npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

A foundational vision transformer improves diagnostic performance for electrocardiograms

Akhil Vaid[1,2,3,4]✉, Joy Jiang[1,2], Ashwin Sawant[5], Stamatios Lerakis[6,7], Edgar Argulian[6,7], Yuri Ahuja[8], Joshua Lampert[6,7], Alexander Charney[1,3,9,10], Hayit Greenspan[11], Jagat Narula[6,7], Benjamin Glicksberg[3,4] and Girish N Nadkarni[1,2,3,4,12]

The electrocardiogram (ECG) is a ubiquitous diagnostic modality. Convolutional neural networks (CNNs) applied towards ECG analysis require large sample sizes, and transfer learning approaches for biomedical problems may result in suboptimal performance when pre-training is done on natural images. We leveraged masked image modeling to create a vision-based transformer model, HeartBEiT, for electrocardiogram waveform analysis. We pre-trained this model on 8.5 million ECGs and then compared performance vs. standard CNN architectures for diagnosis of hypertrophic cardiomyopathy, low left ventricular ejection fraction and ST elevation myocardial infarction using differing training sample sizes and independent validation datasets. We find that HeartBEiT has significantly higher performance at lower sample sizes compared to other models. We also find that HeartBEiT improves explainability of diagnosis by highlighting biologically relevant regions of the EKG vs. standard CNNs. Domain specific pre-trained transformer models may exceed the classification performance of models trained on natural images especially in very low data regimes. The combination of the architecture and such pre-training allows for more accurate, granular explainability of model predictions.

npj Digital Medicine (2023)6:108 ; https://doi.org/10.1038/s41746-023-00840-9

INTRODUCTION

The electrocardiogram (ECG) is a body surface-level recording of electrical activity within the heart. Owing to its low cost, non-invasiveness, and wide applicability to cardiac disease, the ECG is a ubiquitous investigation and over 100 million ECGs are performed each year within the United States alone[1] in various healthcare settings. However, the ECG is limited in scope since physicians cannot consistently identify patterns representative of disease – especially for conditions that do not have established diagnostic criteria, or in cases when such patterns may be too subtle or chaotic for human interpretation.

Deep learning has been applied to ECG data for several diagnostic and prognostic use cases[2–6]. The vast majority of this work has been built upon Convolutional Neural Networks (CNNs)[7]. Like other neural networks, CNNs are high variance constructs[8], and require large amounts of data to prevent overfitting[9]. CNNs must also be purpose-built to accommodate the dimensionality of incoming data, and they have been used for interpreting ECGs both as 1D waveforms and 2D images[10].

In this context, interpreting ECGs as 2D images presents an advantage due to widely available pre-trained models which often serve as starting points for modeling tasks on smaller datasets[11]. This technique is described as *transfer learning* wherein a model that is trained on a larger, possibly unrelated dataset is fine-tuned on a smaller dataset that is relevant to a problem[12]. Transfer learning is especially useful in healthcare since datasets are limited in size due to limited patient cohorts, rarity of outcomes of interest, and costs associated with generating useful labels. As a result, vision models first trained in a supervised manner on natural images[13] often form the basis of models used in healthcare settings. Unfortunately, transfer learning with such natural images is not a universal solution, and it is known to produce suboptimal results when there exist substantial differences in the pre-training and fine-tuning datasets[14].

Transformer-based neural networks utilize the *attention mechanism*[15] to establish and define relationships between discrete units of input data known as tokens[16]. A significant benefit that transformers allow for is unsupervised learning from large corpora of unlabeled data to learn relationships between tokens, and then utilize this information for other downstream tasks[16]. Due to the ease with which unstructured text can be broken down into tokens, transformers have been tremendously successful at Natural Language Processing (NLP) tasks[17,18]. Recent work has extended the functionality of such models into vision-based tasks, leading to the advent of the vision transformer[16,19].

The first vision transformers were pre-trained on immense labeled datasets and then fine-tuned on smaller datasets to indicate better performance over CNNs at natural image classification[20]. More recently, the *Bidirectional Encoder representation from Image Transformers* (BEiT) approach has allowed large unlabeled datasets to be leveraged for pre-training transformer neural networks[21]. This approach consists of converting parts of an input image into discrete tokens or *patches*. Such tokens may be considered analogous to the words within a sentence and be used to pre-train a transformer in much the same way as a language model (Fig. 1). Since transformers consider global dependencies[22]

[1]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Mount Sinai Clinical Intelligence Center, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]The Hasso Plattner Institute for Digital Health at Mount Sinai, New York, NY, USA. [5]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Mount Sinai Heart, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]Department of Cardiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [8]Department of Medicine, NYU Langone Health, New York, NY, USA. [9]The Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [10]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [11]Department of Biomedical Engineering, Tel Aviv University, Tel Aviv 6997801, Israel. [12]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ✉email: akhil.vaid@mssm.edu

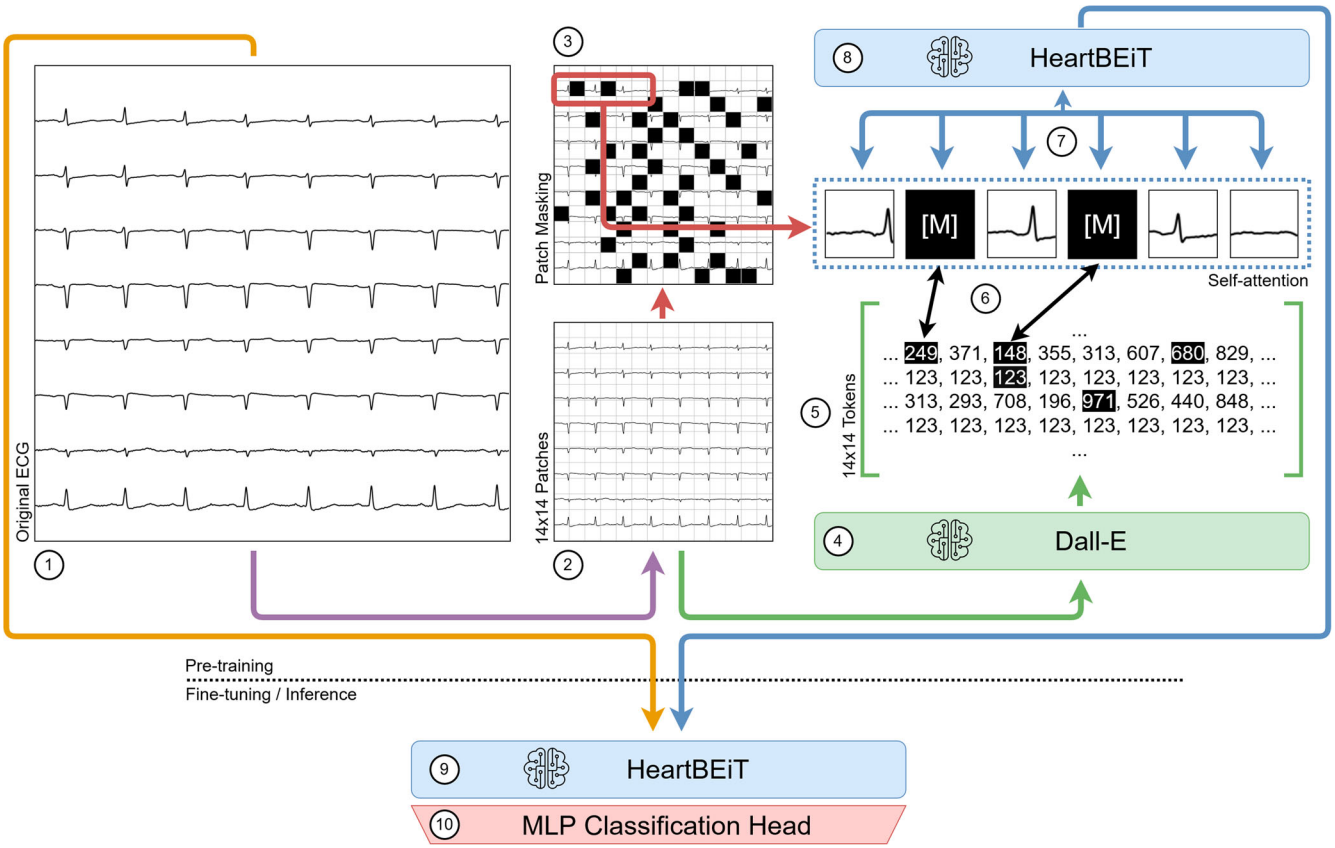

Fig. 1 Modeling workflow. Pre-training of the HeartBEiT model. (1) Each original ECG is partitioned into 14 × 14 patches (2) of 16 × 16 pixels. These patches are tokenized, and some of them are masked (3). The Dall-E model (4) acts as the tokenizer and converts the image into discrete tokens (5) which are then made part of the Masked Image Modeling process (6). This allows for pre-training the HeartBEiT model's attention modules (7), and the model may then be used for downstream fine-tuning and inference (8, 9) upon addition of a Multi-Layer Perceptron classification head (10).

between all features of provided inputs, such pre-training may be especially advantageous for ECGs. Certain pathological patterns such as the S1Q3T3 occur in different parts of a recording[23], and a model which considers only contiguous regions may miss them entirely.

We create a vision transformer model pre-trained on a large corpus of several million ECGs belonging to a diverse population. We utilize this model to create specialized models for use cases where little data may be available. We then compare performance and saliency maps to baseline models subject to similar constraints.

RESULTS

Performance at classification of LVEF

We included 511,491 total ECGs from MSHS in the training or fine-tuning set, 20,448 samples from MSHS in testing, and 1,480 from Morningside in external validation. Low LVEF prevalence was 18% in the training set (Table 1).

HeartBEiT outperformed other CNN models at low LVEF classification at all fractions of training data (Fig. 2; Supplementary Table 1). At 1% of training data (5114 samples), performance (AUROC: 0.86, 95% CI: 0.86–0.86) was 28.4% better than the ViT-B/16 model (AUROC: 0.67, 95% CI 0.67–0.67), 5.2% better than EfficientNet-B4 (AUROC: 0.82, 95% CI: 0.82–0.82), and 2.4% better than ResNet-152 (AUROC: 0.84, 95% CI: 0.84–0.84) in internal testing (Supplementary Fig. 2). These trends were maintained across external validation with HeartBEiT (AUROC: 0.87, 95% CI: 0.87–0.87) outperforming the CNNs by 4–18% (Supplementary Fig. 3).

Using AUPRC as a metric, at 1% of training data and against a prevalence of 18.5% in the internal testing cohort, the HeartBEiT model (AUPRC: 0.59, 95% CI: 0.59–0.59) outperformed ViT-B/16 (AUPRC: 0.31, 95% CI 0.31–0.31) by 90.3%, EfficientNet-B4 (AUPRC: 0.48, 95% CI: 0.48–0.48) by 22.9% and the ResNet-152 (AUPRC: 0.52, 95% CI: 0.52–0.52) by 13.5% (Supplementary Table 2, Supplementary Figs. 4–6). In the external validation cohort, HeartBEiT had the highest AUPRC of 0.73 (95% CI: 0.73–0.73).

With 100% of the training data (511,491 samples), performance across all models became more closely matched. In internal testing, there was no performance differential among HeartBEiT, EfficientNet-B4, and ResNet-152, and a differential of 1.1–4.5% was observed in external validation for AUROC. However, for AUPRC,

Table 1. Dataset Size with Outcome Prevalence.

	Fine-tuning	Testing	External Validation
Low LVEF			
Number of ECGs (*n*)	511,491	128,687	1480
Outcome Prevalence (%)	18.4	18.6	26.6
Hypertrophic Cardiomyopathy			
Number of ECGs (*n*)	78,831	20,448	13,859
Outcome Prevalence (%)	37.4	38.8	36.6
STEMI (PTB-XL database)			
Number of ECGs (*n*)	17,449	4,352	-
Outcome Prevalence (%)	5.7	5.4	-

Cells in bold typeface indicate an outcome.

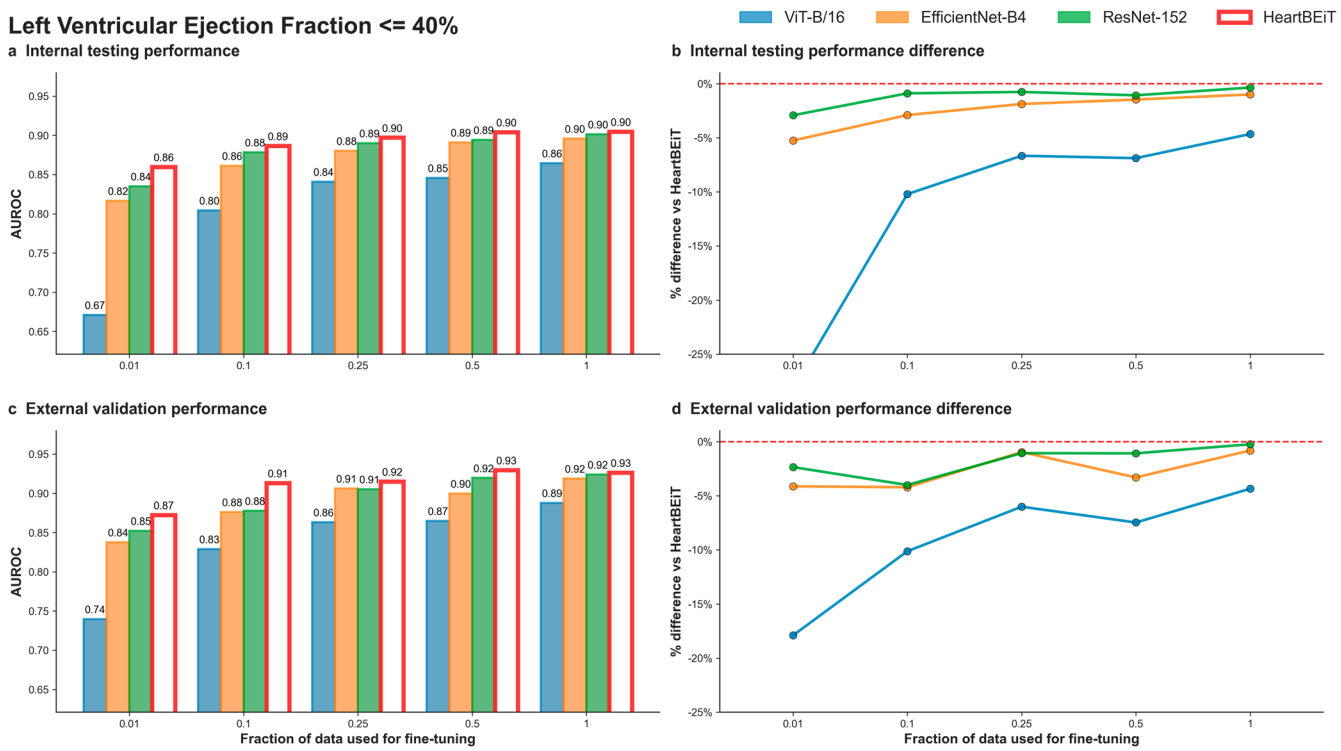

Fig. 2 Left ventricular ejection fraction < = 40% classification on ECGs. a Internal testing performance (4 Mount Sinai facilities). **b** Internal testing performance difference. **c** External validation performance (Morningside patients). **d** External validation performance difference. Red dashed line in (**b**) and (**d**) indicates HeartBEiT performance.

HeartBEiT still had improved performance of 0-17.7% in internal and external datasets.

GRAD-CAM analysis demonstrated areas around the QRS complexes of each lead were highlighted at 1% of training data by HeartBEiT (Supplementary Fig. 7a). When 100% of training data were implemented, foci became more pronounced around the QRS complexes of lead I (Supplementary Fig. 7b).

Performance at diagnosis of HCM

We fine-tuned the HeartBEiT transformer using 78,831 ECGs from four hospitals of the MSHS. Testing was conducted on 20,448 ECGs from these hospitals, and 3,859 ECGs from a holdout set of patients from Morningside were used for external validation (Table 1). The prevalence of HCM in the training set was 38%.

HeartBEiT outperformed the other models at diagnosis of HCM at all fractions of training data (Fig. 3; Supplementary Table 1). At 1% of training data, performance of the HeartBEiT model at AUROC of 0.77 (95% CI: 0.77–0.77) exceeded that of ViT-B/16 by 26.2% and of EfficientNet-B4 and ResNet-152 by 6.9% in internal testing (Supplementary Fig. 2). Similar results were seen for external validation with the HeartBEiT model which had an AUROC of 0.74 (95% CI: 0.74–0.74), outperforming ViT-B/16 (0.61, 95% CI 0.61–0.61) by 21.3%, EfficientNet-B4 (0.69, 95% CI: 0.68–0.70) by 7.2%, and ResNet-152 (0.68, 95% CI: 0.68–0.69) by 8.8% (Supplementary Fig. 3).

Differences in performance were much more profound for AUPRC at 1% of training data in use (Supplementary Table 2; Supplementary Fig. 8). Using 1% of training data, against an outcome prevalence of 38.8% in the internal testing cohort, the HeartBEiT model (AUPRC: 0.67, 95%, CI: 0.67–0.67) exceeded performance of ViT-B/16 (AUPRC: 0.49, 95% CI 0.49–0.49) by 36.7%, EfficientNet-B4 (AUPRC: 0.63, 95% CI: 0.63–0.63) by 6.3% and the ResNet-152 (AUPRC: 0.64, 95% CI: 0.64–0.64) by 4.7% (Supplementary Fig. 5). In external validation, HeartBEiT continued to exhibit the best performance with AUPRC of 0.64 (95% CI: 0.64–0.64) (Supplementary Fig. 6).

The HeartBEiT performance advantage reduced gradually as the amount of training data increased. Compared to 100% of the training data, the performance differential was up to 2.5% in internal testing and 3.9% external validation for AUROC and up to 4.2% and 7.1% for internal testing and external validation, respectively, for AUPRC.

GRAD-CAM analysis revealed that at 1% of the data, the QRS complexes of lead I, V2, and V5 and the ST segment of V6 were denoted as important regions for predicting HCM by HeartBEiT (Supplementary Fig. 9a). In contrast, at 100% of the training data, key areas identified by HeartBEiT became more focused to the beginning of V5 (Supplementary Fig. 9b).

Performance at detection of STEMI

The PTB-XL dataset contains 21,799 total ECGs from 18,869 patients: 17,449 ECGs were used for fine-tuning and 4352 to test the model. The prevalence of STEMI was around 5.7% in the training set and 5.4% in the testing set (Table 1).

The AUROC performance advantage of HeartBEiT was seen to be greater at smaller fractions of training data used for training (Fig. 4; Supplementary Table 1). In internal testing, the AUROC of HeartBEiT was 0.88 (95% CI: 0.88–0.89) with 4.8–10% performance improvement compared to the other models at 1% of training data (Supplementary Fig. 2). This advantage changed to approximately 20.3%, 1.1%, and 2.2% in comparison to ViT-B/16, EfficientNet-B4, and ResNet-152, respectively, when all available training data (17,449 samples) were used.

This performance advantage became much more profound for AUPRC, with HeartBEiT (AUPRC: 0.56, 95% CI 0.56–0.66) outperforming ViT-B/16 (0.27, 95% CI 0.26–37) by 107.4%, ResNet-152 (0.47, 95% CI 0.46–0.47) by 19.1% and the EfficientNet-B4 (0.40, 95% CI 0.40–0.41) by 40.0% at a 1% fraction of training data (Supplementary Table 2; Supplementary Fig. 5; Supplementary Fig. 10). However, at 100% of training data, performance of HeartBEiT

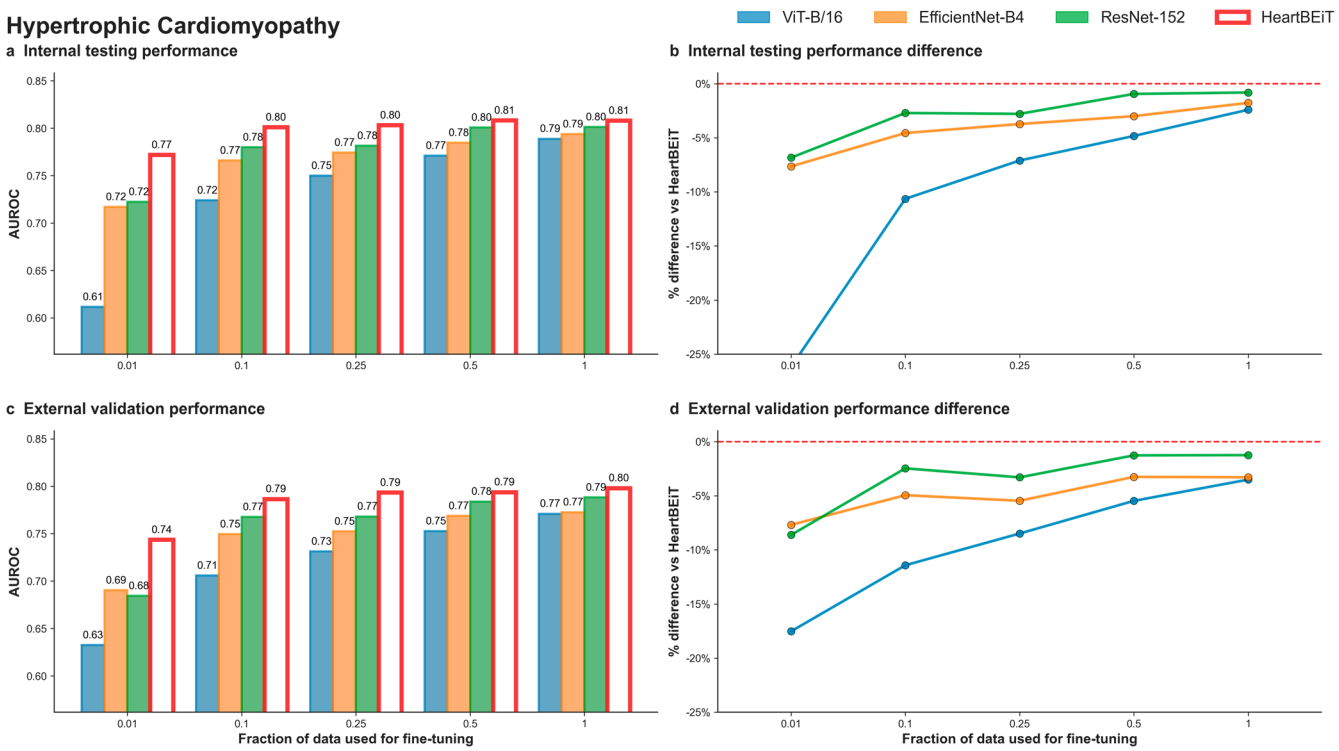

Fig. 3 Hypertrophic cardiomyopathy classification on ECGs. a Internal testing performance (4 Mount Sinai facilities). **b** Internal testing performance difference. **c** External validation performance (Morningside patients). **d** External validation performance difference. Red dashed line in (**b**) and (**d**) indicates HeartBEiT performance.

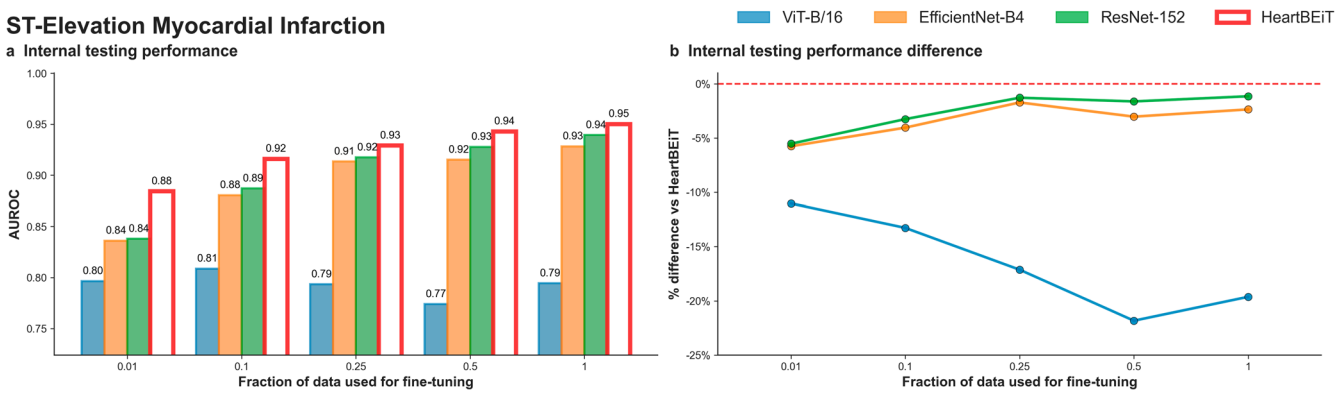

Fig. 4 STEMI detection on ECGs (PTB-XL database). a Internal testing performance. **b** Internal testing performance difference. Dashed red line in (**b**) indicates HeartBEiT performance.

(AUPRC: 0.67, 95% CI 0.66–0.67) became non-significantly lower than that of EfficientNet-B4 (AUPRC: 0.68, 95% CI: 0.67–0.68).

For STEMI detection, the ViT-B/16 vision transformer exhibited training instability when using more than 10% of training data while keeping other hyperparameters such as learning rate constant. This instability was seen only for this outcome, and reported performance corresponds to best metrics achieved prior to the training methods erroring out.

ST segments in each lead were underscored as areas of importance according to GRAD-CAM analysis of HeartBEiT at 1% of the training data (Fig. 5). At 100% of the training data, these areas denoted by HeartBEiT became localized around ST segments of leads V3 and V4 (Supplementary Fig. 11).

Wasserstein distance

The average pairwise Wasserstein distance for the ECG vs ECG set was 2.14. In comparison, this value was 45.48 for the ImageNet vs ImageNet set, and 128.44 for the ECG vs ImageNet set (Supplementary Fig. 12).

DISCUSSION

Using 8.5 million ECGs from 2.1 million patients collected over a period of four decades, we leveraged Masked Image Modeling to create a vision-based transformer (HeartBEiT) model for ECG data that can act as a universal starting point for downstream training on outcomes of interest. We fine-tuned this model against two outcomes using data derived from four hospitals within the Mount Sinai Health System, and externally validated derived models on data from another hospital. We also fine-tuned this model for STEMI detection using data from the publicly available PTB-XL database, followed by testing the derived model against a holdout set of patients. In each case, our model was compared against two CNNs and another vision transformer all subject to the same training conditions. Finally, we evaluated an additional aspect of

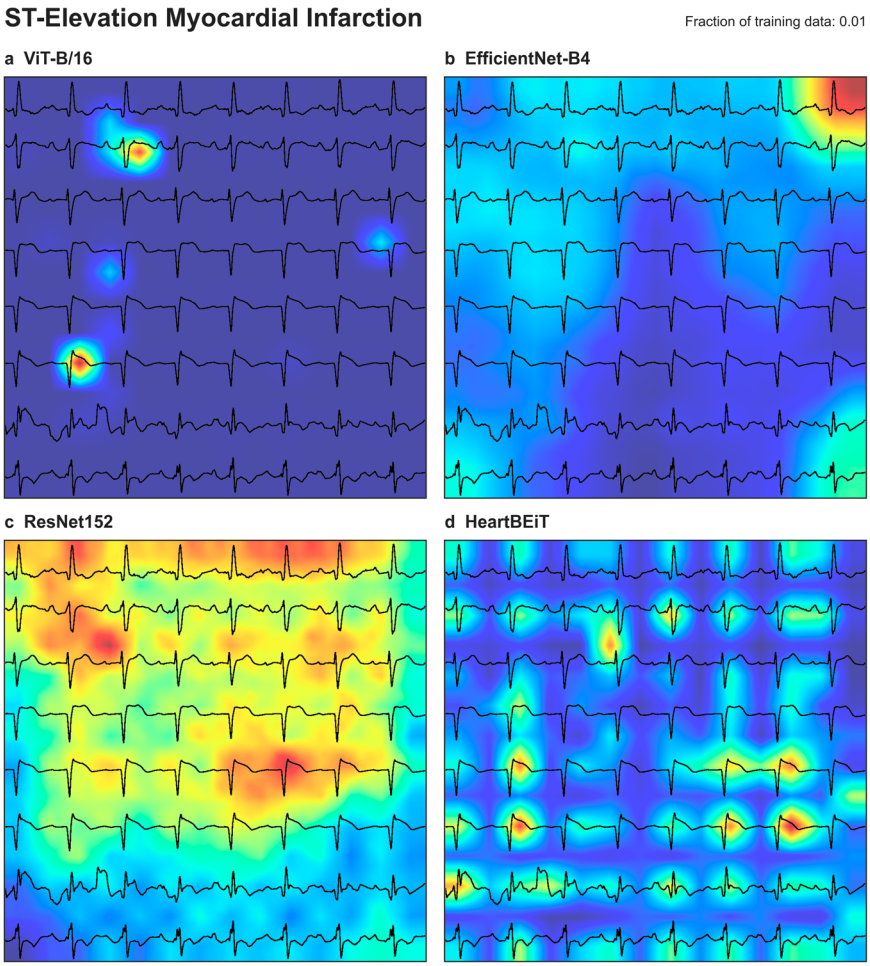

Fig. 5 Saliency mapping for STEMI detection at 1% training data. a ViT-B/16. **b** EfficientNet-B4. **c** ResNet-152. **d** HeartBEiT. HeartBEiT localizes to the ST segments. Other models are more diffuse in highlighting features of importance and may be less useful clinically.

clinical usefulness of these models by creating saliency maps for input samples.

Neural network performance can be heavily influenced by the amount of data available[24], and overfitting can easily result in small data regimes[25]. However, curated labeled data is a scarce resource. This is especially true in the healthcare setting wherein performing testing on patients, detecting pathologies of interest, and gathering data regarding clinical outcomes is laborious and expensive. In addition to the financial costs of acquiring and labelling data, time may be an additional factor that precludes acquisition of larger datasets. During emergent public health concerns, such as the recent COVID-19 pandemic, little data may be available for the development of useful models. In such circumstances, models that can work with a fraction of the data required for other approaches may assist in quicker, more appropriate diagnosis and triage.

Across all outcomes, datasets, and performance metrics, HeartBEiT achieved equivalent performance with an order of magnitude less (100% vs 10%) training data.". Further, in the very low data regime using only 1% of training data, HeartBEiT performance was equivalent to other models using 10 times as much data. This performance was maintained in external validation not only for the fine-tuned models, but also for the pre-trained model when used with an altogether new dataset from an independent dataset comprised of a geographically separated cohort of patients.

Of special importance is the elevated difference in performance in the AUPRC – a better indicator of performance in datasets with heavy class imbalance wherein considering AUROC in isolation may be less useful. Given relatively low event rates, medical datasets tend to have such class imbalances. For example, in detection of STEMI with an outcome prevalence of 5.6%, in the 1% training data regime, HeartBEiT exceeded the AUPRC of the CNNs by 19.1% and 40% respectively, while doubling the performance of the ImageNet vision transformer. These results also indicate that pre-training on natural images isn't always the most optimal solution for creating healthcare related models – a fact further evidenced by the extent of the disparity in the average Wasserstein distance between natural images and ECGs.

An emergent clinical advantage of using transformers with the explainability framework described in this work is the granularity of the saliency mapping. Even at similar levels of performance, the CNNs shown tend to coalesce areas of importance, thereby obfuscating the strongest determinants of a prediction. In comparison, saliency maps for transformers tend to focus on these determinants. Such granular explainability may help both clinician adoption of deep learning models, as well as aid in understanding pathologies for which there are no diagnostic guidelines on an ECG. These factors are demonstrated well for STEMI detection where the pathognomonic pattern is well established, and the ST segment is consistently highlighted even when using 1% of data for fine-tuning (Fig. 5). In the case of LVEF determination, there exist no clear diagnostic guidelines that can assist human physicians. In this case, saliency maps tend to focus on QRS complexes which indicate the net vector of depolarization of the majority of the cardiac ventricular musculature and point

towards the transformer's ability to focus on the mechanisms underlying the disease condition.

Our work must be considered in light of certain limitations. Transformers tend to be very compute intensive to pre-train. We were therefore limited in the size of the transformer model at 86 M parameters, as well as the dimensions of the input data we were able to utilize. However, we believe this work serves as evidence of the viability and advantages of our HeartBEiT model, and future work will deal with scaling up this model to enable better performance prior to live deployment.

In conclusion, pre-trained transformer models enable robust deep learning-based ECG classification even in severely data limited regimes. More specific, better quality, granular saliency maps can aid clinician acceptance of model predictions.

METHODS

Data sources

We utilized all available ECG data from five hospitals within the Mount Sinai Health System (MSHS) to pre-train our model. These hospitals (Mount Sinai Hospital, Morningside, West, Beth Israel, and Brooklyn) serve a large patient population that is reflective of the demographic diversity of New York City. ECG data were retrieved from the GE MUSE system for the years 1980–2021 totaling an approximate 8.5 million discrete ECG recordings for 2.1 million patients. ECG data were obtained as structured XML files containing both raw waveforms as well as metadata associated with patient identifiers, time, place, and indication.

For outcome specific fine-tuning of the model, we collected ground-truth labels for the value of the left ventricular ejection fraction (LVEF) from available echocardiogram reports. The modeling task was classification of patients for an LVEF ≤40%, which defines heart failure with reduced ejection fraction[26]. We also collected labels indicative of a diagnosis of Hypertrophic Cardiomyopathy – a genetic disorder wherein the chambers of the heart undergo a pathological increase in thickness resulting in loss of cardiac function and predisposition to fatal arrhythmias. These labels were generated using Natural Language Processing to parse unstructured echocardiogram reports for any mention of "HCM" / "Hypertrophic Cardiomyopathy" – with or without any intervening qualifiers regarding the obstructive nature of the pathology.

Finally, we utilized the publicly available PTB-XL dataset for additional external validation. This dataset contains 21,799 ECGs from 18,869 patients from October 1989 to June 1996. These data have been annotated by two cardiologists and contain ground-truth diagnostic labels such as whether an ECG is indicative of a normal recording or changes suggestive of acute ischemia. ECG recordings from this database were used to fine-tune models for detection of ST-Elevation Myocardial Infarction (STEMI). STEMIs are caused by acute loss of blood supply to heart tissue, and can result in a plethora of complications ranging from loss of contractile function to death.

Preprocessing

ECGs utilized within this study each contain waveform data recorded from one of twelve leads, with each lead representing a different perspective on the heart's electrical activity. Both datasets contain ECGs with either 5 or 10 s of waveform data per lead sampled at a rate of 500 Hz, for a total of 2500 or 5000 samples. The MSHS dataset does not contain data regarding leads III, aVF, aVL, or aVR. However, these leads are derived since they can be re-created from linear transformations of the vectors representing the other leads. In order to maintain uniformity across samples and datasets, all ECGs were truncated to 2500 samples.

We corrected for noise within ECG recordings through application of a *Butterworth bandpass filter* (0.5 Hz–40 Hz) followed by the application of a *median filter* on raw waveform data. Processed waveform data so derived was organized to maintain order of leads, and plotted to images with each image containing a total of eight leads (I, II, and V1 – V6). Images were saved in the.png (Portable Network Graphics) format at a resolution of 1000×1000 pixels to prevent compression artefacts. Additionally, output images were stored with three channels of color to retain compatibility with CNNs trained on ImageNet.

Tokens and tokenization

Tokens may be defined as discrete pre-defined sequences which are grouped and analyzed together on a semantic basis. In the context of language modeling, tokens may simply be the words comprising a body of text. The process of separating out data into such discrete sequences and assigning unique numeric identifiers to them is referred to as *Tokenization*[27].

Masked image modeling

A method commonly used to pre-train language models is called *Masked Language Modeling* (MLM)[28], wherein a set percentage of the number of tokens input to the model are masked or hidden, and models are pre-trained by having them predict these masked tokens. Collection and labeling of data may be an expensive process, and such costs are amplified for medical datasets. A significant advantage of MLM is that it allows for the usage of large quantities of unlabeled data to pre-train models.

The BEiT approach extends MLM into *Masked Image Modeling* (MIM) wherein 2D input images are separated into patches containing raw pixels which are then converted to tokenized representations of the input image (Fig. 1). This tokenization is accomplished using a separately trained image tokenizer that approximates each patch into a single numeric token. We used the same publicly available image tokenizer (Dall-E) for conversion of ECG images as the original BEiT implementation.

Model selection

We instantiated a 12-layer transformer model with a hidden layer size of 768, and 12 attention heads for a total of approximately 86 M parameters. This model, and its downstream derivatives are referred to as "**HeartBEiT**" within the text of this work.

We compared the downstream problem-specific performance of this model to an equivalently sized ImageNet based vision transformer (ViT-B/16: 86 M parameters), as well as CNN based approaches common to deep learning as applied to ECGs. These include the largest available pre-trained ResNet model (ResNet-152: 60 M parameters), and a computationally more inexpensive architecture (EfficientNet-B4: 19 M parameters) known to demonstrate better performance at image classification despite having fewer parameters. All baselines were pre-trained in a supervised manner on the ImageNet1K dataset containing 1.2 M labeled training images.

Pre-training

Input images were resized to 224×224 pixels, but otherwise subject to no other pre-processing. As opposed to natural images, ECG waveforms require maintenance of morphology and order. Random to loss of information that may only exist within certain segments of an ECG.

Input images were split into square patches of 16 pixels each, for a total of 196 patches per input image (Fig. 5). 40% of the input patches were masked for input into the neural network. We used the *AdamW* optimizer with a learning rate of 5e-4. The HeartBEiT model was pre-trained on a node consisting of 4 NVIDIA A100-40G GPUs. At approximately 6 h per epoch, pre-training the model for 300 epochs took around 2.5 months. Model parameters saved at

the 300th epoch were used for downstream fine-tuning in all cases (Supplementary Fig. 1).

Fine-tuning and statistical analysis

Pre-trained models were subjected to a fine-tuning task to demonstrate and compare performance at ECG based classification. We used data from 4 hospitals for detection of LVEF of < 40%, and diagnosis of HCM. In either case, the performance of the fine-tuned model was externally validated on data from Morningside hospital. Data from the PTB-XL database were used to fine-tune the pre-trained HeartBEiT model, as well as the other models for detection of STEMI.

Data were separated into a training dataset, an internal testing dataset, and where applicable, an external validation dataset. We modeled conditions of extreme data paucity by reducing training data to either 1%, 10%, 25%, 50%, or 100%, and then testing resulting models against common testing data. In all cases, *Group Shuffle Splitting* with a constant random seed was employed to ensure no patients were present in both training and testing data, and that the same patients were part of either dataset across runs.

We set the classification head of each model to a size of two neurons and utilized *CrossEntropy* loss. The *Adam* optimizer on a *OneCycle* learning rate schedule between 3e-4 and 1e-3 over 30 epochs was utilized for fine-tuning and reported performance metrics correspond to the best performance achieved across these epochs. Threshold independent Area Under the Receiver Operating Characteristic curve (AUROC) and Area Under the Precision Recall Curve (AUPRC) metrics were used to calculate and compare model performance. 95% confidence intervals for areas under the curve were generated through 500 iterations of the bootstrap.

Wasserstein distance

The Wasserstein distance[29] is a metric of the cost required to transform one distribution into another. Given two discrete images, the magnitude of the Wasserstein distance between them is directly proportional to how dissimilar they are. Higher Wasserstein distances between pre-training and fine-tuning data may lead to sub-optimal results with transfer learning.

We randomly sampled 1000 images each from both the ImageNet and ECG datasets. All samples from within each cohort were resized to 224×224 pixels and paired against all other samples from the same cohort, as well as the other cohort for a total of 3 such combinations: ECG vs ECG, ECG vs ImageNet, ImageNet vs ImageNet. Each such operation yielded a total of 10^6 pairs. The Wasserstein distance was calculated for each resulting pair of images and averaged across the combination of cohorts.

Explainability

Model explainability was generated using the *Gradient-weighted Class Activation Mapping* (GradCAM) library[30]. Generated attributions were plotted as an overlay upon the original input image to demonstrate which part of an input contributed most to a prediction.

Software

All analyses were performed using the *pandas, numpy, Python Image Library (PIL), SciPy, scikit-learn, torchvision, timm*, and *PyTorch* libraries. Plotting was performed using the *matplotlib* and *seaborn* libraries. All code was written for and within the 3.8.x version of the Python programming language.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Mount Sinai data utilized in this study are not publicly available due to patient privacy concerns. The PTB-XL dataset is publicly available for download at: https://doi.org/10.13026/kfzx-aw45 The HeartBEiT model may be released to other researchers on IRB-approved agreement with Mount Sinai Intellectual Partners.

CODE AVAILABILITY

Model creation code is not dataset specific, and is available at: https://github.com/akhilvaid/HeartBEiT.

Received: 13 January 2023; Accepted: 5 May 2023;
Published online: 06 June 2023

REFERENCES

1. Drazen, E., Mann, N., Borun, R., Laks, M. & Bersen, A. Survey of computer-assisted electrocardiography in the United States. *J. Electrocardiol.* **21**, S98–S104 (1988).
2. Vaid, A. et al. Automated Determination of Left Ventricular Function Using Electrocardiogram Data in Patients on Maintenance Hemodialysis. *Clin. J. Am. Soc. Nephrol.* **17**, 1017–1025 (2022).
3. Vaid, A. et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. *Cardiovasc. Imaging* **15**, 395–410 (2022).
4. Vaid, A. et al. Multi-center retrospective cohort study applying deep learning to electrocardiograms to identify left heart valvular dysfunction. *Commun. Med.* **3**, 24 (2023).
5. Mincholé, A., Camps, J., Lyon, A. & Rodríguez, B. Machine learning in the electrocardiogram. *J. Electrocardiol.* **57**, S61–S64 (2019).
6. Aziz, S., Ahmed, S. & Alouini, M.-S. ECG-based machine-learning algorithms for heartbeat classification. *Sci. Rep.* **11**, 18738 (2021).
7. Hong, S., Zhou, Y., Shang, J., Xiao, C. & Sun, J. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers Biol. Med.* **122**, 103801 (2020).
8. Geman, S., Bienenstock, E. & Doursat, R. Neural networks and the bias/variance dilemma. *Neural Comput.* **4**, 1–58 (1992).
9. Alzubaidi, L. et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **8**, 53 (2021).
10. Gu, J. et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **77**, 354–377 (2018).
11. Weimann, K. & Conrad, T. O. F. Transfer learning for ECG classification. *Sci. Rep.* **11**, 5251 (2021).
12. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J. Big Data* **3**, 9 (2016).
13. Deng, J. et al. In *2009 IEEE conference on computer vision and pattern recognition*. 248–255 (Ieee).
14. Gavrilov, A. D., Jordache, A., Vasdani, M. & Deng, J. Preventing model overfitting and underfitting in convolutional neural networks. *Int. J. Softw. Sci. Comput. Intell. (IJSSCI)* **10**, 19–28 (2018).
15. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* Vol. 30 (eds Guyon, I. et al.) (Curran Associates, Inc, 2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
16. Khan, S. et al. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)* **54**, 1–41 (2022).
17. Wolf, T. et al. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.
18. Kalyan, K. S., Rajasekharan, A. & Sangeetha, S. Ammus: A survey of transformer-based pretrained models in natural language processing. *Preprint at* https://arxiv.org/abs/2108.05542 (2021).
19. Liu, Z. et al. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.
20. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint at* https://arxiv.org/abs/2010.11929 (2020).
21. Bao, H., Dong, L. & Wei, F. Beit: Bert pre-training of image transformers. *Preprint at* https://arxiv.org/abs/2106.08254 (2021).
22. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. & Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **34**, 12116–12128 (2021).
23. Shahani, L. S1Q3T3 pattern leading to early diagnosis of pulmonary embolism. *BMJ Case Rep.* **2012** https://doi.org/10.1136/bcr-2012-006569 (2012).

24. Raudys, S. J. & Jain, A. K. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 252–264 (1991).
25. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
26. Bozkurt, B. et al. Universal definition and classification of heart failure: a report of the heart failure society of America, heart failure association of the European society of cardiology, Japanese heart failure society and writing committee of the universal definition of heart failure. *J. Card. Fail.* **27**, 387–413 (2021).
27. Webster, J. J. & Kit, C. In *COLING 1992 volume 4: The 14th international conference on computational linguistics*.
28. Ghazvininejad, M., Levy, O., Liu, Y. & Zettlemoyer, L. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 6112–6121. https://arxiv.org/abs/1904.09324 (Association for Computational Linguistics, Hong Kong, China, 2019).
29. Rubner, Y., Tomasi, C. & Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Computer Vis.* **40**, 99–121 (2000).
30. Selvaraju, R. R. et al. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

ACKNOWLEDGEMENTS

This study was funded by R01HL155915 and Clinical and translational award for infrastructure UL1TR004419. The authors would like to thank Wei Guo, Lili Gai, and Eugene Fluder of the High Performance Computing group at Mount Sinai for making the infrastructure underlying this study possible.

AUTHOR CONTRIBUTIONS

The study was designed by A.V.; The code was written by A.V.; Underlying data were collected, analyzed, and visualized by A.V.; the first draft of the manuscript was written by A.V. and J.J.; G.N.N. supervised the project. A.V. and G.N.N. had access to and verified the data. All authors provided feedback and approved the final draft for publication.

COMPETING INTERESTS

Dr. Nadkarni reports consultancy agreements with AstraZeneca, BioVie, GLG Consulting, Pensieve Health, Reata, Renalytix, Siemens Healthineers, and Variant Bio; research funding from Goldfinch Bio and Renalytix; honoraria from AstraZeneca, BioVie, Lexicon, Daiichi Sankyo, Meanrini Health and Reata; patents or royalties with Renalytix; owns equity and stock options in Pensieve Health and Renalytix as a scientific cofounder; owns equity in Verici Dx; has received financial compensation as a scientific board member and advisor to Renalytix; serves on the advisory board of Neurona Health; and serves in an advisory or leadership role for Pensieve Health and Renalytix. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-023-00840-9.

Correspondence and requests for materials should be addressed to Akhil Vaid.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.