**Notes**

This study assumes that the underlying data has been transformed into an OMOP CDM like object. OMOP CDMs are distributed table architectures which usually reflect primary key linked relational tables stored in a traditional, on-site enterprise database. The example used in this directory is the largest CDM we have and in turn code that works on this is likely to work on smaller CDMs in our collection. The DX\_ICD\_CMS subset was the first, largest and most complicated. Not all CDM study code is presented here for brevity; but the DX\_ICD\_CMS code should be instructive, as AOR, CPRD and UKBB code is a 99% match to it with minor changes to reflect the data sources.

**Directory Contents**

00. Base File

Here Spark SQL is used to harvest data from our databricks instance (which holds the CMS CDM) and transform choice elements into the study base file. Not all of this code is necessary, depending on the cardinality of your OMOP CDM.

01. Base File Prep

The Base file needs to be prepared and several features need to be added to it.

02. Model

This code, is the CMS.r code except the user name has been redacted for privacy and security reasons. This is the code that BioWulf ran.

03. BioWulf Benchmark for CMS ICD

Figuring out how much memory H2o.ai, as a java program, needs within a high performance computing facility like BioWulf is a non trivial task. We do not recommend using interactive session as wall time and java heap space issues were observed. Note that H2o.ai does not use all memory allocated to it, and frugal memory allocation is dangerous; as it causes heap space issues. Smaller memory allocations with the largest dataset, CMS ICD, did not finish or aborted due to overflow issues. The session described in this directory completed, and reproduction may depend on overallocation of memory.

04. BioWulf Benchmark Configuration

Here you can see the actual batch script that called the code in 02\_Model. You can also reuse this batch script to get BioWulf to run R scripts if you like.