Arbitrary Choices Matter: A Case of Cohort Selection in N3C

Atefehsadat Haghighathoseini, Sai S. Vinnakota, Robert J. Thomas, Cara L. Frankenfeld, Timothy F. Leslie, Hua Min, Nirup M. Menon, Janusz Wojtusiak
George Mason University, Fairfax, VA, US

Introduction

When processing health data, random choices are often made with respect to cohort inclusion/exclusion criteria, as well as encoding and selection of variables. Researchers often make their best guess based on what seems reasonable, clinical knowledge and what was done in the previously published literature (that may also be based on someone's previous guess). It is rarely tested in the published literature how these choices affect quality of models, and if they lead to optional and unbiased models. Data analysis used by the healthcare insurance companies, health systems, clinicians at hospitals or ambulatory settings should be performed without bias. The users of such analytics should be aware of the limitations under which constructed models and results of analysis may not perform fairly. For example, too strict exclusion criteria may eliminate substantial portions of populations and thus prevent generalizability of the methods. The hypothesis tested here is that seemingly unimportant choices made during data processing may introduce selection bias and result in datasets with significantly different characteristics.

Methods

Data used in this analysis came from the National COVID Cohort Collaborative (N3C), and analyses were conducted within N3C enclave [1]. The goal was to create cohorts for subsequent application of machine learning methods to predict outcomes for hospitalized COVID-19 patients. A series of decisions needed to construct the cohort was applied to about 18M patients in N3C from August 2020 – December 2021. These decisions included: identification of COVID-19 positive cases, identification of inpatient hospitalization records, identification of COVID-19 related hospitalizations, and potential exclusion of records with inaccurate timestamps. The choices made for these decisions result in 24 potential datasets. The results presented here show a comparison of two extreme cases in which always the most selective or always the most inclusive choices were made (Figure 1).

Results: The most inclusive cohort was about 354K COVID-19 positive patients (left branch in Figure 1) and the most exclusive consisted of about 112K patients (right branch in Figure 1). The inclusive cohort was composed of 50% female patients overall, dropping by 3% in the exclusive cohort. The difference is extreme in the case of some states. For example, the percent of females in North Dakota and Hawaii drops to 0% in the exclusive cohort, compared to 66% and 41% in the most inclusive cohort, respectively. The analysis indicated that the inclusive and exclusive cohorts are statistically different ($p<0.001$) in the composition of gender, ethnicity, race, age and length of stay as tested with the Mann-Whitney U Test and Chi-squared test.

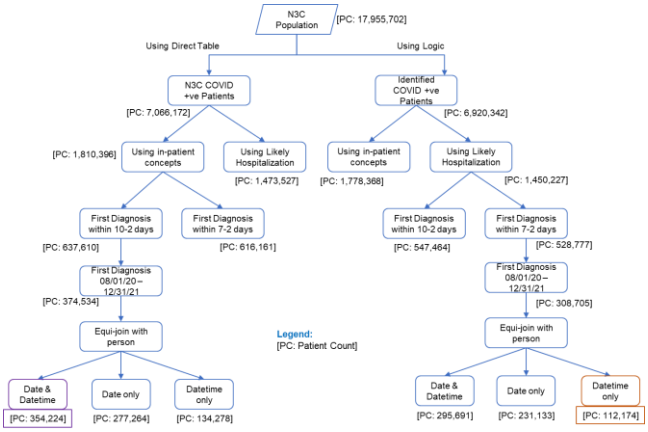

Figure 1: Partial exclusion tree outlining two extreme choices. PC is patient count.

Discussion and Conclusions

The results indicate that potentially arbitrary decisions in data preprocessing resulted in cohorts that are significantly different in size, have different characteristics and introduce bias. While more exclusive cohorts are preferred as giving higher quality of data for analytics and machine learning applications, one needs to carefully consider alternatives and the impact of such decisions. Further work is needed to evaluate the impact of the choices on potential biases within machine learning-based models constructed from the data.

References

1. National Institutes of Health (NIH). National Center for Advancing Translational Sciences (NCATS). National COVID Cohort Collaborative Data Enclave Repository. Bethesda, Maryland: U.S. Department of Health and Human Services, National Institutes of Health, 2023 [https://covid.cd2h.org]