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](https://doi.org/10.13026/ae9a-f727) (image shows patient with pneumonia)
CheXclusion: Fairness gaps in deep chest X-ray classifiers

Laleh Seyyed-Kalantari, 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 work provides a base level of understanding to drive the next steps in reducing bias in healthcare applications.
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

Google Search for eICU Collaborative Research Database

9 data sets found

1. 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 Raffie, Leo Anthony Cell, Omar Badawi, Roger Mark

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

3. eICU-CRD Dataset
Integrated viewers

Select record to plot
wave_1

Input signals (8 maximum)

123 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

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.

```python
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.

```python
#title Input Datasets (run: "auto")
GCP_ANALYSIS_PROJECT = 'aarhus-critical-2019-team'
TRAIN_TFRECORDS = 'gs://mimic_cxr_derived/tfrecords/trai
VALID_TFRECORDS = 'gs://mimic_cxr_derived/tfrecords/val
# VIEW should be one of 'frontal', or 'lateral'
VIEW = 'frontal'  #param ['frontal', 'lateral'] (type:
```
Recommended repository for leading journals

- Springer Nature, eLife, PLOS...

Research data policies

Health sciences

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

- ClinicalTrials.gov
- National Anticancer & HIV Data Archive Program (NAHDAP)
- National Institute of Mental Health Data Archive (NIDA)

PhysoNet
Who is using PhysioNet?
Diverse, active community

- Regular workshops, challenges, and datathons based around PhysioNet datasets
<table>
<thead>
<tr>
<th>Country</th>
<th>Users</th>
<th>% of Total</th>
<th>% of Total Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>110,732</td>
<td>26.92%</td>
<td>(32%)</td>
</tr>
<tr>
<td>United States</td>
<td>75,734</td>
<td>18.41%</td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>37,302</td>
<td>9.07%</td>
<td></td>
</tr>
<tr>
<td>South Korea</td>
<td>10,671</td>
<td>2.59%</td>
<td></td>
</tr>
<tr>
<td>United Kingdom</td>
<td>10,324</td>
<td>2.51%</td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>9,536</td>
<td>2.32%</td>
<td></td>
</tr>
<tr>
<td>Canada</td>
<td>8,561</td>
<td>2.08%</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>8,208</td>
<td>2.00%</td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>7,640</td>
<td>1.86%</td>
<td></td>
</tr>
<tr>
<td>Iran</td>
<td>7,359</td>
<td>1.79%</td>
<td></td>
</tr>
</tbody>
</table>
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

https://github.com/MIT-LCP/physionet-build
Creating a network of repositories (pilots in Toronto and Uganda)
Thanks!

Benjamin Moody
Prof Roger Mark
Dr Alistair Johnson
Dr Leo Celi
Brian Gow
Dr Li-Wei Lehman
Dana Moukheiber
Lama Moukheiber
Dr Ken Paik
Shulammite Lim
Chen Xie