

PhysioNet team

Tom Pollard, Benjamin Moody, Li-Wei Lehman; Brian Gow, Chen Xie, Dana Moukheiber, Lama Moukheiber; Ken Paik, Leo Celi, Alistair Johnson, Roger Mark

Open Data
@ MIT

PhysioNet : Open Data@MIT

Tom Pollard (@tompollard)

Friday, 28 October 2022





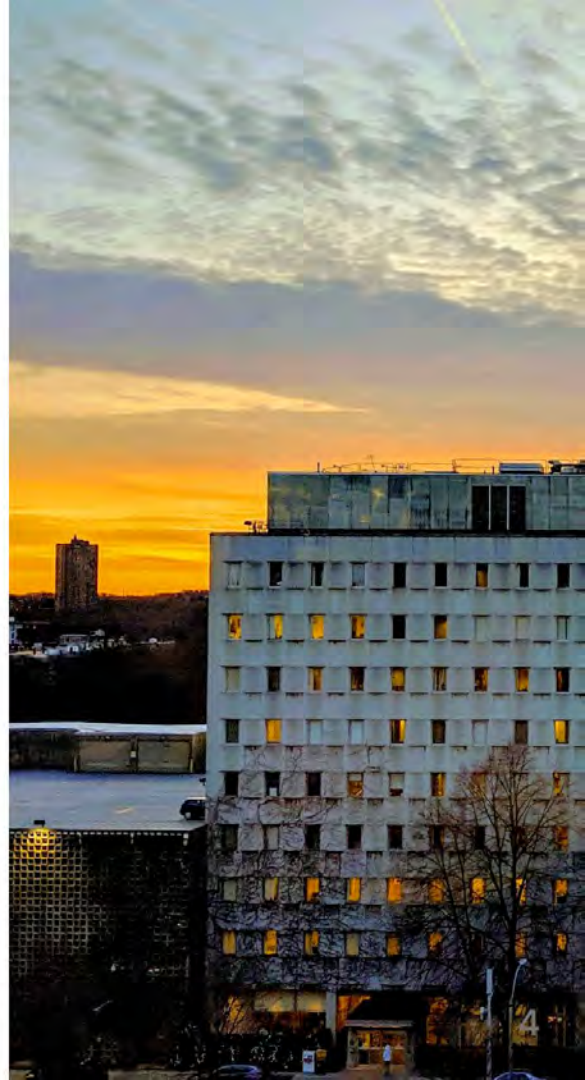
PhysioNet

- Data sharing platform built and maintained by MIT
- Began as outreach arm of a research project
- Rebuilt in 2019 following "FAIR principles"
- >55,000 registered, active users since 2019
- >30TB ecgs, x-rays, echocardiograms...

What are some examples of content shared on PhysioNet?

MIMIC-IV

- Highly-detailed critical care database for >40k patients in the US, comprising:
 - Vital signs, medications, labs..
 - Chest X-rays
 - ECGs, waveforms
 - Free text notes
 - Echocardiograms
- Extensively used across education, research, and industry



BRAX, a Brazilian labeled chest X-ray dataset

- 40,967 DICOM images
- De-identified to protect patient privacy
- Annotated by trained radiologists
- Adds geographic diversity in public chest x-ray data

BRAX, a Brazilian labeled chest X-ray dataset (v 1.0.0). PhysioNet. <https://doi.org/10.13026/ae9a-f727>
(image shows patient with pneumonia)



CheXclusion: Fairness gaps in deep chest X-ray classifiers

Laleh Seyyed-Kalantari¹, Guanxiong Liu, Matthew McDermott, Irene Y Chen, Marzyeh Ghassemi

Affiliations + expand

PMID: 33691020

Free article

Abstract

Machine learning systems have received much attention recently for their ability to achieve expert-level performance on clinical tasks, particularly in medical imaging. Here, we examine the extent to which state-of-the-art deep learning classifiers trained to yield diagnostic labels from X-ray images are biased with respect to protected attributes. We train convolution neural networks to predict 14 diagnostic labels in 3 prominent public chest X-ray datasets: MIMIC-CXR, Chest-Xray8, CheXpert, as well as a multi-site aggregation of all those datasets. We evaluate the TPR disparity - the difference in true positive rates (TPR) - among different protected attributes such as patient sex, age, race, and insurance type as a proxy for socioeconomic status. We demonstrate that TPR disparities exist in the state-of-the-art classifiers in all datasets, for all clinical tasks, and all subgroups. A multi-source dataset corresponds to the smallest disparities, suggesting one way to reduce bias. We find that TPR disparities are not significantly correlated with a subgroup's proportional disease burden. As clinical models move from papers to products, we encourage clinical decision makers to carefully audit for algorithmic disparities prior to deployment. Our

ACTIONS

“ Cite

☆ Favorites

SHARE



PAGE NAVIGATION

< Title & authors

Abstract

Similar articles

Cited by

Publication types

Multi-task Prediction of Organ Dysfunction in ICUs

Thursday, July 22, 2021

Posted by Subhrajit Roy, Research Scientist and Diana Mincu, Research Software Engineer, Google Research

The intensive care unit (ICU) of a hospital looks after the most medically vulnerable patients, many of whom require organ support, such as [mechanical ventilation](#) or [dialysis](#). While always critical, the demand on ICU services during the COVID-19 pandemic has further underscored the importance of data-driven decision-making in healthcare. Furthermore, the ability to accurately predict the clinical outcomes of ICU patients has the potential to guide therapy and may inform decisions about most effective care, including staffing and triage support.

Why do people choose to share on PhysioNet?

Expert review and curation

- > 20 years experience in safely sharing clinical data
- Assistance with data preparation
- State-of-the-art tools in de-identification

Fine-grained access control

- **Open data**
- **Restricted:**
 - Data Use Agreement.
- **Credentialed:**
 - Data Use Agreement
 - Training in human research.
 - Identity check.
- **Contributor-managed:**
 - Approval of the contributor.

Enhanced discovery

The screenshot shows a Google search interface with the search query 'eICU Collaborative Research Database'. The search results are filtered by 'Free' and 'Usage rights'. Three results are listed on the left, with the first one highlighted. The main content area displays details for the 'eICU Collaborative Research Database', including links to explore it on various platforms, citation information, and metadata.

Google

eICU Collaborative Research Database

Last updated Download format Usage rights Topic Free Saved data sets

9 data sets found

eICU Collaborative Research Database
physionet.org
commons.datacite.org
+1 more
Updated Apr 15, 2019

eICU Collaborative Research Database Demo
physionet.org
Updated May 6, 2021

eICU-CRD Dataset

eICU Collaborative Research Database

Explore at physionet.org Explore at commons.datacite.org
Explore at search.datacite.org

476 scholarly articles cite this dataset (View in Google Scholar)

Unique identifier
<https://doi.org/10.13026/C2WM1R>

Data set updated
Apr 15, 2019

Authors
Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Anthony Celi, Omar Badawi, Roger Mark

Licence
<https://github.com/MIT-LCP/icecap-and-dix/tree/master/drafts>

Integrated viewers

Select record to plot

wave_1

Input signals (8 maximum)

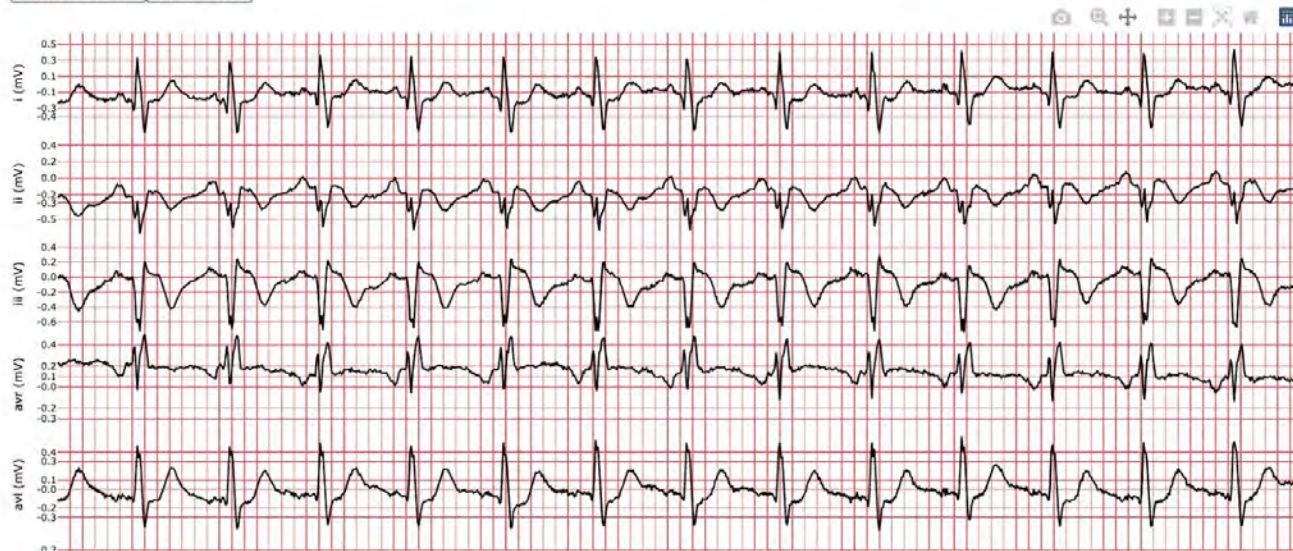
i ii iii avr avl avf v1 v2 v3 v4 v5 v6 vx vy vz

Go to time (HH:MM:SS) 00:00:00

Display annotations

On Off

Previous Record Next Record



Cloud integration

The screenshot shows a Google Colab notebook interface. At the top, the title is "mimic-cxr-train-aarhus.ipynb". Below the title is a menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". A status bar indicates "Cannot save changes". On the right, there are "Share", "Settings", and "Terminal" icons. The notebook content is divided into sections:

- Training a Convolutional Neural Network to Classify Chest X-rays**

This notebook shows how to train a state of the art Convolutional Neural Network (CNN) to classify chest X-rays images from the MIMIC CXR Dataset. Its approach is influenced by [CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison](#).

```
import datetime
import os
import tensorflow as tf
```
- Understanding the dataset**

First, we need to specify where the training and validation datasets are located. Labelled images are provided in [TFRecord](#) format.

```
! @title Input Datasets (run: "auto")
GCP_ANALYSIS_PROJECT = 'aarhus-critical-2019-team' #@param {type: string}
TRAIN_TFRECORDS = 'gs://mimic_cxr_derived/tfrecords/training' #@param {type: string}
VALID_TFRECORDS = 'gs://mimic_cxr_derived/tfrecords/validation' #@param {type: string}
# VIEW should be one of 'frontal', or 'lateral'
VIEW = 'frontal' #@param ["frontal", "lateral"] {type: string}
```

Input Datasets

```
GCP_ANALYSIS_PROJECT: aarhus-critical-2019-team
TRAIN_TFRECORDS: gs://mimic_cxr_derived/tfrecords/training
```

Recommended repository for leading journals

- Springer Nature, eLife, PLOS...

SPRINGER NATURE Search FR

Authors

Research data

Research data policies

Research Data Policies

Health sciences

Data policy types

Data availability statements

Data policy FAQs

Journal policies & services

Data repository guidance

Research Data Helpdesk

Some repositories in this section are suitable for datasets requiring restricted data access, which may be required for the preservation of study participant anonymity in clinical datasets. We suggest contacting repositories directly to determine those offering the data access controls which are best suited to your specific requirements. Authors should also consider whether they have access to a national, funder or project-specific repository that can facilitate data access controls, and which could therefore be suitable for hosting sensitive health science data.

Health science repository examples Information about data access options where available

[ClinicalTrials.gov](#)

[National Addiction & HIV Data Archive Program \(NAHDAP\)](#) restricted data access possible

[National Institute of Mental Health Data Archive \(NDA\)](#)




[PhysioNet](#)

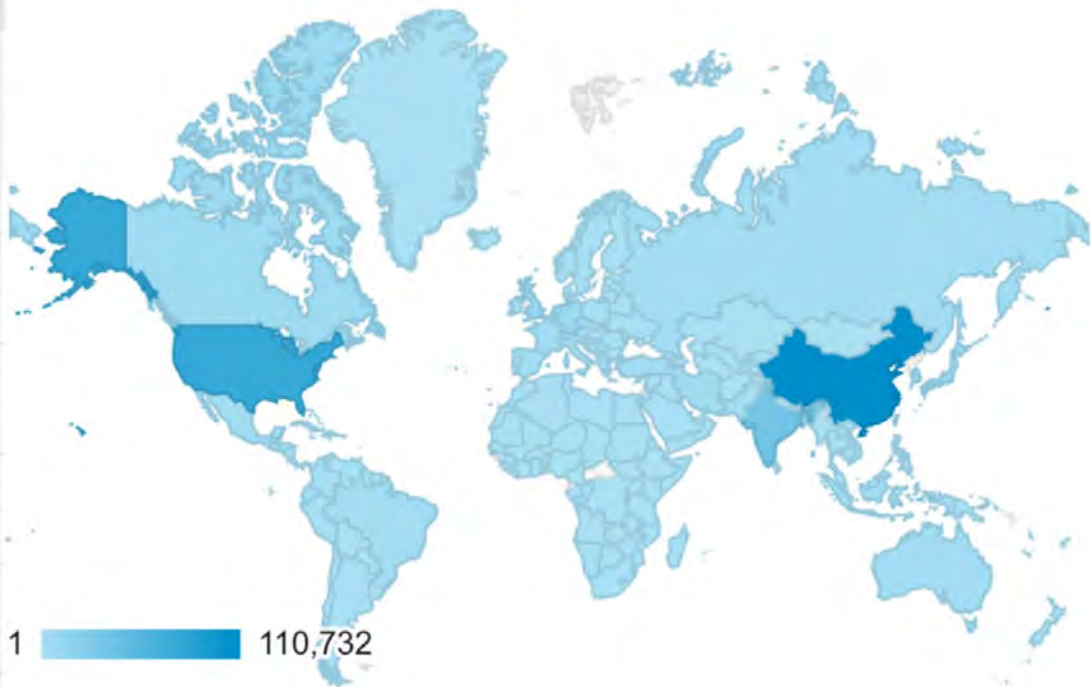
Who is using PhysioNet?

Diverse, active community

- Regular workshops, challenges, and datathons based around PhysioNet datasets



Country ?	Users ? ↓
	394,415 % of Total: 100.00% (394,415)
1.  China	110,732 (26.92%)
2.  United States	75,734 (18.41%)
3.  India	37,302 (9.07%)
4.  South Korea	10,671 (2.59%)
5.  United Kingdom	10,324 (2.51%)
6.  Germany	9,536 (2.32%)
7.  Canada	8,561 (2.08%)
8.  Japan	8,208 (2.00%)
9.  Taiwan	7,640 (1.86%)
10.  Iran	7,359 (1.79%)



PhysioNet in 2021

In the year 2021:

- 18,553 people created an account on PhysioNet
- 7,399 people were granted "credentialed" access
- ~7 TB of new data
- ~6,610 academic citations

What's next?

Open source code

Search or jump to... Pull requests Issues Marketplace Explore

MIT-LCP / physionet-build Public Unwatch 20 Star 28 Fork 14

<> Code Issues 204 Pull requests 24 Discussions Actions Projects 5 Wiki Security 4 Insights Settings

dev 38 branches 0 tags Go to file Add file + Code - About

tompollard Merge pull request #1421 from MIT-LCP/anonymous-url-pre... ✓ d765e1 3 days ago 3,481 commits

File/Folder	Description	Time Ago
.github/workflows	physionet-build-test.yml: only run flake8 on modified lines.	5 days ago
config	begin move towards a configurable platform by adding a config fol...	6 months ago
demo-files	Add an example file for the demo ArchivedProject.	6 months ago
deploy	Add single-Interpreter option to PhysioNet uWSGI configs	13 days ago
docker	Remove disallowed host logging	13 days ago
media	make cleaner zip function and other utilities. Move urls to search a...	3 years ago
physionet-django	Merge pull request #1431 from MIT-LCP/anonymous-url-prefix	3 days ago
.dockernignore	Add Dockerfile and docker-compose	4 months ago
.env.example	Set GOOGLE_APPLICATION_CREDENTIALS only if undefined	13 days ago
flake8	Add Makefile.	3 days ago
.gitattributes	.gitattributes: use diff-python for *.py	3 years ago
.gitignore	Add a settings folder that gets created by IntelliJ/JetBrains/PyChar...	9 months ago
CODE_OF_CONDUCT.md	Create CODE_OF_CONDUCT.md	3 years ago
Dockerfile	Add uWSGI config for k8s deployment with JSON logging	13 days ago
LICENSE	add license. closes #11	3 years ago
Makefile	Add Makefile.	3 days ago
README.md	PEP8 code style check only on modified lines of code relative to P...	10 days ago

The new PhysioNet platform.

[physionet.org](#)

[physionet](#)

Readme

BSD-3-Clause license

Code of conduct

Contributors 17

+ 6 contributors

Environments 1

dockerhub (Active)

Languages

Language	Percentage
JavaScript	66.9%
Python	8.5%
HTML	6.5%
C	6.3%
TeX	3.6%
Other	4.5%
Self	3.7%

Creating a network of repositories (pilots in Toronto and Uganda)

Infrastructure Overview



Thanks!

Benjamin Moody

Dr Alistair Johnson

Dr Leo Celi

Brian Gow

Lucas Bulgarelli

Dr Li-Wei Lehman

Dana Moukheiber

Lama Moukheiber

Dr Ken Paik

Lucas Mccullum

Chen Xie

Felipe Torres

Dr Jesse Raffa

Prof Roger Mark

