Project Description.

There is a dearth of clinical data available about COVID-19 patients. There are a few studies that describe (at a population level) the clinical symptoms and comorbidities of diagnosed COVID-19 patients, that have been hospitalized. Much interest has been expressed about the apparent high severity and mortality for COVID-19 patients with particular comorbidities. But the question remains, is that information of clinical significance and what actions should we take based upon that information. This is a problem of causality. There are tools and techniques in the Causal Inference space, including building Bayesian Networks that try to solve this problem. (Mckenzie, 2018)

I would like to try to build a Bayesian Network to help us understand the clinical significance of the comorbidities associated with severe COVID-19 cases. I am particularly interested in understanding a causal model for COPD, since the mortality rate in the Guan study was 25% for patients with COPD. (Guan W-jie, 2020) This is a particularly interesting comorbidity, because the numbers for mortality in China are dramatically different than for New York, and that fact might allow us to make inference about the best clinical allocation of ventilators.

I hope to combine data gleaned from research studies, and published literature on disease prevalence to build a useful clinical model, or at least guide us towards the questions we should be asking in order to have the information we need to make better clinical decisions. This specific problem is described as a data fusion problem (a component of causality). (Pearl, 2016)

Relevance

Clinical Decision Support (CDS) ultimately is about presenting the clinician with tools to select the best option for a specific patient, given the patient's circumstance. Today CDS is dominated by predictions based upon population data, which may or may not translate to the patient. It is the clinician's job to take input from multiple data sources, and to try to synthesize that information to make the best decision for the patient. Unfortunately, often this task is done by using the most well-known results of trials, which may not reflect the patient and may not combine sufficient information. If we can build causal models, it will allow us to tailor treatment to the patient and use counterfactual techniques to build "crash-test dummies" to estimate the impact upon the patient in advance of treatment.

In our current COVID-19 predicament where we are dealing with limited resources, like ventilators, causal effect is even more important because clinicians need to make the decision about who needs the ventilator.

Data Sources and Tools

Specifically, I plan to use tables from studies by Grasseli and Guan. (Giacomo Grasselli, Alberto Zangrillo, Alberto Zanella, & al, 2020) (Guan W-jie, 2020). I hope to combine this with information for population prevalence from a variety of sources like the CDC and the WHO, as well as data from simulations using probabilistic programming techniques.

I plan to use the BN learn library of R, the Casual Fusion package by Elias Bareinboim (Pearl, 2016), and perhaps Pyro libraries in Python.

Team.

Rose Glavin.

I collaborated with Dashyang Kachru and Ramya Mounika on scaping the data from the studies and putting it in a spreadsheet.

I have spoken with Jeremy Zucker about the project and approach. I am using many of Robert Ness's ideas and sample code as presented in his online Causal Inference class.

I will be tackling this problem alone, but maybe consult with others as necessary.