

Statistical Analysis

Summary statistics are presented as counts and percentages for categorical elements and median and interquartile range (IQR) for continuous elements. A paired t-test was used to compute the

probability of overlap between the pairwise and baseline cosine similarity of reference and model-generated statements. All analyses were performed using Python 3.11.3, and the significance level was set at an alpha of 0.05.

DATA AVAILABILITY

The development dataset from Yale and the validation datasets from Mount Sinai and Lake Regional Hospital are not publicly available, given the stipulations of the relevant Institutional Review Boards. The external datasets from Brazil, UK Biobank, and PTB-XL are available directly from the respective groups and are outside the purview of the authors.

ACKNOWLEDGEMENTS

Author contributions: RK conceived the study and accessed the data. AK, VS, and RK developed the model. AK, VS, and RK pursued the statistical analysis. AK, VS, and LSD drafted the manuscript. All authors provided feedback regarding the study design and made critical contributions to the manuscript writing. RK supervised the study, procured funding, and is the guarantor.

Funding: This study was supported by research funding awarded to Dr. Khera by the Yale School of Medicine and grant support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award K23HL153775) and the Doris Duke Charitable Foundation (under award, 2022060). Dr. Oikonomou receives support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award 1F32HL170592). The funders had no role in the design and conduct of the study; collection, management, analysis, and

interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Competing Interests: Mr. Khunte, Mr. Sangha, and Dr. Khera are the coinventors of U.S. Provisional Patent Application No. 63/428,569. Mr. Sangha and Dr. Khera are the coinventors of U.S. Pending Patent Application No. 63/346,610, and are co-founders of Ensight-AI with Dr. Krumholz. Dr. Khera is the coinventor of U.S. Provisional Patent Application No. 63/177,117 (unrelated to current work) and is a co-founder of Evidence2Health, a precision health platform for evidence-based care. He is also an associate editor at JAMA, and received support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award K23HL153775) and the Doris Duke Charitable Foundation (under award, 2022060). He also receives research support, through Yale, from Bristol-Myers Squibb, Novo Nordisk, and BridgeBio. Dr. Oikonomou receives support from the National Heart, Lung, and Blood Institute of the National Institutes of Health (under award 1F32HL170592). He is an academic co-founder of Evidence2Health LLC, a co-inventor in patent applications (US17/720,068, 63/177,117, 63/580,137, 63/606,203, WO2018078395A1, WO2020058713A1) and has been an ad hoc consultant for Caristo Diagnostics Ltd. Dr. Nadkarni is a founder of Renalytix, Pensieve, and Verici and provides consultancy services to AstraZeneca, Reata, Renalytix, Siemens Healthineer, and Variant Bio, and serves a scientific advisory board member for Renalytix and Pensieve. He also has equity in Renalytix, Pensieve, and Verici. Dr. Krumholz works under contract with the Centers for Medicare & Medicaid Services to support quality measurement programs, was a recipient of a research grant from Johnson & Johnson, through Yale University, to support clinical trial data sharing; was a recipient of a research agreement, through Yale University, from the Shenzhen Center for Health Information for work to advance intelligent disease prevention

and health promotion; collaborates with the National Center for Cardiovascular Diseases in Beijing; receives payment from the Arnold & Porter Law Firm for work related to the Sanofi clopidogrel litigation, from the Martin Baughman Law Firm for work related to the Cook Celect IVC filter litigation, and from the Siegfried and Jensen Law Firm for work related to Vioxx litigation; chairs a Cardiac Scientific Advisory Board for UnitedHealth; was a member of the IBM Watson Health Life Sciences Board; is a member of the Advisory Board for Element Science, the Advisory Board for Facebook, and the Physician Advisory Board for Aetna; and is the co-founder of Hugo Health, a personal health information platform, and co-founder of Refactor Health, a healthcare AI-augmented data management company, and Ensight-AI, Inc. All other authors declare no relevant competing interests. Dr. Bhatt discloses the following relationships - Advisory Board: Angiowave, Bayer, Boehringer Ingelheim, CellProthera, Cereno Scientific, Elsevier Practice Update Cardiology, High Enroll, Janssen, Level Ex, McKinsey, Medscape Cardiology, Merck, MyoKardia, NirvaMed, Novo Nordisk, PhaseBio, PLx Pharma, Stasys; Board of Directors: American Heart Association New York City, Angiowave (stock options), Bristol Myers Squibb (stock), DRS.LINQ (stock options), High Enroll (stock); Consultant: Broadview Ventures, GlaxoSmithKline, Hims, SFJ, Youngene; Data Monitoring Committees: Acesion Pharma, Assistance Publique-Hôpitaux de Paris, Baim Institute for Clinical Research (formerly Harvard Clinical Research Institute, for the PORTICO trial, funded by St. Jude Medical, now Abbott), Boston Scientific (Chair, PEITHO trial), Cleveland Clinic, Contego Medical (Chair, PERFORMANCE 2), Duke Clinical Research Institute, Mayo Clinic, Mount Sinai School of Medicine (for the ENVISAGE trial, funded by Daiichi Sankyo; for the ABILITY-DM trial, funded by Concept Medical; for ALLAY-HF, funded by Alleviant Medical), Novartis, Population Health Research Institute; Rutgers University (for the NIH-funded MINT

Trial); Honoraria: American College of Cardiology (Senior Associate Editor, Clinical Trials and News, ACC.org; Chair, ACC Accreditation Oversight Committee), Arnold and Porter law firm (work related to Sanofi/Bristol-Myers Squibb clopidogrel litigation), Baim Institute for Clinical Research (formerly Harvard Clinical Research Institute; RE-DUAL PCI clinical trial steering committee funded by Boehringer Ingelheim; AEGIS-II executive committee funded by CSL Behring), Belvoir Publications (Editor in Chief, Harvard Heart Letter), Canadian Medical and Surgical Knowledge Translation Research Group (clinical trial steering committees), CSL Behring (AHA lecture), Cowen and Company, Duke Clinical Research Institute (clinical trial steering committees, including for the PRONOUNCE trial, funded by Ferring Pharmaceuticals), HMP Global (Editor in Chief, Journal of Invasive Cardiology), Journal of the American College of Cardiology (Guest Editor; Associate Editor), K2P (Co-Chair, interdisciplinary curriculum), Level Ex, Medtelligence/ReachMD (CME steering committees), MJH Life Sciences, Oakstone CME (Course Director, Comprehensive Review of Interventional Cardiology), Piper Sandler, Population Health Research Institute (for the COMPASS operations committee, publications committee, steering committee, and USA national co-leader, funded by Bayer), WebMD (CME steering committees), Wiley (steering committee); Other: Clinical Cardiology (Deputy Editor); Patent: Sotagliflozin (named on a patent for sotagliflozin assigned to Brigham and Women's Hospital who assigned to Lexicon; neither I nor Brigham and Women's Hospital receive any income from this patent); Research Funding: Abbott, Acesion Pharma, Afimmune, Aker Biomarine, Alnylam, Amarin, Amgen, AstraZeneca, Bayer, Beren, Boehringer Ingelheim, Boston Scientific, Bristol-Myers Squibb, Cardax, CellProthera, Cereno Scientific, Chiesi, CinCor, Cleerly, CSL Behring, Eisai, Ethicon, Faraday Pharmaceuticals, Ferring Pharmaceuticals, Forest Laboratories, Fractyl, Garmin, HLS Therapeutics, Idorsia, Ironwood,

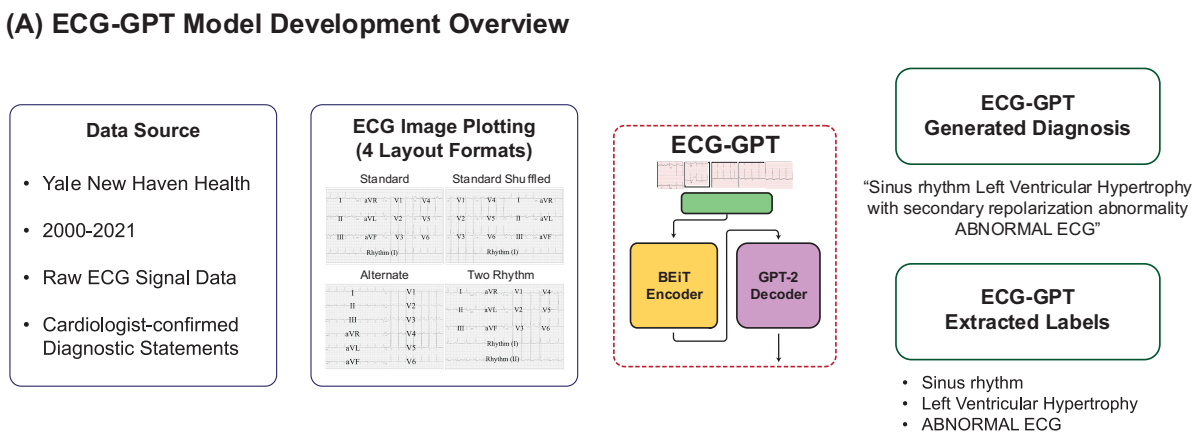
Ischemix, Janssen, Javelin, Lexicon, Lilly, Medtronic, Merck, Moderna, MyoKardia, NirvaMed, Novartis, Novo Nordisk, Otsuka, Owkin, Pfizer, PhaseBio, PLx Pharma, Recardio, Regeneron, Reid Hoffman Foundation, Roche, Sanofi, Stasys, Synaptic, The Medicines Company, Youngene, 89Bio; Royalties: Elsevier (Editor, Braunwald's Heart Disease); Site Co-Investigator: Abbott, Biotronik, Boston Scientific, CSI, Endotronix, St. Jude Medical (now Abbott), Philips, SpectraWAVE, Svelte, Vascular Solutions; Trustee: American College of Cardiology; Unfunded Research: FlowCo.

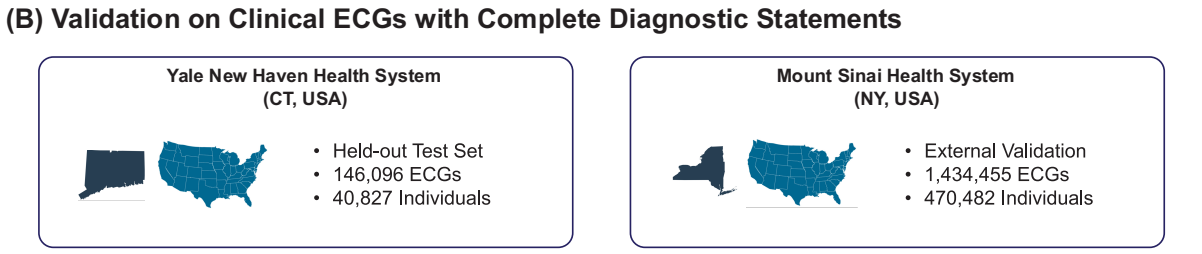
REFERENCES

1. Schlant, R. C. *et al.* Guidelines for electrocardiography. A report of the American college of cardiology/American heart association task force on assessment of diagnostic and therapeutic cardiovascular procedures (committee on electrocardiography). *Circulation* **85**, 1221–1228 (1992).
2. Shah, A. P. & Rubin, S. A. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *J. Electrocardiol.* **40**, 385–390 (2007).
3. Guglin, M. E. & Thatai, D. Common errors in computer electrocardiogram interpretation. *Int. J. Cardiol.* **106**, 232–237 (2006).
4. Schläpfer, J. & Wellens, H. J. Computer-interpreted electrocardiograms: Benefits and limitations. *J. Am. Coll. Cardiol.* **70**, 1183–1192 (2017).
5. Jones, C., Park, T. & United Economic Research Service (Ers). *Health status and health care access of farm and rural populations*. (Bibliogov, 2012).
6. Aneja, S. *et al.* US cardiologist workforce from 1995 to 2007: modest growth, lasting geographic maldistribution especially in rural areas. *Health Aff. (Millwood)* **30**, 2301–2309 (2011).
7. Ribeiro, A. H. *et al.* Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat. Commun.* **11**, 1760 (2020).
8. Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
9. Hughes, J. W. *et al.* Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA Cardiol.* **6**, 1285–1295 (2021).
10. Sangha, V. *et al.* Automated multilabel diagnosis on electrocardiographic images and signals. *Nat. Commun.* **13**, 1583 (2022).
11. Sangha, V. *et al.* Detection of left ventricular systolic dysfunction from electrocardiographic images. *Circulation* (2023) doi:10.1161/CIRCULATIONAHA.122.062646.
12. Sattar, Y. & Chhabra, L. Electrocardiogram. in *StatPearls* (StatPearls Publishing, 2023).
13. Bao, H., Dong, L., Piao, S. & Wei, F. BEiT: BERT Pre-Training of Image Transformers. *arXiv [cs.CV]* (2021).
14. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. https://life-extension.github.io/2020/05/27/GPT%E6%8A%80%E6%9C%AF%E5%88%9D%E6%8E%A2/language-models.pdf.
15. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv [cs.CL]* (2019).
16. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings. *arXiv [cs.CL]* (2019).
17. Wang, G. *et al.* Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat. Med.* **29**, 2633–2642 (2023).
18. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. in *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, 2004).
19. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. https://aclanthology.org/P02-1040.pdf.
20. Banerjee, S. & Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* 65–72 (Association for Computational Linguistics, 2005).

21. Vedantam, R., Zitnick, C. L. & Parikh, D. CIDEr: Consensus-based image description evaluation. *arXiv [cs.CV]* 4566–4575 (2014).
22. Selivanov, A. *et al.* Medical image captioning via generative pretrained transformers. *Sci. Rep.* **13**, 4171 (2023).
23. Ayesha, H. *et al.* Automatic medical image interpretation: State of the art and future directions. *Pattern Recognit.* **114**, 107856 (2021).
24. Bartels, M. G. G. *et al.* Learning to Automatically Generate Accurate ECG Captions. in *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning* (eds. Konukoglu, E. et al.) vol. 172 86–102 (PMLR, 06--08 Jul 2022).
25. Ribeiro, A. L. P. *et al.* Tele-electrocardiography and bigdata: The CODE (Clinical Outcomes in Digital Electrocardiography) study. *J. Electrocardiol.* **57S**, S75–S78 (2019).
26. Wagner, P. *et al.* PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data* vol. 7 Preprint at https://doi.org/10.1038/s41597-020-0495-6 (2020).
27. Kashou, A. H. *et al.* A comprehensive artificial intelligence-enabled electrocardiogram interpretation program. *Cardiovasc Digit Health J* **1**, 62–70 (2020).
28. Siontis, K. C., Noseworthy, P. A., Attia, Z. I. & Friedman, P. A. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat. Rev. Cardiol.* **18**, 465–478 (2021).
29. ECG-GPT. *CarDS Lab* cards-lab.org/ecg-gpt (2024).
30. Wolf, T. *et al.* Transformers: State-of-the-Art Natural Language Processing. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, 2020).
31. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
32. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *arXiv [cs.LG]* (2017).
33. Ecg-plot. *PyPI* https://pypi.org/project/ecg-plot/ (2021).
34. Pyenchant. *PyPI* https://pypi.org/project/pyenchant/ (2021).
35. Kligfield, P. *et al.* Recommendations for the standardization and interpretation of the electrocardiogram: part I: The electrocardiogram and its technology: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Computerized Electrocardiology. *Circulation* **115**, 1306–1324 (2007).
36. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).
37. von Werra, L. *et al.* Evaluate & evaluation on the Hub: Better best practices for data and model measurements. *arXiv [cs.LG]* (2022).
38. Chen, X. *et al.* Microsoft COCO captions: Data collection and evaluation server. *arXiv [cs.CV]* (2015).
39. Sun, X. & Xu, W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Process. Lett.* **21**, 1389–1393 (2014).
40. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).
41. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979).

Figure 1. Overview of ECG-GPT's development and evaluation.
Abbreviations: ECG, electrocardiogram; BEiT, Bidirectional Encoder representation from Image Transformers; GPT-2, Generative Pretrained Transformer-2.





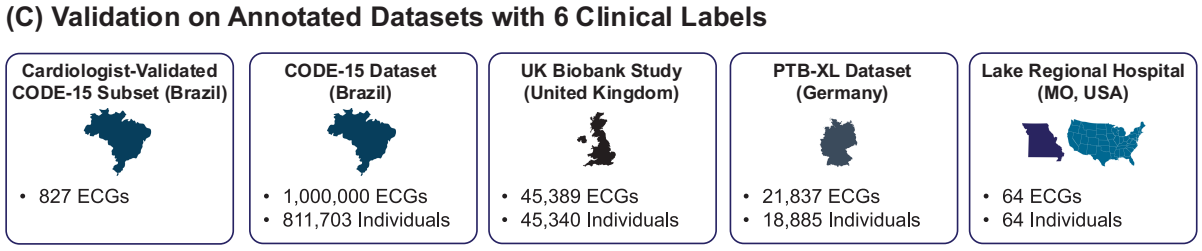


Figure 2. Vision Encoder-Decoder Model Architecture and Sample ECG-GPT Output Diagnosis Statement. Abbreviations: ECG, electrocardiogram; BEiT, Bidirectional Encoder representation from Image Transformers; GPT-2, Generative Pretrained Transformer-2.

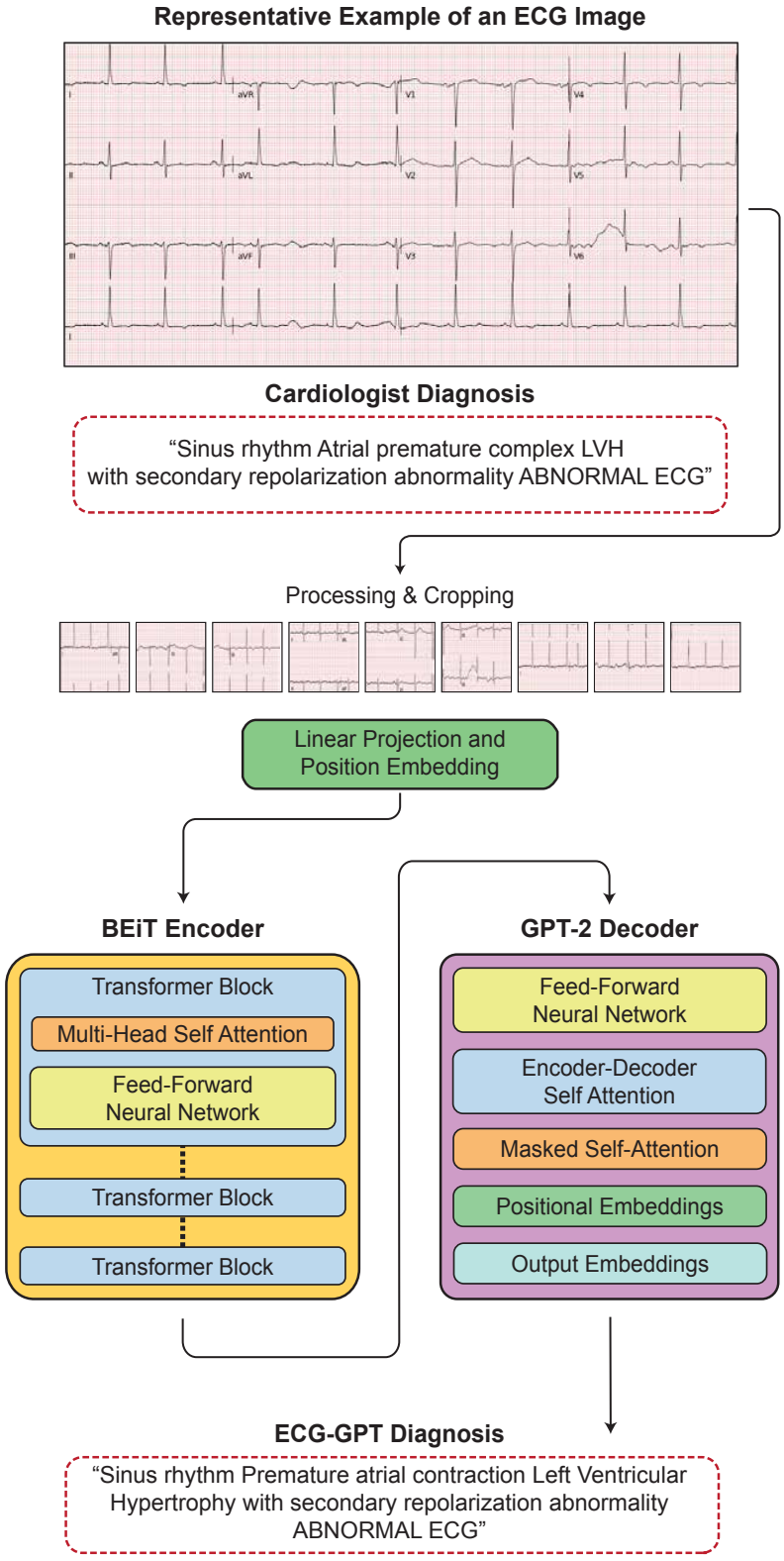


Table 1. Pairwise and baseline similarity between reference and model-generated diagnosis statements in the held-out test set. Abbreviations: AF, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; LBBB, left bundle branch block; RBBB, right bundle branch block; AVb, atrioventricular block; LAFB, left anterior fascicular block; LPFB, left posterior fascicular block; SVT, supraventricular tachycardia; PAC, premature atrial complexes; PVC, premature ventricular complexes; LAE, left atrial enlargement; LVH, left ventricular hypertrophy; MI, myocardial infarction.

Labels	Pairwise Similarity (IQR)	Baseline Similarity (IQR)	P-Value
Normal Sinus Rhythm	0.923 (0.855-0.987)	0.783 (0.735-0.833)	<0.001
AF	0.908 (0.834-1.000)	0.803 (0.738-0.858)	<0.001
Atrial Flutter	0.937 (0.877-1.000)	0.809 (0.759-0.858)	<0.001
ST	0.872 (0.792-0.949)	0.758 (0.673-0.861)	<0.001
SB	0.925 (0.857-0.988)	0.797 (0.730-0.859)	<0.001
Sinus Arrhythmia	0.903 (0.816-0.952)	0.822 (0.741-0.872)	<0.001
LBBB	0.847 (0.783-0.896)	0.743 (0.669-0.799)	<0.001
RBBB	0.917 (0.846-0.973)	0.756 (0.697-0.815)	<0.001
AVb	0.912 (0.839-1.000)	0.753 (0.694-0.811)	<0.001
LAFB	0.956 (0.870-1.000)	0.791 (0.734-0.847)	<0.001
LPFB	0.885 (0.807-0.959)	0.750 (0.695-0.811)	<0.001
SVT	0.915 (0.842-0.972)	0.754 (0.690-0.812)	<0.001
PAC	0.912 (0.843-0.958)	0.765 (0.696-0.823)	<0.001
PVC	0.957 (0.894-1.000)	0.837 (0.758-0.896)	<0.001
LAE	0.932 (0.867-0.980)	0.775 (0.703-0.841)	<0.001
LVH	0.881 (0.795-0.951)	0.749 (0.687-0.808)	<0.001
Low Voltage	0.930 (0.871-0.981)	0.793 (0.743-0.843)	<0.001
Left Axis Deviation	0.939 (0.881-1.000)	0.821 (0.762-0.873)	<0.001
Acute MI	0.867 (0.782-0.896)	0.757 (0.669-0.845)	<0.001

Table 2. Clinical assessment of model-generated diagnosis statements in the held-out test set. Abbreviations: PPV, Positive Predictive Value; NPV, Negative Predictive Value; AUROC, area under the receiver operator characteristic; AUPRC, area under precision-recall curve; AF, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; LBBB, left bundle branch block; RBBB, right bundle branch block; AVb, atrioventricular block; LAFB, left anterior fascicular block; LPFB, left posterior fascicular block; SVT, supraventricular tachycardia; PAC, premature atrial complexes; PVC, premature ventricular complexes; LAE, left atrial enlargement; LVH, left ventricular hypertrophy; MI, myocardial infarction.

Labels	Accuracy	PPV	NPV	Specificity	Sensitivity	AUROC	AUPRC	F1
Normal Sinus Rhythm	0.964	0.973	0.941	0.930	0.977	0.954 (0.953-0.955)	0.967 (0.966-0.968)	0.975
AF	0.986	0.912	0.992	0.992	0.911	0.952 (0.949-0.954)	0.839 (0.832-0.846)	0.912
Atrial Flutter	0.989	0.686	0.994	0.994	0.692	0.843 (0.834-0.852)	0.480 (0.459-0.499)	0.689
ST	0.991	0.929	0.995	0.995	0.936	0.965 (0.963-0.968)	0.873 (0.868-0.881)	0.932
SB	0.989	0.906	0.993	0.995	0.876	0.935 (0.932-0.939)	0.799 (0.789-0.808)	0.890
Sinus Arrhythmia	0.983	0.568	0.996	0.987	0.797	0.892 (0.885-0.899)	0.457 (0.443-0.472)	0.664
LBBB	0.992	0.877	0.995	0.997	0.829	0.913 (0.907-0.919)	0.732 (0.717-0.745)	0.852
RBBB	0.990	0.921	0.995	0.995	0.922	0.958 (0.956-0.961)	0.854 (0.847-0.861)	0.922
AVb	0.960	0.710	0.983	0.974	0.786	0.880 (0.876-0.884)	0.574 (0.564-0.582)	0.746
$2^{nd}/3^{rd}$ degree AVb	0.998	0.403	0.999	0.998	0.634	0.816 (0.789-0.847)	0.256 (0.204-0.305)	0.493
LAFB	0.977	0.733	0.990	0.987	0.780	0.883 (0.878-0.888)	0.581 (0.569-0.594)	0.756
LPFB	0.994	0.372	0.999	0.995	0.794	0.894 (0.878-0.911)	0.296 (0.268-0.323)	0.507
SVT	0.994	0.358	0.999	0.995	0.740	0.868 (0.849-0.887)	0.266 (0.244-0.297)	0.483
PAC	0.979	0.761	0.988	0.989	0.746	0.868 (0.862-0.873)	0.579 (0.564-0.591)	0.753
PVC	0.983	0.836	0.991	0.991	0.834	0.913 (0.908-0.917)	0.706 (0.696-0.718)	0.835
LAE	0.934	0.636	0.968	0.959	0.691	0.825 (0.821-0.829)	0.469 (0.461-0.477)	0.662
LVH	0.954	0.716	0.984	0.965	0.852	0.908 (0.905-0.911)	0.624 (0.617-0.631)	0.778
Low Voltage	0.950	0.622	0.985	0.962	0.808	0.885 (0.881-0.889)	0.516 (0.507-0.525)	0.703
Left Axis Deviation	0.948	0.651	0.978	0.966	0.747	0.856 (0.852-0.860)	0.506 (0.498-0.516)	0.696
Acute MI	0.991	0.271	0.998	0.993	0.629	0.811 (0.792-0.830)	0.172 (0.151-0.191)	0.379

Table 3. Pairwise and baseline similarity between reference and model-generated diagnosis statements in the Mount Sinai Health System. Abbreviations: AF, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; LBBB, left bundle branch block; RBBB, right bundle branch block; AVb, atrioventricular block; LAFB, left anterior fascicular block; LPFB, left posterior fascicular block; SVT, supraventricular tachycardia; PAC, premature atrial complexes; PVC, premature ventricular complexes; LAE, left atrial enlargement; LVH, left ventricular hypertrophy; MI, myocardial infarction.

Labels	Pairwise Similarity	Baseline Similarity	P-Value
Normal Sinus Rhythm	0.854 (0.785-0.911)	0.747 (0.674-0.810)	<0.001
AF	0.882 (0.807-0.945)	0.787 (0.731-0.841)	<0.001
Atrial Flutter	0.834 (0.755-0.900)	0.767 (0.707-0.822)	<0.001
ST	0.828 (0.754-0.894)	0.748 (0.684-0.806)	<0.001
SB	0.888 (0.819-0.960)	0.786 (0.728-0.837)	<0.001
Sinus Arrhythmia	0.919 (0.849-0.963)	0.843 (0.774-0.896)	<0.001
LBBB	0.927 (0.821-0.993)	0.818 (0.749-0.874)	<0.001
RBBB	0.897 (0.838-0.941)	0.803 (0.755-0.848)	<0.001
AVb	0.898 (0.829-0.947)	0.794 (0.728-0.851)	<0.001
LAFB	0.901 (0.846-0.939)	0.804 (0.752-0.850)	<0.001
LPFB	0.867 (0.794-0.917)	0.777 (0.720-0.828)	<0.001
SVT	0.804 (0.742-0.851)	0.763 (0.705-0.811)	<0.001
PAC	0.868 (0.793-0.925)	0.782 (0.720-0.835)	<0.001
PVC	0.886 (0.810-0.937)	0.806 (0.739-0.856)	<0.001
LAE	0.914 (0.846-0.958)	0.821 (0.760-0.871)	<0.001
LVH	0.883 (0.817-0.942)	0.805 (0.740-0.857)	<0.001
Low Voltage	0.832 (0.742-0.902)	0.731 (0.657-0.801)	<0.001
Left Axis Deviation	0.878 (0.811-0.930)	0.772 (0.711-0.824)	<0.001
Acute MI	0.787 (0.713-0.871)	0.752 (0.681-0.827)	<0.001

Table 4. Clinical assessment of model-generated diagnosis statements in the Mount Sinai Health System. Abbreviations: PPV, Positive Predictive Value; NPV, Negative Predictive Value; AUROC, area under the receiver operator characteristic; AUPRC, area under precision-recall curve; AF, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; LBBB, left bundle branch block; RBBB, right bundle branch block; AVb, atrioventricular block; LAFB, left anterior fascicular block; LPFB, left posterior fascicular block; SVT, supraventricular tachycardia; PAC, premature atrial complexes; PVC, premature ventricular complexes; LAE, left atrial enlargement; LVH, left ventricular hypertrophy; MI, myocardial infarction.

Labels	Accuracy	PPV	NPV	Specificity	Sensitivity	AUROC	AUPRC	F1
Normal Sinus Rhythm	0.926	0.951	0.877	0.902	0.938	0.920 (0.919-0.920)	0.933 (0.933-0.933)	0.944
AF	0.982	0.903	0.988	0.993	0.846	0.919 (0.918-0.921)	0.776 (0.773-0.778)	0.874
Atrial Flutter	0.987	0.599	0.994	0.993	0.651	0.822 (0.819-0.825)	0.396 (0.389-0.401)	0.624
ST	0.974	0.937	0.978	0.993	0.813	0.903 (0.902-0.904)	0.782 (0.780-0.784)	0.871
SB	0.961	0.771	0.991	0.965	0.934	0.950 (0.949-0.950)	0.728 (0.726-0.730)	0.845
Sinus Arrhythmia	0.960	0.534	0.989	0.968	0.774	0.871 (0.869-0.873)	0.423 (0.402-0.426)	0.632
LBBB	0.989	0.714	0.998	0.991	0.915	0.953 (0.951-0.954)	0.656 (0.651-0.660)	0.802
RBBB	0.983	0.818	0.997	0.985	0.955	0.970 (0.969-0.971)	0.784 (0.782-0.787)	0.881
AVb	0.945	0.592	0.987	0.954	0.838	0.896 (0.895-0.897)	0.508 (0.505-0.511)	0.694
$2^{nd}/3^{rd}$ degree AVb	0.994	0.266	0.999	0.995	0.640	0.818 (0.810-0.825)	0.171 (0.164-0.180)	0.376
LAFB	0.970	0.589	0.985	0.984	0.615	0.799 (0.797-0.802)	0.376 (0.372-0.380)	0.602
LPFB	0.987	0.140	0.998	0.989	0.429	0.709 (0.703-0.715)	0.062 (0.059-0.066)	0.211
SVT	0.991	0.232	0.999	0.992	0.692	0.842 (0.836-0.849)	0.161 (0.155-0.169)	0.347
PAC	0.964	0.751	0.975	0.987	0.614	0.800 (0.799-0.802)	0.484 (0.481-0.488)	0.675
PVC	0.976	0.761	0.987	0.987	0.758	0.873 (0.871-0.874)	0.589 (0.585-0.593)	0.759
LAE	0.693	0.233	0.993	0.665	0.953	0.809 (0.809-0.810)	0.227 (0.226-0.228)	0.375
LVH	0.914	0.637	0.965	0.936	0.765	0.850 (0.850-0.852)	0.517 (0.515-0.519)	0.695
Low Voltage	0.954	0.900	0.955	0.999	0.221	0.610 (0.608-0.611)	0.243 (0.241-0.246)	0.355
Left Axis Deviation	0.930	0.717	0.942	0.983	0.412	0.698 (0.696-0.699)	0.351 (0.348-0.353)	0.524
Acute MI	0.953	0.126	0.997	0.956	0.674	0.815 (0.811-0.819)	0.088 (0.086-0.090)	0.213

Table 5. Clinical assessment of model-generated diagnosis statements on external validation sets. Abbreviations: ECG, electrocardiogram; PPV, Positive Predictive Value; NPV, Negative Predictive Value; Spec, specificity; Sens, sensitivity; AUROC, area under the receiver operator characteristic; AUPRC, area under precision-recall curve. AFIB, atrial fibrillation; ST, sinus tachycardia; SB, sinus bradycardia; LBBB, left bundle branch block; RBBB, right bundle branch block; AVb, atrioventricular block.

	Labels	Accuracy	PPV	NPV	Specificity	Sensitivity	AUROC	AUPRC	F1
Cardiologist Validated CODE15	AF	0.995	0.800	0.999	0.996	0.923	0.960 (0.884-1)	0.740 (0.528-0.930)	0.857
	ST	0.990	0.872	0.996	0.994	0.919	0.956 (0.912-1)	0.805 (0.657-0.917)	0.895
	SB	0.886	0.139	0.999	0.885	0.938	0.911 (0.849-0.974)	0.131 (0.067-0.197)	0.242
	LBBB	0.996	0.966	0.997	0.999	0.933	0.966 (0.921-1)	0.904 (0.778-1)	0.949
	RBBB	0.992	0.846	0.999	0.992	0.971	0.982 (0.953-1)	0.822 (0.689-0.927)	0.904
	AVb	0.978	0.647	0.992	0.985	0.786	0.885 (0.808-0.963)	0.516 (0.329-0.713)	0.71
CODE15	AF	0.991	0.711	0.997	0.994	0.860	0.927 (0.924-0.929)	0.613 (0.607-0.620)	0.778
	ST	0.986	0.619	0.997	0.988	0.881	0.935 (0.933-0.937)	0.548 (0.542-0.553)	0.727
	SB	0.908	0.141	0.999	0.907	0.952	0.930 (0.928-0.931)	0.135 (0.133-0.137)	0.246
	LBBB	0.993	0.758	0.997	0.996	0.819	0.907 (0.904-0.910)	0.624 (0.615-0.632)	0.787
	RBBB	0.988	0.723	0.997	0.990	0.912	0.951 (0.949-0.953)	0.662 (0.658-0.666)	0.807
	AVb	0.971	0.326	0.997	0.974	0.815	0.894 (0.891-0.897)	0.268 (0.263-0.273)	0.465
UK Biobank	AF	0.998	0.961	0.998	0.999	0.896	0.948 (0.936-0.959)	0.862 (0.827-0.889)	0.927
	ST	0.999	0.743	1	0.999	0.848	0.924 (0.887-0.961)	0.630 (0.515-0.738)	0.792
	SB	0.953	0.680	0.995	0.953	0.952	0.952 (0.949-0.956)	0.651 (0.636-0.662)	0.793
	LBBB	0.998	0.851	1	0.998	0.990	0.994 (0.989-0.999)	0.842 (0.805-0.874)	0.915
	RBBB	0.993	0.755	1	0.994	0.984	0.989 (0.985-0.993)	0.743 (0.718-0.767)	0.854
	AVb	0.963	0.607	0.996	0.965	0.938	0.951 (0.946-0.956)	0.573 (0.557-0.589)	0.737
PTB-XL	AF	0.985	0.887	0.992	0.992	0.895	0.943 (0.936-0.951)	0.801 (0.783-0.820)	0.891
	ST	0.988	0.799	0.997	0.991	0.927	0.959 (0.950-0.968)	0.744 (0.716-0.772)	0.858
	SB	0.910	0.218	0.994	0.913	0.802	0.858 (0.842-0.873)	0.181 (0.167-0.196)	0.343
	LBBB	0.993	0.806	0.999	0.994	0.958	0.976 (0.967-0.985)	0.773 (0.740-0.811)	0.876
	RBBB	0.988	0.671	1	0.988	0.985	0.986 (0.981-0.992)	0.661 (0.631-0.692)	0.798
	AVb	0.941	0.354	0.990	0.948	0.759	0.853 (0.838-0.868)	0.277 (0.256-0.296)	0.482
LRH	AFIB	0.984	0.929	1	0.980	1	0.990 (0.971-1)	0.929 (0.750-1)	0.963
	ST	0.969	1	0.964	1	0.800	0.900 (0.769-1)	0.831 (0.618-1)	0.889
	SB	0.922	0.857	0.930	0.981	0.600	0.791 (0.630-0.952)	0.577 (0.300-0.849)	0.706
	LBBB	1	1	1	1	1	1 (1-1)	1 (1-1)	1
	RBBB	1	1	1	1	1	1 (1-1)	1 (1-1)	1
	AVb	0.906	0.692	0.961	0.925	0.818	0.871 (0.747-0.996)	0.598 (0.329-0.917)	0.75