**Methods**

Study data was sourced from the several large-scale clinical data sets. These datasets contain either identifiable or various vintages of de-identified encounter level longitudinal data. The study datasets were split into cases who showed evidence of clinical status or lack of clinical status. The clinical status used here was HIV infection. HIV infection was assigned differently depending on the source dataset. Cases that did not present with assignment were considered HIV- throughout the studies. Status denominated cases were further split on survival status, where cases with an observed date of death were assigned ‘Dead’ in all study years. Cases without a date of death were assigned to ‘Alive’. Birth cohort (year of birth) was assigned using the latest populated year of birth declared on any claim type or program in the case of discrepancy. Gender (Male or Female) was assigned using Female preference if multiple genders were declared across claim types and programs. Cases were aggregated (counted) within Birth cohort, Survival, Gender and clinical Status (BSGS) for each distinct clinical code observed in a given study year. Cases within a BSGS group could have a maximum additive count of one code (annualized utilization) and a minimum of zero for a given clinical code in a given study year.

To aid machine learning of the differences that support clinical distinctiveness within status, and not simply (over fitting) statistical significance, for sone analyses (CMS datasets) clinical codes were assigned to a given compartment. One compartment per clinical insult (code string) was sought. Candidate SNOMED-CT terms had descendent terms joined to them using OHDSI ATHENA. If a descendent term was mapped to a given ICD code, then the ancestor SNOMED-CT code became a candidate compartment. In the case of multiple candidate compartments ties were broken using a ranked preference algorithm. The algorithm prioritized cancer followed by thoracic compartment candidacy, abdominal candidacy and ended with external causes of injury and sequela candidacy as the lowest prioritized. The aggregated, compartmentalized dataset had an attack rate computed out of the observed cases within BSGS and a given condition out of the total BSGS population over the study period. Disenrollment was not controlled for to prevent reidentification; a consequence of this is attack rates in this study are likely lower than observed in the source data.

The resulting study dataset was processed by a machine learning model, Gradient Boosting Machine (GBM) using H2o.ai software. GBM guessed the clinical status of a given compartmentalized aggregate and then weighted and re-weighted its guesses to minimize its error over iterations. GBM uses tree models to build and rebuild its decision-making justifications (guesses) over model iterations. GBM has parameter requirements including the number of trees to produce and the depth of each tree branch. These features were learned computationally (grid search). The GBM was set to a maximum depth of 17 nodes per leaf, and to produce 50 trees while considering five cross validation runs. GBM attempted to understand what made HIV+ BSGS aggregate records different from HIV- BSGS aggregates. The resulting model was scored for accuracy, recall and variable importance. The final model was used to ‘predict’ the study dataset. This predictive step results in leaf-node assignment, or the study dataset row number assignment within a given GBM tree. This Leaf-Node-Assignment is the equitant of a cluster identification number in K-means, where a given cluster’s members are segmented together by the model to achieve decision-making.

**What you will find in this repository**

This study made several information sources that use several different kinds of information. Telling kinds of information and sources apart is key to proper use.

**Kinds of information**

DX: Diagnostic

Diagnostic information means any information encoded using a controlled vocabulary whose Athena Domain is ‘Condition’.

RX: Medication

Diagnostic information means any information encoded using a controlled vocabulary whose Athena Domain is ‘Drug’. At this time no Medication data is presented here.

**Kinds of information sources**

CMS: The Centers for Medicare and Medicaid Services (CMS) Virtual Research Data Center (VRDC) collection contains populated claim forms and administrative meta data describing individual providers, facilities, patients, care plans and transactions known to CMS. The data is sourced from Medicare, Medicaid, Child Health Insurance Program (CHIP) and Social Security Disability Insurance (SSDI) encounters among others. Third party data integration (National Death Index, Social Security Death Index) punctuates the product as well. Medicare includes Parts A, B, C and D. Medicaid includes both MAX(Y1999-Y2015) and TAF(Y2014-Ongoing) era claims. The VRDC provides ‘Research Identifiable Files’ (RIF) to authorized users. These RIF files contain complex clinical and financial information about individuals over time. You can learn more about CMS data here:

https://lhncbc.nlm.nih.gov/CCOI/datasets/CMS/index.html

Note there are several flavors of CMS data. We present ICD denominated data as ICD collection. ICD\_CL collection is ICD denominated data with some compartments and the case definition codes removed. This has the effect of lowering the amount of pruning one might perform. We also include PHE collection, which is phenotype denominated, as proposed to ICD denominated data. Individuals using Phenome Wide Association Studies methods should use this collection. For more information about PheWAS see:

https://phewascatalog.org/

CPRD: Clinical Practice Research Datalink (CPRD) is a is a real-world research service supporting retrospective and prospective public health and clinical studies. CPRD includes de-identified Electronic Health Record patient level data from a network of UK based general practitioners (GPs). You can learn more about CPRD here:

https://lhncbc.nlm.nih.gov/CCOI/datasets/CPRD/index.html

AOU: All of Us (AOU) is a program that integrates Electronic Health Records (EHR) with survey questionnaire to develop a diverse, information rich database that serves as a central point for many secondary research studies and reduce the need for developing individual single use study specific data collection protocols. You can learn more about AOU here:

https://lhncbc.nlm.nih.gov/CCOI/datasets/allofus/index.html

UKBB: The UK Biobank program is a large health and biomedical database that serves multiple retrospective, observational studies and includes over half a million participants between the ages of 40 and 69 from the United Kingdom. UK Biobank contains a combination of health, questionnaire and genetic data that is regularly updated and enriched with new data fields. You can learn more about UKBB here:

https://lhncbc.nlm.nih.gov/CCOI/datasets/UKbb/index.html