Overview
We created a causal discovery pipeline for use with the carevue subset of the MIMIC-III dataset [Johnson et al., 2022]. Custom R and command line scripts were written and run on DNANexus. Users define parameters of the causal search and our pipeline automates the data preprocessing steps, causal search, and data output visualization.

Prepare database
Provide a script that will stand up a full database of the data we're using (or other user submitted data, there is an understanding that public health data is typically all in the same format-similar to MIMIC-III). Shifting the data from large and unwieldy CSVs into a sql database will make further steps easier. This is being performed through RmariaDB, though the exact database being created will be a 'normal' sql database and agnostic to accession tools.

Parse user input
We created a custom yaml file to serve as template for user input. The user provides patient filtering criteria (e.g., based on diagnosis code or diagnosis (human-readable), a set of outcomes (e.g., 30-day mortality), and a set of features over which to perform the causal discovery search (e.g., common laboratory values). We are working on a custom R script to parse user input and perform a series of dataset transformations, joins, and variable selection steps.

For microbiological events, we subsetted the most common 50 organisms, including common pathogens such as Pseudomonas aeruginosa, Staphylococcus, and Enterobacteriaceae.

For laboratory values, we included the 50 most common lab tests, including common components like glucose, hemoglobin, and red blood cell counts.

Run tetrad query (causal discovery)
According to user input and generated data, run (a series) of causal discovery searches using tetrad. The user may have specified parameters of the causal search including algorithm, significance threshold, and conditional independence test.

Parse tetrad output
TODO