SOC555 Final Project: Proposal

Group Members
Jamie Chiu, Greg Guggenmos, Meet Patel, Nicholas Toth

Project Aim
Our aim is to build and examine the usefulness of different baseline models using sociodemographic variables in predicting health and wellbeing outcomes.

Existing research shows that sociodemographic variables are associated with health and wellbeing outcomes. However, much research exploring determinants of health largely ignores these factors, beyond simply adjusting or stratifying statistical analyses. Further, current practices typically only involve comparing models against pure-chance-models or an underperforming model where every variable is thrown in. Thus, there is much room for improvement in how model performance is evaluated.

Why is identifying a good baseline model important to predicting health and wellbeing outcomes? How is this related to limits of prediction?
Having an established baseline that is accepted in the field as a norm would lead to the identification of more useful predictors scientifically and clinically – where the variables/features need to be more meaningful above and beyond what we already know about a person's outcome given their sociodemographic background. Claims about a model or its features would not be considered meaningful or useful for solely beating a pure-chance dummy model.

We consider this to be a starting point for bridging the two cultures as discussed in class – where we want to use existing literature to inform our baseline models, so that any new models proposed will need to outperform what we already know in order to be considered meaningful.

From a practical standpoint, sociodemographics are routinely collected information and so, using it as a baseline also forces model-builders to justify the additional (costly) collection of information about an individual.

LISS Panel
In this proposed project, we will be utilizing the LISS Panel from Centerdata, an independent non-profit research institute at Tilburg University. The data set includes approximately 7,500 individuals over 5,000 households, aged 16 years and older with data points on socio-economic, demographic, health, and wellbeing variables.

Proposed Methodology
Build several baseline models and see which ones performs best:

- Try with different variables (e.g. using individual item responses; applying PCA first for dimensionality reduction; using different domains/constructs from the surveys)
- Try with different architecture (classifiers, neural networks)
- Try for cross-sectional prediction
- Try for longitudinal (within-person changes) prediction
- Try with different outcome variables (there are a few in the dataset we could use, such as diagnoses received, hospitalization in the past 12 months, self-reported wellbeing ratings, etc.)

To evaluate the performance of our baseline models:

- Compare performance to both a weighted random and a true random model (as a sanity check that our baseline is indeed better than chance)
- Compare constructed models to one another
- Test on held-out portions of the data set, consider using k-fold cross-validation

We are interested in seeing how performance differs based on the researcher choices made for a given model.

Potential limitations:
Generalisability of findings could be a potential issue. More U.S. adults have received mental health diagnoses (~25%) than adults in other high-income countries (<10% in France, Netherlands, Germany). However, the scope of this project is to examine the performance of different baseline models constructed using sociodemographic data, and the model frameworks generated can be applied to other datasets; especially since sociodemographic data are routinely collected in almost all datasets that can be used for health and wellbeing prediction tasks.

What we would like feedback on:
Thank you for your feedback on our previous proposal. In our updated proposal, we tried to more clearly position the project in relation to limits of prediction. We would be interested in your feedback on whether it is more aligned with the scope of the class, as well as hearing any suggested changes (additions or subtractions) you may have for our proposed project.