Starter Kit for Bias Detection Tools for Clinical Decision-Making Challenge

Table of Contents

Starter Kit for Bias Detection Tools for Clinical Decision-Making Challenge

Thank you for registering for this challenge. To get you started on developing your bias detection tool, this document contains some introductory information about health information you may want to use to develop and test your tool. We will continually add to this document and welcome any additions that you would like to share with everyone.

Patients' clinical data is sensitive and data exchange must follow HIPAA Privacy rules. To overcome challenges with data access and limitations to the applications, in this starter kit we describe different types of data that you may choose to use. Note that the first example presents real data at the aggregate level and the second one presents synthetic data at the individual level. Therefore, the access to these records is free from cost, privacy, and security restrictions. Throughout this challenge we hope to add more sample data sources, so please refer to this document and watch the Slack channel for those updates. The first two options show clinical and demographic data that can be used in this NIH Challenge for bias detection within Machine Learning (ML) models. Also, the presented tools and methodologies are just examples of ways you may choose to use to start building or testing your bias detection tool. **You are not required to use any of these data sources, but they are available to you to help you get started.** Please refer to the Slack channel for additional information on other data sources, but it is your responsibility to find and decide what data sets you want to use.

1) Columbia Open Health Data (COHD)

About

Data provides access to counts and patient prevalence (i.e., prevalence from electronic health records) of conditions, procedures, drug exposures, and patient demographics, and the co-occurrence frequencies between them. Count and frequency data were derived from the Columbia University Irving Medical Center's OHDSI database including inpatient and outpatient data. Counts are the number of patients with the concept, e.g., diagnosed with a condition, exposed to a drug, or who had a procedure. Frequencies are the number of patients with the concept divided by the total number of patients in the dataset. Clinical concepts (e.g., conditions, procedures, drugs) are coded by their standard concept ID in the OMOP Common Data Model. To protect patient privacy, all concepts, and pairs of concepts where the count ≤ 10 were excluded, and counts were randomized by the Poisson distribution. Demographic attributes covered in the COHD dataset includes sex, race, and ethnicity. See the referenced webpage for more details https://cohd.io/about.html.

Data Structure

The data are presented at the concept level, in which every instance shows a concept identifier that refers to a clinical finding, procedure, or other OMOP ID (Observational Medical Outcomes Partnership for a standard concept ID). You can choose to look at the data from a single concept view (to see a view such as the average and standard deviation of the annual prevalence for the "Asthma" concept), or from a pair-concepts view to look at views such as conditions associated with "Aspirin" (data presents the count and the relative frequency of this occurrence). Also, data are broken down at two levels, 5-year period and life-time period, users may choose the data that best fit their needs.

Attributes, definitions, and access:

1) Single-concept level:
   - Concepts Deviation View (direct links: 5-year data, lifetime data)
     - concept id: Unique numeric code, use the “Concept name” column from the “Concepts” table for the full concept description.
     - mean: mean of annual prevalence for that concept
     - std (standard deviation): std of annual prevalence for that concept
   - Concepts Counts View (direct links: 5-year data, lifetime data)
     - concept id: Unique numeric code ¸use the “Concept name” column from the “Concepts” table for the full concept description.
     - count: count of each concept, includes patients from descendant concepts, i.e., SNOMED follows a hierarchical structure, in which concepts like “Lymphocytic Enteritis” are descendants of “Enteritis”, or “Inflammation of the small intestine”.
     - prevalence: The number of patients with this concept divided by the total number of patients in this data set (1.0 is 100%)
2) Pair-concepts level:
   - Concepts Deviation View (direct links: 5-year data, lifetime data)
     - Same as single-concept view, except that it breaks it down per the concept correlation with another concept listed in column “**concept id2**”
   - Concepts Counts View (direct links: 5-year data, lifetime data)
     - Same as single-concept view, except that it breaks it down per the concept correlation with another concept listed in column “**concept id2**”
3) Concepts: (direct link: concepts)
   - concept id: Unique numeric code identifying each concept
   - concept name: The descriptive name of each concept
   - Use the following search engine to search for a "concept id” by name https://athena.ohdsi.org/search-terms/start

Access

a) Direct download: through the above links

b) API request: https://smart-api.info/ui/9fbeaeabd19b334fa0f1932aa111bf35

Coverage and insights

The dataset covers the following categories of concepts, refer to this link for more insights: project written in Python, https://github.com/WengLab-InformaticsResearch/cohd_api/blob/master/notebooks/COHD_API_Example.ipynb

	dataset_id	domain_id	count
0	1	Condition	10159
1	1	Device	170
2	1	Drug	10264
3	1	Ethnicity	2
4	1	Gender	4
5	1	Measurement	188
6	1	Observation	870
7	1	Procedure	8270
8	1	Race	32
9	1	Relationship	5

2) SyntheticMass

About

SyntheticMass contains realistic but fictional residents of the state of Massachusetts. The synthetic population aims to statistically mirror the real population in terms of demographics, disease burden, vaccinations, medical visits, and social determinants. SyntheticMass establishes a risk-free environment for experimenting with large-scale HL7 FHIR® data. [https://synthea.mitre.org/about]

Data Structure

The SyntheticMass dataset is structured at the patient-level, in which it follows the FHIR standard that ensures interoperability during any electronic health records exchange. Data is encoded in FHIR, C-CDA, and CSV. In this dataset you will find information on patients, demographics, procedures, medications, observations, conditions, allergies, patient's health plans, healthcare providers, and others. Previous studies demonstrated the use of these datasets and reported many realistic insights [1-3].

Attributes, definitions, and access:

Data can be downloaded directly from the SyntheticMass website (https://synthea.mitre.org/downloads) or using their API (https://synthea.mitre.org/fhir-api).

3) Integrated Clinical and Environmental Exposure Data (ICEES)

About

ICEES is a novel, regulatory-compliant interface for presenting aggregated clinical data. It was developed by the Renaissance Computing Institute (RENCI) and researchers at the North Carolina Translational and Clinical Sciences Institute "for openly exposing clinical data on patients from UNC Health Care".

Data Structure

ICEES data is structured to support four types of data search. Each of these is briefly discussed below.

Cohort discovery: users define a cohort using any number of defined feature variables as input parameters, and the service returns a sample size.

Feature-rich cohort discovery: users select a predefined cohort as the input parameter, and the service returns a profile of that cohort in terms of the available feature variables.

Hypothesis-driven 2 x 2 feature associations: users select a predefined cohort and two feature variables, and the service returns a 2 x 2 feature table with a corresponding Chi Square statistic and P value.

Exploratory 1 X N feature associations: users select a predefined cohort and a feature variable of interest, and the service returns a 1 x N feature table with corrected Chi Square statistics and associated P values.

Attributes, definitions, and access:

Further information about the dataset can be found here, and here.
Data can be accessed via an API: https://github.com/ExposuresProvider/icees-api

ML Algorithms

The following are some examples of widely used ML algorithms (the most popular algorithms in the medical domain [4]) that can be used towards this competition. Best practice to minimize social bias when using these algorithms will be of a great impact:

- Artificial Neural Networks (Conventional or Deep Learning), Logistic Regression, Random Forest, decision tree, and k-nearest neighbor:

The following ML algorithms have also shown to be top performers in many applications:

- Bayesian models (such as Naive Bayes) and Gradient boosting (such as XGBoost)

ML examples:

- Refer to the following ML applications as examples to use to identify bias:
  - https://www.kaggle.com/code/drscarlat/predict-mortality
  - https://github.com/synthetichealth/syntheticmass/blob/master/modeling/US%20Data.R
  - https://github.com/Alvearie/patient_pathway_extractor/
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8861740/

Using the aggregated data in ML:

The COHD dataset was presented at the aggregate level instead of the individual-patient level. In order to use this data set in any ML applications, you may build your individualized data instances from these aggregates (utilizing attributes such as counts and prevalence), or by learning from aggregated data through incorporating methodologies designed to work with aggregates, such as Multilevel models or by compressing your data in some format that enables the application of ML algorithms in common Python libraries (scikit-learn, Keras, and others).

References

1) Chen, J., Chun, D., Patel, M. *et al.* The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak* 19, 44 (2019). https://doi.org/10.1186/s12911-019-0793-0
2) Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing
3) Jason Walonoski, Sybil Klaus, Eldesia Granger, Dylan Hall, Andrew Gregorowicz, George Neyarapally, Abigail Watson, Jeff Eastman, Synthea™ Novel coronavirus (COVID-19) model and synthetic data set,
4) Kim, Jong Taek. "Application of machine and deep learning algorithms in intelligent clinical decision support systems in healthcare." Journal of Health & Medical Informatics 9.05 (2018).