

Reproducible Research with Clinical Databases

Workshop

Wei-Hung Weng (courtesy of Alistair Johnson and Tom Pollard)

TMU
Sep 27, 2019



Massachusetts
Institute of
Technology

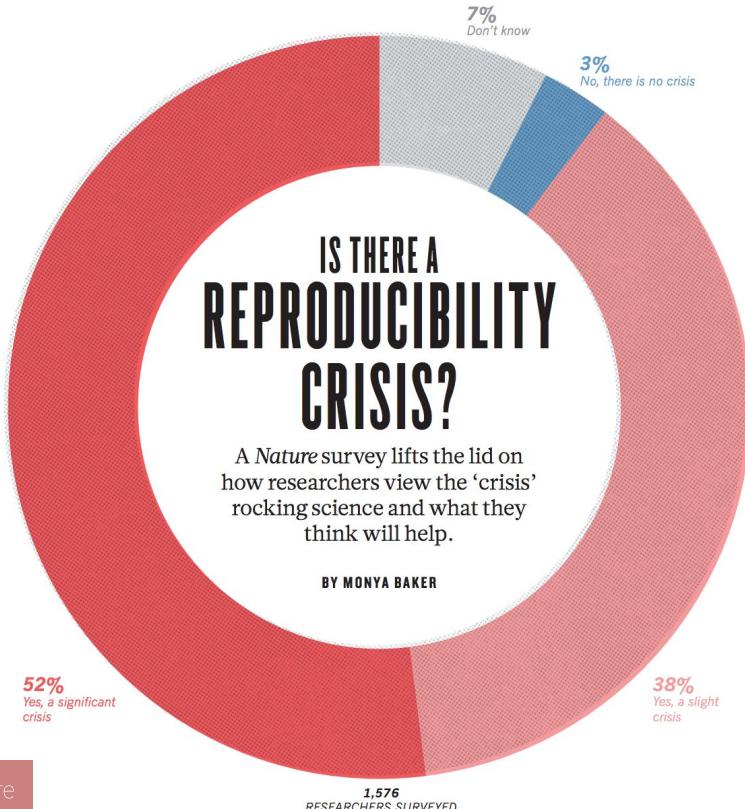


Outline

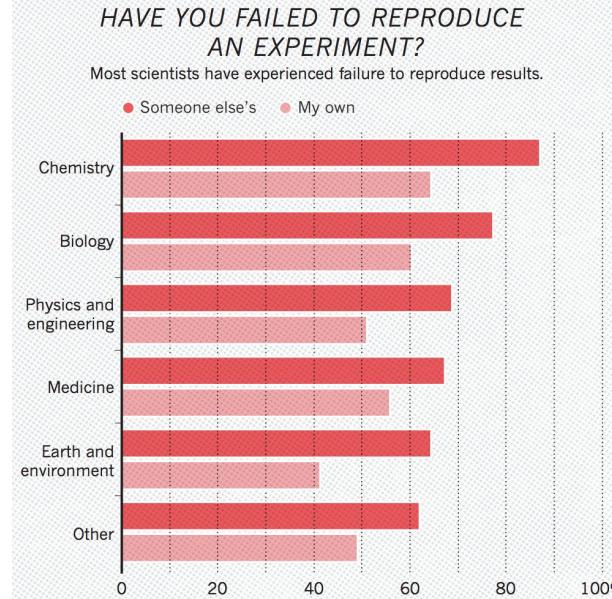
- Why reproducibility? How to make this happen?
- PhysioNet / eICU / MIMIC
- Collaboration

Reproducibility

Reproducibility

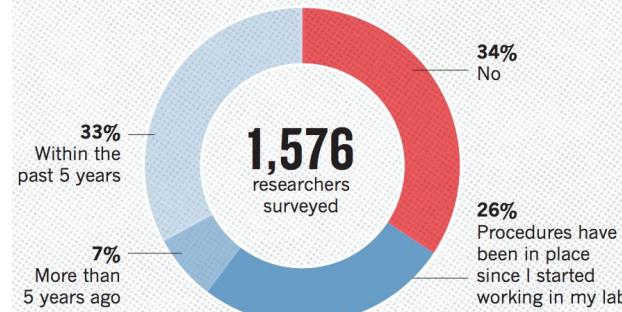


Baker, 2015 Nature



HAVE YOU ESTABLISHED PROCEDURES FOR REPRODUCIBILITY?

Among the most popular strategies was having different lab members redo experiments.



ICLR 2018 Reproducibility Challenge

Background:

One of the challenges in machine learning research is to ensure that published results are reliable and reproducible. In support of this, the goal of this challenge is to investigate reproducibility of empirical results submitted to the [2018 International Conference on Learning Representations](#).

We are choosing ICLR for this challenge because the timing is right for course-based participants (see below), and because papers submitted to the conference are automatically made available publicly on [Open Review](#).

The Challenge is inspired by discussions at the ICML 2017 [Workshop on Reproducibility in Machine Learning](#).

Task Description

You should select a paper from the 2018 ICLR submissions, and aim to replicate the experiments described in the paper. The goal is to assess if the experiments are reproducible, and to determine if the conclusions of the paper are supported by your findings. Your results can be either positive (i.e. confirm reproducibility), or negative (i.e. explain what you were unable to reproduce, and potentially explain why).

Essentially, think of your role as an inspector verifying the validity of the experimental results and conclusions of the paper. In some instances, your role will also extend to helping the authors improve the quality of their work and paper.

Reproducibility

- Data + code (+ documentation!)
- For...
 - Validation
 - Learning
 - Collaboration



Katerina Borodina @kathyra_ · Jul 7

my first coding job was with a small company, replacing their only developer who had recently quit. the code base was massive, in a language I had never used before. there was only one comment, at the end of an 1100 line function, and it said:

//that'll do pig. that'll do.

97 721 5.9K

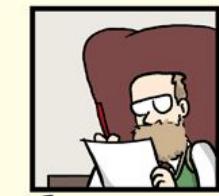
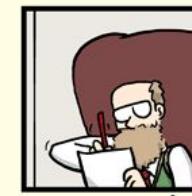
Tools

- Version control
 - Git
- Executable notebooks
 - Jupyter notebook
 - Google colab
- Code publishing platform
 - GitHub
 - Gitlab

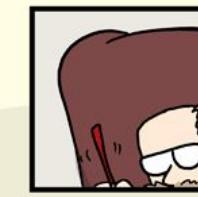
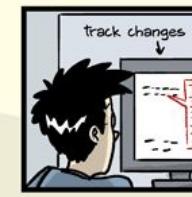
Version Control

- git add .
- git status
- git commit -m 'first commit'
- git push
- git reset '.DS_Store'
- ...

"FINAL".doc



JORGÉ CHAM © 2012



Jupyter Notebook

Execute (**Shift + Enter**) code cells and get your output underneath the cells

The screenshot shows a Jupyter Notebook interface with the following components:

- Title Bar:** Shows the logo, "Untitled", and a note about the last checkpoint being a minute ago (unsaved changes).
- Toolbar:** Includes buttons for File, Edit, View, Insert, Cell, Kernel, Help, and various cell-related icons like new cell, run, and cell toolbar.
- Code Cells:** Three code cells are visible:
 - In [1]:** `import sys
import os
import math
import numpy as np`
 - In [2]:** `x = [1, 4, 7, 10, 15]
np.mean(x)`
 - In [3]:** `print np.sqrt(sum(x))`
- Output:** The output for cell In [2] is **Out[2]: 7.4**. The output for cell In [3] is **6.082762530298219**.

CoLab (CoLaboratory)

<http://g.co/colab>

Write code just as you would on a Jupyter Notebook

Use one of google's virtual machines to carry out your tasks

Free

Can write shell commands preceded with a ‘!’

- !pip install gensim
- !ls

Code Publishing (scary but worth it)

ckbjimmy / p2c

Unwatch 1 Star 3 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Security Insights Settings

No description, website, or topics provided. Edit

Manage topics

6 commits 1 branch 0 releases 1 contributor MIT

Branch: master New pull request Create new file Upload files Find File Clone or download

Wei-Hung Weng exec Latest commit b544296 on May 12

data	exec	4 months ago
.gitignore	exec	4 months ago
LICENSE	Initial commit	8 months ago
README.md	exec	4 months ago
run.sh	exec	4 months ago

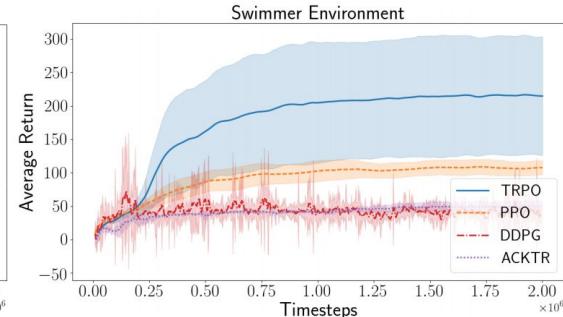
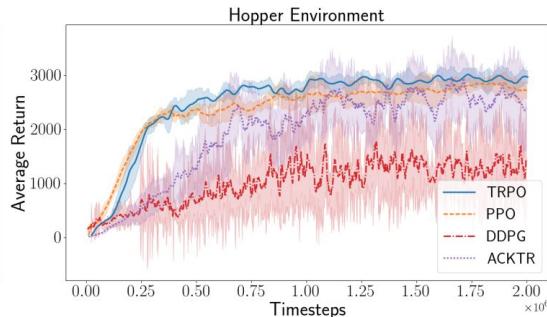
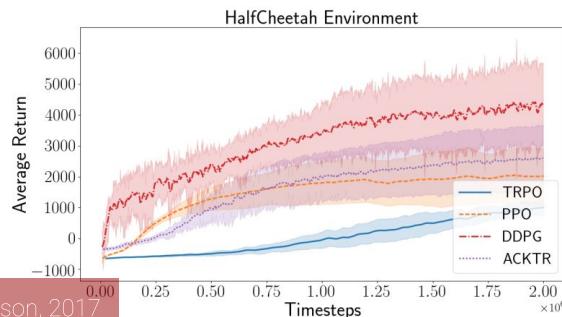
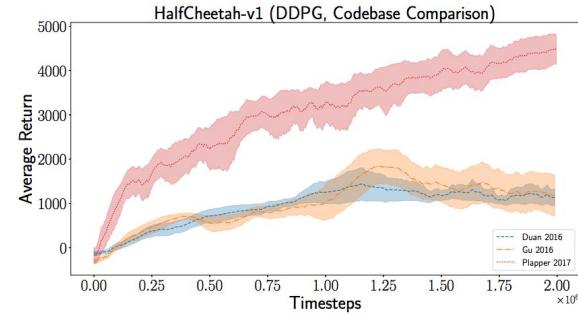
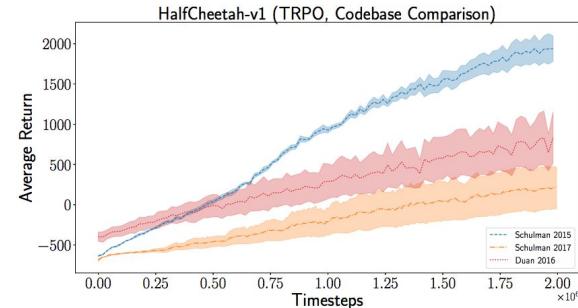
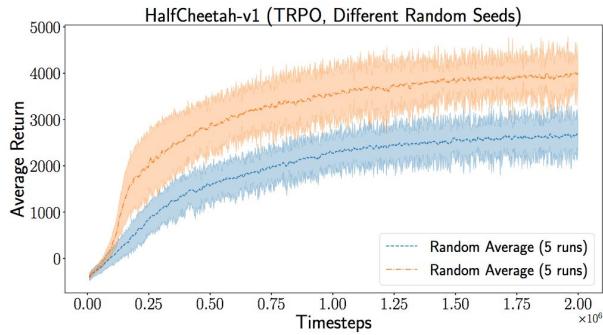
README.md

Unsupervised Clinical Language Translation

- Edited by Wei-Hung Weng (MIT CSAIL)
- Created: Jan 16, 2019
- Latest update: May 10, 2019
- Please contact the author with errors found.
- ckbjimmy {AT} mit {DOT} edu

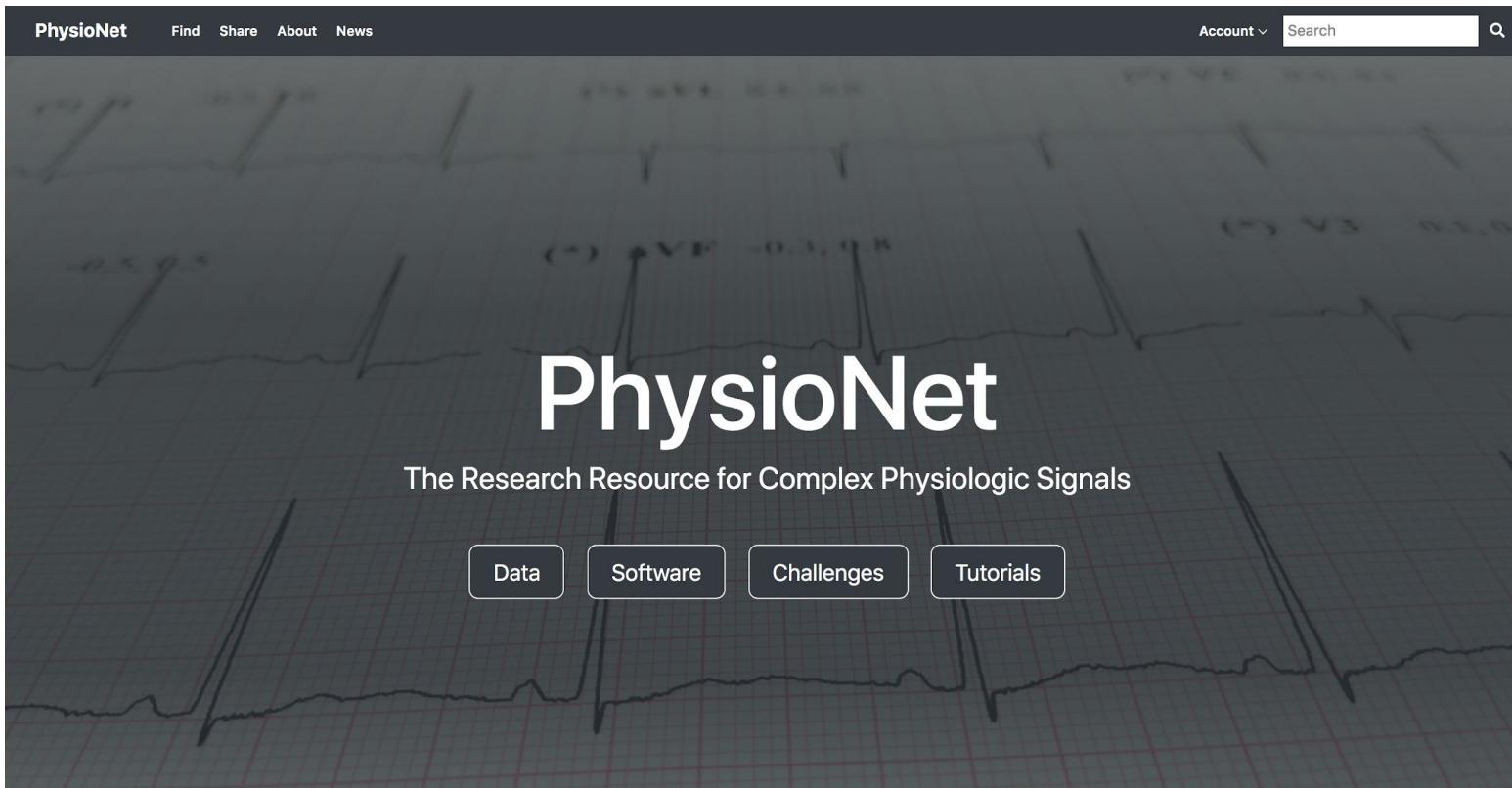
This repository contains the codes for the clinical language translation task using unsupervised SMT method presented in the paper [Unsupervised Clinical Language Translation \(KDD 2019\)](#).

Model Robustness



PhysioNet / eICU / MIMIC-III

physionet.org



The PhysioNet logo is overlaid on this image, featuring the word "PhysioNet" in a large, bold, white sans-serif font. Below it, the tagline "The Research Resource for Complex Physiologic Signals" is written in a smaller, white sans-serif font.

PhysioNet

Find Share About News

Account Search

Data Software Challenges Tutorials

 [Database](#) [Credentialed Access](#)

eICU Collaborative Research Database

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

Published: April 15, 2019. Version: 2.0

When using this resource, please cite:

Pollard, T., Johnson, A., Raffa, J., Celi, L. A., Badawi, O., Mark, R. (2019). eICU Collaborative Research Database. PhysioNet. doi:10.13026/C2WM1R

Additionally, please cite the original publication:

[The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG and Badawi O. Scientific Data \(2018\), DOI: <http://dx.doi.org/10.1038/sdata.2018.178>.](#)

Please include the standard citation for PhysioNet:

Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.

Contents [▼](#)

Share



Access

Access Policy:

Only PhysioNet credentialed users who sign the specified DUA can access the files.

License (for files):

[PhysioNet Credentialed Health Data License 1.5.0](#)

Abstract

The eICU Collaborative Research Database is a multi-center database comprising deidentified health data associated with over 200,000 admissions to ICUs across the United States between 2014–2015. The database includes vital sign measurements, care plan documentation, severity

[Database](#) [Credentialled Access](#)

MIMIC-III Clinical Database

Alistair Johnson , Tom Pollard , Roger Mark

Published: Sept. 4, 2016. Version: 1.4

When using this resource, please cite:

Johnson, A., Pollard, T., Mark, R. (2016). MIMIC-III Clinical Database. PhysioNet.
doi:10.13026/C2XW26

Additionally, please cite the original publication:

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.

Please include the standard citation for PhysioNet:

Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). *Circulation*. 101(23):e215-e220.

Abstract

MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

[Contents](#) [▼](#)**Share****Access****Access Policy:**

Only PhysioNet credentialled users who sign the specified DUA can access the files.

License (for files):

[PhysioNet Credentialled Health Data License 1.5.0](#)

MIMIC-CXR

PhysioNet Find Share About News Account ▾ Search 

 Database  Credentialed Access

MIMIC-CXR Database

Alistair Johnson , Tom Pollard , Roger Mark , Seth Berkowitz , Steven Horng 

Published: Sept. 19, 2019. Version: 2.0.0

When using this resource, please cite:

Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S. (2019). MIMIC-CXR Database. PhysioNet. doi:10.13026/C2JT1Q

Please include the standard citation for PhysioNet:

Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals (2003). Circulation. 101(23):e215-e220.

Contents ▾

Share



Access

Access Policy:

Only PhysioNet credentialed users who sign the specified DUA can access the files.

License (for files):

[PhysioNet Credentialed Health Data License 1.5.0](#)

Abstract

The MIMIC Chest X-ray (MIMIC-CXR) Database v2.0.0 is a large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. The dataset is de-identified to satisfy the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements. Protected health information (PHI) has been removed. The dataset is intended to support a wide body of research in medicine including image understanding, natural language processing, and decision support.

Warning

- Not in a clean, tidy spreadsheet
- Not generated for us to do research



r/MachineLearning



Find a community, post, or user

LOG IN

captkrob 2 points · 1 year ago

If you want an idea of a "real world" disease dataset, try looking at the MIMIC-III database (<https://mimic.physionet.org/>).

It's probably significantly more work to request the data and set it up in an easy format for ML than you had ever planned on, but this is the kind of thing actual data scientists and informaticians working on this kind of problem are used to dealing with. The dataset is tremendously rich, but also incredibly noisy. In other words, much like medical practice.

Share Save

Getting Access

- Sign data use agreement (DUA)
- Take online courses (citi)

Sign Data Use Agreement - eICU Collaborative Research Database v2.0

Sign the following data use agreement to access the files in [eICU Collaborative Research Database v2.0](#).

PhysioNet Credentialled Health Data Use Agreement 1.5.0

If I am granted access to the database:

1. I will not attempt to identify any individual or institution referenced in PhysioNet restricted data.
2. I will exercise all reasonable and prudent care to avoid disclosure of the identity of any individual or institution referenced in PhysioNet restricted data in any publication or other communication.
3. I will not share access to PhysioNet restricted data with anyone else.
4. I will exercise all reasonable and prudent care to maintain the physical and electronic security of PhysioNet restricted data.
5. If I find information within PhysioNet restricted data that I believe might permit identification of any individual or institution, I will report the location of this information promptly by email to PHI-report@physionet.org, citing the location of the specific information in question.
6. I have requested access to PhysioNet restricted data for the sole purpose of lawful use in scientific research, and I will use my privilege of access, if it is granted, for this purpose and no other.
7. I have completed a training program in human research subject protections and HIPAA regulations, and I am submitting proof of having done so.
8. I will indicate the general purpose for which I intend to use the database in my application.
9. If I openly disseminate my results, I will also contribute the code used to produce those results to a repository that is open to the research community.
10. This agreement may be terminated by either party at any time, but my obligations with respect to PhysioNet data shall continue after termination.

The screenshot shows the CITI Program website. At the top right, there are links for '+1 888.529.5929', 'English', 'Register', and 'Log In'. The main navigation menu includes 'Subscriptions', 'Courses', 'CE/CMEs', 'Tools', and 'Support'. A search icon is also present. The central content area features a large blue banner with the text 'Human Subjects Research (HSR)' in yellow. Below the banner, a sub-section titled 'HSR provides foundational training in human subjects research and includes the historical development of human subject protections, ethical issues, and current regulatory and guidance information.' is visible. On the left, a sidebar lists various training categories: 'View All', 'CE Certified Courses', 'Animal Care and Use (ACU)', 'Bioethics', 'Biomedical PI', 'Biosafety and Biosecurity (BSS)', 'Clinical Research Coordinator (CRC)', and 'Clinical Trial Billing Compliance (CTBC)'. At the bottom right, there are 'ORGANIZATIONS' and 'LEARNERS' sections with 'LEARN MORE' and 'BUY NOW' buttons, along with 'Questions?' and 'Contact Us' links.

Data Sharing

- More people seeing the data → more knowledge
- Education
- Accelerating research
- You can also share the data through PhysioNet!

MIMIC-III

- Single center (>60K ICU stay, >40K patients)
- ICU data in details (almost everything incl. notes and waveforms)
- Limited out-of-ICU data (medication, DOD, ...)



MIMIC-III, a freely accessible critical care database. Johnson AEW,
Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P,
Celi LA, and Mark RG. *Scientific Data* (2016). DOI: 10.1038/sdata.2016.35.
Available from: <http://www.nature.com/articles/sdata201635>

eICU

- Multicenter (>200 hospitals, >200K ICU stay, 2 years)
- ICU data (no free text notes)
- Heterogeneous data (every hospital has their own input format)



If you use the eICU Collaborative Research Database in your work, please cite the following publication:

The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG and Badawi O. *Scientific Data* (2018). DOI: <http://dx.doi.org/10.1038/sdata.2018.176>. Available from: <https://www.nature.com/articles/sdata2018176>

eICU [vs. MIMIC-III]

- Strengths
 - Latest data (2014-2015)
 - Larger dataset, multicenter
 - Severity score for all patients (APACHE)
 - Treatment plan in **carePlanGeneral** table
 - Diagnosis available during stay
- Limitations
 - Documentation, codebase and community support
 - Less literature
 - Varying between hospital
 - No clinical notes

Navigating eICU [eicu-crd.mit.edu]

The screenshot shows a navigation sidebar on the left with links like About, Getting started, and various tables. The main content area displays the apachePredVar page, which includes a purpose section and a detailed text about APACHE predictions. A sidebar on the right contains sections for Important considerations, Table columns, and Further reading.

eICU Collaborative Research Database eICU Collaborative Research Database

About >

Getting started >

Tables in eICU-CRD ▾

- admissiondrug
- admissiondx
- allergy
- apacheApsVar
- apachePatientResult
- apachePredVar**
- carePlanCareProvider
- carePlanEOL
- carePlanGeneral
- carePlanGoal
- carePlanInfectiousDisease

apachePredVar

Purpose: Provides variables underlying the APACHE predictions. Acute Physiology Age Chronic Health Evaluation (APACHE) consists of a groups of equations used for predicting outcomes in critically ill patients. APACHE II is based on the APS or acute physiology score (which uses 12 physiologic values), age, and chronic health status within one of 56 disease groups. APACHE II is no longer considered valid due to inadequate case mix index adjustments and over estimates mortality because it is based on models from the 1970s-1980s. APACHE III, introduced in 1991, improved the equation by changing the number and weights of the APS and revising the measurement of chronic health status. APACHE IVa further improved the equations and has been described as having the highest discrimination of any other adult risk adjustment model (SAPS 3, SOFA, MPM III).

apachePredVar

Important considerations

Table columns **Further reading**

Patient Tracking

- `patientunitstayid` → ICU stay
- `patienthealthsystemstayid` → hospital stay
- `uniquepid` → patient
- `hospitalid` → hospital
- ...offset

Tables

- **apacheapsvar**
 - First day aggregated data for APACHE
- **apachepredvar**
- **diagnosis** → offset, string, ICD code
- **infusiondrug** → fluid, insulin, vasopressor, sedative
- **intakeoutput** → urine output
- **patient** → demographics
- **lab, medication, pasthistory, treatment** → offset, string
- **vitalperiodic, vitalaperiodic**

- github.com/ckbjimmy/hst953_iomed

Codebase [github.com/MIT-LCP/eicu-code]

MIT-LCP / eicu-code

Unwatch 35 ⚡ Unstar 94 Fork 76

Code Issues 29 Pull requests 3 Projects 0 Wiki Security Insights

Code and website related to the eICU Collaborative Research Database <https://eicu-crd.mit.edu>

database eicu-crd mimic physionet ehr healthcare

187 commits 3 branches 1 release 8 contributors MIT

Branch: master New pull request Create new file Upload files Find File Clone or download

File	Commit Message	Time
build-db/postgres	add constraints for missing tables	last month
concepts	add chloride	4 months ago
demo	init info on demo	last year
notebooks	add pdvega for interactive plots	last year
website	update link for black website	2 months ago
.gitignore	add ds_store	2 years ago
LICENSE	clean up	3 years ago
README.md	add doi badge	last year
styleguide.md	remove external links	last year

README.md

eICU Collaborative Research Database Code Repository

DOI 10.5281/zenodo.1249016

Concepts

- In BigQuery
- Derived tables
 - `/eicu-code/concepts/pivoted/`

Branch: master ▾ [eicu-code](#) / [concepts](#) / [pivoted](#) /

 alistairewj	add chloride	Latest commit <code>c66c0fe</code> on May 12
..		
pivotedsqll	add pivot tables	last year
pivotedsqll	add pivot tables	last year
pivotedsqll	add chloride	5 months ago
pivotedsqll	remove superfluous column	last year
pivotedsqll	add pivoted views	last year
pivotedsqll	add pivoted views	last year
pivotedsqll	add pivot tables	last year
pivotedsqll	add pivot tables	last year
pivotedsqll	add pivoted views	last year
pivotedsqll	add pivoted views	last year

```
, max(case
    when drugname in
        (
            'EPI (mcg/min)'
            , 'Epinepherine (mcg/min)'
            , 'Epinephrine'
            , 'Epinephrine ()'
            , 'EPINEPHrine(Adrenalin)MAX 30 mg Sodium Chloride 0.9% 250 ml (mcg/min)'
            , 'EPINEPHrine(Adrenalin)STD 4 mg Sodium Chloride 0.9% 250 ml (mcg/min)'
            , 'EPINEPHrine(Adrenalin)STD 4 mg Sodium Chloride 0.9% 500 ml (mcg/min)'
            , 'EPINEPHrine(Adrenalin)STD 7 mg Sodium Chloride 0.9% 250 ml (mcg/min)'
            , 'Epinephrine (mcg/hr)'
            , 'Epinephrine (mcg/kg/min)'
            , 'Epinephrine (mcg/min)'
            , 'Epinephrine (mg/hr)'
            , 'Epinephrine (mg/kg/min)'
            , 'Epinephrine (ml/hr)'
        ) then 1 else 0 end)
    as epinephrine
```

Are You the First Person to Ask → GitHub Issues

MIT-LCP / [eicu-code](#)

[Unwatch](#) 35 [Unstar](#) 94 [Fork](#) 76

[Code](#) [Issues 29](#) [Pull requests 3](#) [Projects 0](#) [Wiki](#) [Security](#) [Insights](#)

[Filters](#) [Labels 7](#) [Milestones 0](#) [New issue](#)

Author	Labels	Projects	Milestones	Assignee	Sort
ragi24					
acanakoglu					
remifol					
RyuheiSo					
eruca					
shong95					

Dialysis
#83 opened 6 days ago by ragi24

Mechanical ventilation in eICU
#82 opened 14 days ago by acanakoglu 1 comment

APACHE score calculation for missing data
#81 opened 23 days ago by remifol 2 comments

How to differentiate cardiac deaths from non-cardiac deaths?
#80 opened 26 days ago by RyuheiSo

how to define patients first access to the icu stay?
#79 opened on Aug 9 by eruca

Stuck in the middle of procedure of eICU setting up
#78 opened on Jul 23 by shong95 10 comments

https://github.com/ckbjimmy/gcp/blob/master/tutorial_py.ipynb

```
from google.colab import auth
from google.cloud import bigquery
auth.authenticate_user()

project_id='...'
os.environ["GOOGLE_CLOUD_PROJECT"]=project_id

def run_query(query):
    return pd.io.gbq.read_gbq(query, project_id=project_id,
verbose=False, configuration={'query':{'useLegacySql': False}})
```

Collaborative Data Science in Medicine

- Find good problems, collect reliable data > Big data and algorithms
- Collaboration > Expert or AI only
 - Expert: high-level integration, interpretation and decision making (learn Bayesian logic, statistics, and data science and be aware of other sources of information!)
 - AI: pattern recognition and massive repetitive tasks
- Sharing
 - AI researchers open-source platforms and algorithms
 - You can open-source data and knowledge
 - We can do nothing without sharing!
- Experience
 - Learning from AI
 - Communicating with AI

Collaborative Data Science in Medicine

- Courses, Workshops, Symposium, Datathon
- Japan, China, Singapore, Philippines, Denmark, Bhutan

