

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☐ | ☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Tabular data was obtained from the EHR and echocardiography lab sources and processed using standard dataframe manipulation software in Python (pandas) and R (tidyverse). No specialized software was used. The video data processing pipeline is contained in the Github repo https://github.com/broadinstitute/ml4h .
Data analysis	Model training was performed using Python version 3.6.9 with modules tensorflow 2.5.0 and ml4h 0.0.7. Model training and inference code with synthetic test data is available at https://github.com/broadinstitute/ml4h/tree/master/model_zoo/DROID-RV . Statistical analyses were performed using R (version 4.2.1) with packages survival (3.4-0), survminer (0.4.9), cmprsk (2.2-11), yardstick (1.1.0), mcr (1.3.3), and pROC (1.18.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

MGB clinical and echocardiogram data contain protected health information and cannot be shared publicly. Limited, deidentified source data sufficient to recreate the presented figures is provided in the supplementary information. Additionally, synthetic test data is provided at https://github.com/broadinstitute/ml4h/tree/master/model_zoo/DROID-RV to demonstrate the model training and inference pipelines. Reasonable requests for collaboration will be considered on a case-by-case basis by direct correspondence with the authors. RVNet data can be obtained for non-commercial research use by request from <https://rvenet.github.io/dataset/>.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

We have used the term sex to describe the patients in our cohorts. We report relatively a relatively balanced distribution of sex in our samples. As our work focused on deep learning estimation of echocardiogram measurements, which can be made from echocardiogram videos without knowledge of the patient's sex, we did not conduct separate analyses by sex. We did include sex as an adjustment when testing associations with downstream cardiovascular disease due to known associations between sex and CVD.

Reporting on race, ethnicity, or other socially relevant groupings

Race and ethnicity were self-reported and derived from EHR data. We did not use race and or ethnicity as adjustment variables in our analyses, as our models were designed to predict RV measurements which can be made from echocardiogram videos without knowledge of the patient's race or ethnicity.

Population characteristics

We have included information about age, sex, cardiovascular comorbidities and height and weight for our patient population. No adjustments were made for deep learning predicted values including RVEDD, predicted probability of RV dysfunction, and RVEF. For outcomes associations analyses, we report analyses adjusted for age, sex, height, weight, and left ventricular ejection fraction, as these are relevant confounders that affect both RV measurements and future CVD.

Recruitment

Our study employed two large retrospective, EHR-derived cohorts. Significant effort was made to validate a high level of comorbidity ascertainment and minimize data missingness. Namely, patients were required to have at least two visits with either a cardiologist or PCP (depending on the cohort) in the MGB healthcare system spaced by one year, to ensure longitudinal followup. The retrospective cohort design uses the second office visit as the start of follow up date to ensure adequate ascertainment. Additionally, natural language processing was used to extract variables including vital signs, height, and weight that have high missingness in tabular EHR datasets. This work was the subject of a previous publication "Cohort design and natural language processing to reduce bias in electronic health records research" (Khurshid et al npj Digital Medicine 2022). We note that the selection of patients in primary care and cardiology cohorts changes the distribution of RV measurements and comorbidities, which may bias the training of our model.

Ethics oversight

Our study was approved by the Mass General Brigham (MGB) Institutional Review Board (Protocol #2017P001650).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Our sample size was determined by the size of our cohort and available echocardiogram videos and associated measurements. The development of our cohort has been previously described (Khurshid et al npj Digital Medicine 2022; Lau et al JACC 2023). We believe these sample sizes are sufficient for our deep learning model given the strong model performance metrics and generalizability that we report in the manuscript.

Data exclusions

By necessity, we excluded patients without available echocardiogram videos that have necessary associated measurements to train our

Data exclusions	model. For relevant outcome analyses, we excluded patients with missing covariate data.
Replication	We demonstrate generalizability of our model to two cohorts distinct from the training cohort, including one from an outside institution. We find that our model's ability to predict RV size and function is replicated in these two distinct cohorts.
Randomization	Randomization is not relevant to our study because we aimed to predict RV measurements from echocardiogram videos, so all patients with available measurements could be included. We do not require a comparator group to define our model. Patient splits for model training/validation/testing were chosen randomly.
Blinding	Blinding is not applicable to our study as there was no randomization of groups. This was a retrospective study developing a machine learning model, not a randomized trial of an intervention.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a