Selection Bias from Data Processing in N3C

Atefehsadat Haghighathoseini
Department of Health Administration and Policy
George Mason University
Fairfax, Virginia
ahoseini@gmu.edu

Mohammad Qodrati
Department of Health Administration and Policy
George Mason University
Fairfax, Virginia
mqodrati@gmu.edu

Hua Min
Department of Health Administration and Policy
George Mason University
Fairfax, Virginia
hmin3@gmu.edu

Timothy Leslie
Geography & Geoinformation Science Department
George Mason University
Fairfax, Virginia
tleslie@gmu.edu

Cara Frankenfeld
Center for Interdisciplinary & Population Health Research
GeorMaineHealth Institute for Research
Scarborough, Maine
cara.frankenfeld@mainehealth.org

Nirup M Menon
School of Business
George Mason University
Fairfax, Virginia
nmenon@gmu.edu
On behalf of the N3C Consortium[1]

Janusz Wojtusiak
Department of Health Administration and Policy
George Mason University
Fairfax, Virginia
jwojtusi@gmu.edu

2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) | 979-8-3503-8373-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICHI61247.2024.00038

Abstract**—** *This study investigates potential selection bias in outcome prediction within the National COVID Cohort Collaborative (N3C) resulting from arbitrarily made decisions. In the processing of health data, decisions regarding cohort criteria and variable selection are often arbitrarily made, potentially introducing selection bias. This work explores if such decisions affect results of data analysis and potential conclusions of research studies. An experiment is conducted in which four arbitrary decisions are made. Results demonstrate significant differences in the obtained datasets and indicate a high potential for bias based on inclusion or exclusion decisions. The findings contribute to informed healthcare policies, better decision-making, and improved patient outcomes, emphasizing the necessity for testing assumptions and decisions in ongoing research that uses clinical data.*

Keywords**—** *Prediction, Selection Bias, Data Processing, National COVID Cohort Collaborative (N3C)*

I. Introduction

This work addresses the issues of decision made by researchers and data analysts when analyzing health data, and whether they may affect study conclusions. It is not intended to provide "perfect" set of decisions which does not exist, instead argues that when designing a study, one needs to consider and examine alternatives and their effect on results.

In the processing of health data, decisions related to cohort and variable selection are often made in a non-standardized manner or in a way that lacks reproducibility, relying on researchers' intuition and existing literature. This common approach creates a significant gap in evaluating the impact of such choices on data analysis models. Healthcare stakeholders need heightened awareness to conduct unbiased data analyses, acknowledging potential limitations arising from arbitrary decisions. Neglecting seemingly minor choices during data processing can lead to selection bias and dataset alterations. Strict inclusion/exclusion criteria may unintentionally omit important subpopulations, compromising generalizability. On the other hand, relaxed inclusion/exclusion criteria may lead to noise in the data. A rigorous evaluation of these choices is an important step in data processing.

Selection bias is a type of bias that results from the selection of a cohort that does not closely represent the greater population for which the study is conducted and results in reduced external validity or generalizability. It is introduced by the selection of individuals, groups, or data for analysis in such a way that proper randomization is not achieved, thereby failing to ensure that the sample obtained is representative of the population [1]. James Heckman's work *"Varieties of Selection Bias"* in econometrics discusses the impact of selection bias on estimating the impact of certain variables [2]. Additionally, selection bias can lead to inflated effect sizes and inaccurate results, ultimately impacting the reliability of statistical tests and the validity of research outcomes [3].

To minimize selection bias, researchers can use randomization or probability sampling techniques to ensure that all eligible participants have an equal chance of being included in the sample [4], though it is only possible when underlying distributions in the population are known. Also, adjusting for selection bias may involve the construction of a model that incorporates additional bias-breaking variables to account for differences between the study population and the target population [5]. Another strategy is to adjust for factors that can break the biasing paths linking the exposure and the outcome, thereby controlling for selection bias in cohort studies [6].

Electronic medical records (EMR) can provide a rich source of data to evaluate health outcomes. However, the processing of EMRs and other health data in clinical settings can introduce various forms of bias, which can significantly impact the results of data analysis, subsequent healthcare policies, and patient outcomes. These biases can arise from

[1] Authorship was determined using ICMJE recommendations. The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave covid.cd2h.org/enclave and supported by CD2H - The National COVID Cohort Collaborative (N3C) IDeA CTR Collaboration 3U24TR002306-04S2 NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data from participating organizations (covid.cd2h.org/dtas) and the organizations and scientists (covid.cd2h.org/duas) who have contributed to the on-going development of this community resource. See Hanendel et al 2021 for details.

2575-2634/24/$31.00 ©2024 IEEE
DOI 10.1109/ICHI61247.2024.00038

arbitrary decisions made during cohort criteria and variable selection, patient selection, and data completeness [7], [8], [9], [10]. Selection bias, for instance, can occur during the process of defining populations and outcomes, or during the linkage of different databases. Missed and false links during the linkage process can lead to missing data or misclassification [8]. Attrition bias, another form of bias, refers to systematic differences between groups in withdrawals from a study, leading to incomplete outcome data [11].

A significant amount of work has been done to study and mitigate selection bias in traditional statistics and experimental studies. For example, in a study where participants made arbitrary choices between two equally preferred options, it was found that these choices influenced their preferences [12]. This suggests that such arbitrary decisions can have significant impacts on the results of a study. Bias can also have a major impact on clinical practice, research, and decision-making. For instance, implicit bias in healthcare can lead to disparities which affects patient outcomes [13], [14]. Further, errors in selecting study participants, including non-response bias and under-/over-representation, introduce selection bias [15]. Addressing systematic differences between groups and attrition bias requires rigorous criteria for patient selection, standardized/blinded data collection, and inclusion of all randomized participants in final analyses [16] [17].

In order to mitigate these biases in experimental studies, several strategies have been proposed. These include using objective criteria for recruitment, blind evaluations, and salary disclosures [14]. Additionally, a general framework for addressing selection bias in EHR-based settings has been proposed, which involves grounding analysis in a pre-specified ideal study and decomposing data provenance into manageable components [9].

In the context of outcome prediction, selection bias can lead to systematic errors in the predictions made by machine learning models. For example, if a model is trained on data that is not representative of the population, it may make inaccurate predictions when applied to the actual population [18]. A significant amount of work has been done in to address bias in machine learning, but little work exists in selection bias. Suri [19] and MacNamee [20] both highlight the issue of bias in machine learning systems for outcome prediction, particularly in the context of cardiovascular disease risk assessment and anticoagulant drug therapy. Suri [19] emphasizes the need for stronger outcomes and multiethnic group representation, while MacNamee [20] suggests stratified sampling and boosting as potential solutions. Fernández-Castilla [21] and Zhu [22] further explore the issue of bias in meta-analyses and gene selection in supervised classification, respectively, underscoring the complexity of the problem and the need for ongoing research to address it. Various publications have extensively delved into the wide range of datasets available for analyzing bias in clinical data [23], [24]. Researchers have explored various data sources such as electronic health records (EHRs), claims data, and registries [25], [26], [27]. This diverse array of datasets facilitates thorough analysis and assessment of biases, empowering researchers to achieve a more profound comprehension of factors impacting the quality and dependability of clinical data. Recognizing these biases is crucial for devising interventions and policies that uphold health equity.

Thus, understanding and addressing biases in health data processing is crucial for improving healthcare policies and patient outcomes. Selection bias in clinical data arises from errors in study design and patient recruitment [16]. This issue is particularly problematic in case-control and retrospective cohort studies. Performance bias involving differential treatment, care, and follow-up during a trial can contribute to bias [28].

Research presented here specifically focuses on selection bias and demonstrates potential consequences of applying seemingly arbitrary decisions when processing data. The main part of the work aims at illustrating the problem in the context of National COVID Cohort Collaborative (N3C) data. The presented research is part of a larger project aimed at understanding biases and fairness of machine learning models applied to prediction of outcomes for hospitalized patients, specifically in the context of N3C. The following sections discuss data, choices made, and results indicating impact of these choices.

II. Methods

N3C is a crucial resource in the realm of COVID-19 research providing a comprehensive dataset for analyses within its dedicated platform, the N3C Data Enclave [29]. The effort aims to leverage diverse data from various healthcare facilities and provides a coherent framework for researchers towards quality information. More specifically, the presented study used Limited Data Set ("LDS") which provides access to individual patient-level data with removed patient identifiers.

The problem considered here is one of predicting outcomes for patients hospitalized for COVID-19, and potential biases in such prediction. The unit of analysis in the work is patient hospitalization, for which an outcome is predicted. The N3C data are organized in multiple tables in a relational database following the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standard that need to be converted to format in which one hospitalization is exactly one record of data (flat table

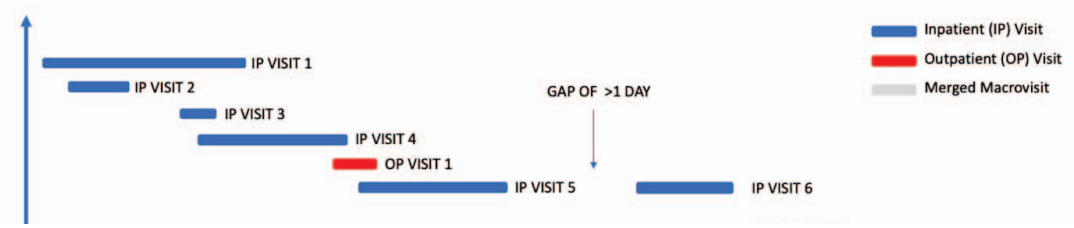

Fig. 1. a demonstration of how N3C defines the concept of macro-visit incorporating a number of small visits, and the criterion of delineation which is a gap of at least one day between two sequential visits. Taken from [31].

sometimes referred to as analytic file). Such format is typical for applying machine learning methods as well as most traditional statistical tools. Within the inpatient data "Macro-visit" is a key concept used in the N3C [30], defined as the merge of chronological, overlapping inpatient and other longitudinal facility visits, with further addition of outpatient visits, telehealth appointments, and any other types of visits occurring during the whole interval (Fig. 1). Thus, a macro-visit makes the hospitalizations out of the inpatient visits, observation stays, multi-day facility stays following outpatient hospital procedures, emergency room visits, and co-occurring outpatient and telehealth visits during the COVID-19 surge [30].

First, and one of the most important steps of any data analysis, including machine learning model construction, is cohort construction. To construct the cohorts, a series of decisions were applied to about 21 million N3C patients. Depicted in Fig. 2, these decisions included identification of COVID-19 cases, identification of inpatient hospitalization records, identification of COVID-19 related hospitalizations, and potential exclusion of records with missing data about specific time of admission. The figure also includes counts of care sites (CC), which may seem like a good indicator of cohort distribution, yet may be misleading due to very high number of missing values. The seemingly arbitrary choices can result in 16 potential datasets of different sizes and properties. The presented work focuses on analyzing the extreme cases of the most inclusive and most exclusive cohort selection, but can be generalized to all potential datasets.

Decision 1: Identifying COVID-19 patients - There have been a wide range of tests and choices available that allow for identifying COVID-19 positive cases. Some patients are laboratory-confirmed cases with details of the tests and results available in N3C. Others have only assigned with the diagnosis code. We considered two choices:

Choice A: Only people with a positive COVID-19 lab test result - This involves getting the records that are of a positive value for the lab tests indicating COVID-19 infection. Diverse ways that care provider sites use to report a such results (e.g., Positive, Detected, Reactive, Presumptive Positive, etc.) were gathered in the N3C's "ResultPos" concept set. The lab tests were composed of SARS-CoV-2 RNA PCR or *antigen test*, as well as the antibody retry test.

Choice B: People with positive lab test result or those with coded diagnosis [31] - These records are indicated by the ICD-10-CM code of U07.1 indicating "COVID-19 with virus identified".

Decision 2: Identifying hospitalization records among patient encounters - We investigated two options:

Choice A: Wildcard search for the visit concepts that incorporate "inpatient", "observ", and "hospital". A total of 8,374 concepts were identified and manually reviewed for accuracy.

Choice B: Use of the N3C-defined variable which marks the records of hospitalization - It is calculated by N3C team as encounter entered by a reliable site and/or the records with either an attributed diagnosis-related group (DRG), a Centers

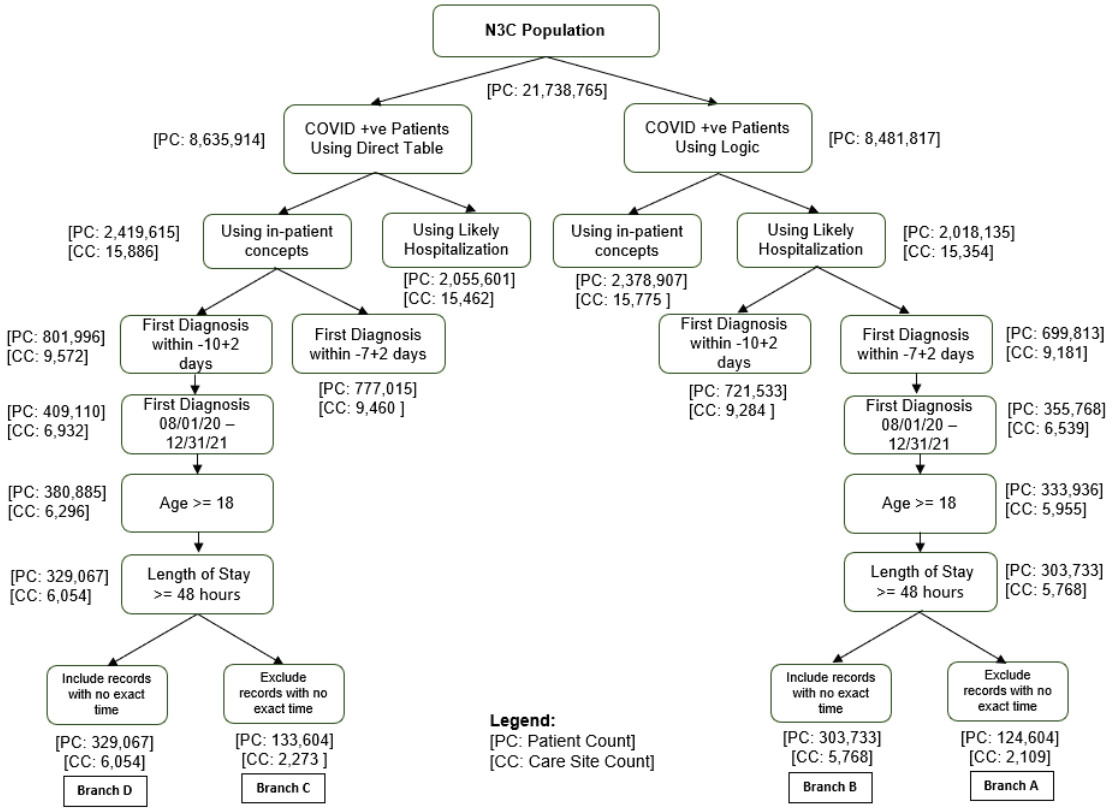

Fig. 2. Partial inclusion/exclusion tree outlining extreme choices for the four decisions. PC indicates patient counts, CC indicates caresite counts.

for Medicare & Medicaid Services (CMS) inpatient-only procedure, an inpatient Evaluation and Management (E&M) Healthcare Common Procedure Coding System (HCPC) or concept, an inpatient intensive critical unit (ICU) HCPC or concept, or a minimum of 50 combined attributed resources. Resources in this context are comprised of all attributed diagnoses, procedures, labs, and medications [30].

Decision 3: Selecting time window for a COVID-19 positive test to be considered relevant to a macro-visit - We compared a (Choice A) *7-day* with a (Choice B) *10-day* window prior to the beginning of the inpatient records, and until the second day of the hospitalization. These two time windows are depicted in Fig. 3.

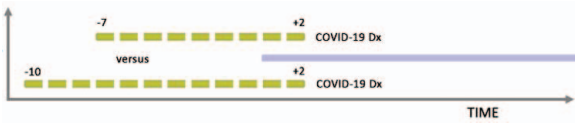

Fig. 3. Two time windows used as options in COVID-19 identification

Decision 4: The last decision point is purely data processing-related. One may treat records without an exact time string (i.e., those with only a date) of admission to have a missing value, possibly dropping it, rather than keeping all records regardless of their timestamp string. One of the criteria in the current project would be to consider data during either a 48-hour or mathematically (but not practically) equivalent 2-day period for further analysis.

Choice A: Keeping only those records that include exact timestamp of admission - This allows for precise calculation of 48-hour observation window after admission that is used for further analysis. Dropping such records from cohorts could introduce issues in generalizability of the results to the population.

Choice B: Inclusion of all records - This choice requires calculation of 2-day hospitalization period that may not align with the exact 48-hours, i.e., calculation of midnight to midnight using only date.

Other decisions: The data preprocessing included other decisions that were not investigated for simplicity of the presented work. After decision 3, data were filtered to include only those records with a first COVID-19 diagnosis between August 1, 2020, and December 31, 2021, inclusively, which constitutes our study target period. Data prior to August 2020 are not considered reliable as corresponding to early efforts in the pandemic, and those after 2021 were removed to keep the cohort fixed in time. There is no specific reason why these exact dates were selected, albeit reasonable. Subsequently, hospitalization records are merged with other basic data to be eventually analyzed using the key unique identifier, while the features include gender, date of birth, race, ethnicity, and age at death. At the same time, non-adult patients were excluded from all of the cohorts. We used patients 18 years and older, although there is no specific clinical reason to use 18, instead of 21 or some other numbers.

III. Results & Discussions

Application of the four decisions results in 16 potential datasets. As one can argue for which option is the best at each decision level, it is not clear which of the datasets is to be used in the final analysis. The sections below compare two "most extreme" datasets, as well as their two companions to specifically look at the impact of timestamp selection (Decision 4). These decisions along with corresponding annotation are depicted in the previously described Fig. 2.

Cohort Sizes

There is a clear difference in size of the cohorts, ranging from 124,604 to 329,067 patients, indicating an almost three-fold increase in population size. The largest impact on the cohort size results from the last decision that is purely data-driven (see Fig. 2), which accounts for approximately 60% drop in cohorts. In addition, there is a significant difference in number of care sites in different branches of the study, yet as noted before die to missingness these values may not be representative.

While there is 3-fold difference in cohort sizes, it is not immediately clear if that difference results in different data characteristics.

Geographical Distribution

To discern the most notable distinctions among cohorts, it is a common practice to analyze the percentages within each state or geographical region. Higher percentages within a state imply a more significant presence of a particular cohort, whereas lower percentages suggest a lower representation. The differences of where patients come from are directly linked to the number of distinct care sites in the data. It is important to emphasize that during the application of both the logic and the LDS table to identify COVID-19-positive patients, we encountered nearly eight million instances of missing care site data.

There are differences across the four cohorts in terms of geographic distribution of patient counts. After New York (NY) which has the highest relative counts across all cohorts, North Carolina (NC) is the second most commonly represented state in Cohort D, while Illinois (IL) is the counterpart of which in Cohort A. Additionally, in Cohort C, Hawaii (HI), and North Dakota (ND) consistently demonstrate lower patient counts.

Fig. 4 illustrates cohorts and decisions across states. Note that the figures exclude a significant portion of data, the states for which were not available or marked as other (locations outside the individual U.S. states, encompassing Mexico, Canada, Central American countries, and five territories - American Samoa, Guam, the Northern Mariana Islands, Puerto Rico, and the U.S. Virgin Islands).

Directly following the cohort size, the highest impact on geography was the last decision, that eliminated a significant portion of data coming from Indiana.

Demographics

In the inclusive cohorts in terms of the timestamp decision (Cohorts B & D) there were more males, Asians, and Hispanic/Latinos or unknown-ethnicity than their complements. Table I contains descriptive statistics for gender, race, and ethnicity across four cohorts, revealing distinct disparities among the groups.

The age distribution in the four cohorts seems similar (**Error! Reference source not found.**), but closer examination shows statistically significant differences. There are also different gender distributions. As can be seen, Cohort D has the lowest percentage of males and the highest percentage of females.

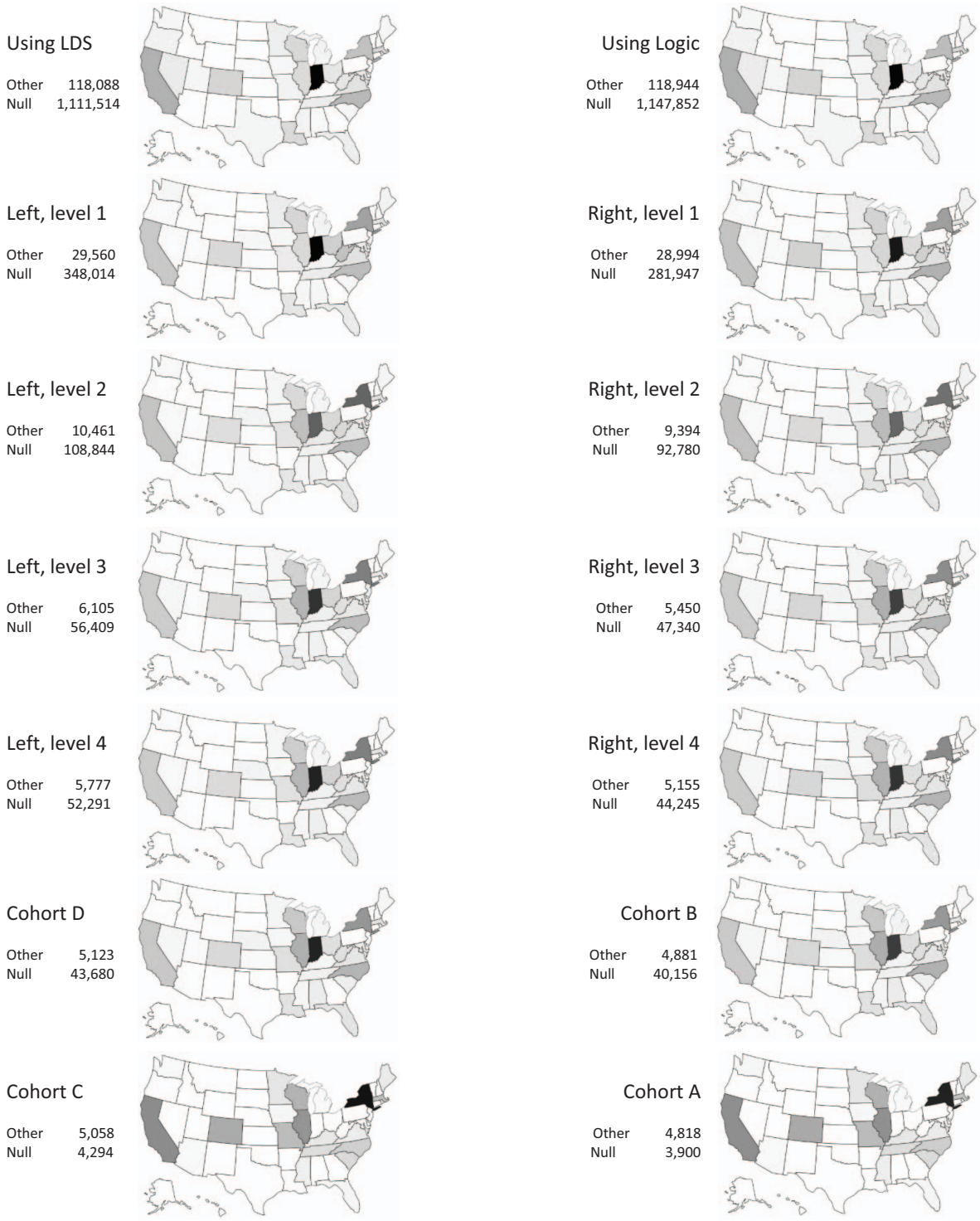

Fig. 4. Geographic distribution of data following the data processing decisions.

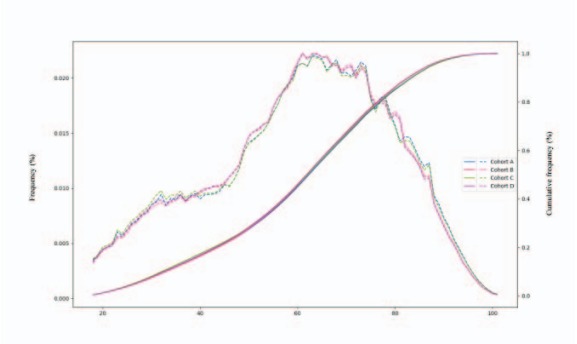

Fig. 5. Distributions of the COVID-19 patients along the years of age (The graphs are truncated at age 102 years following the data providers stipulation to not show counts of less than 20.)

Examining the racial distribution across cohorts, we observe Cohort D generally has the highest counts across all racial categories, indicating a larger overall patient population in that cohort. Cohort A consistently has the lowest counts across most racial categories. The distribution of races is significantly different among the cohorts, as indicated by the low p-values from the Chi-squared test. This suggests that the racial composition varies significantly between the cohorts. Cohort D has also the highest number of patients across all categories of ethnicity, which is consistent with it being the largest cohort. Cohort A has the lowest counts of Hispanic/Latino and NOT Hispanic/Latino patients, and also the lowest count of patients with unknown ethnicity. However, when considering the proportions relative to the total number of patients in each cohort, Cohort A has a higher percentage of Hispanic/Latino patients compared with Cohort D.

Outcome Differences

In addition to conducting another analysis, our objective was to pinpoint outcome disparities within Cohorts A, B, C, and D. These cohorts are distinguished by the percentages of individuals in each group, categorized according to an "Expiry Flag", where *0* denotes survival and *1* indicates death.

In Cohort A, 85.06% survived and 14.94% expired. Cohorts B, C, and D exhibit similar trends, with survival rates of 83.51%, 85.44%, and 83.96%, respectively. The Chi-square test with a p-value of 0.000 for all cohorts indicates a significant association between cohort membership and the Expiry Flag. Despite small variations in survival and death percentages among cohorts, the consistent low p-values suggest that these differences are not due to chance. Therefore, there is a statistically significant association between cohort membership and survival outcome. Further investigation may be needed to identify the factors contributing to the observed variations in survival rates among the cohorts.

Variable-by-Variable Comparison

Understanding and addressing potential selection bias based on gender, race, ethnicity, and geographical location is crucial for developing predictive models within the N3C. Adjusting for these demographic factors using appropriate techniques is essential to ensure that predictive models are not biased and can be effectively generalized across diverse cohorts.

The analysis indicated that the cohorts are statistically different ($p<0.001$) in the composition of age, gender, race, ethnicity (Table I), and length of stay as tested with the Mann-Whitney *U* Test and the Chi-squared test.

Comparing the same cohort across different states can unveil variations. Notably, NY consistently exhibits elevated percentages across all cohorts, hinting at possible regional disparities. States with the lowest percentages, such as ND and Vermont (VT), and HI may suffer from limited representation, potentially introducing selection bias into the analysis.

IV. Conclusions

The findings reveal that potentially arbitrary decisions in data processing stage can result in significantly different cohort sizes and characteristics, introducing biases that may impact the quality of research conclusions. While there is a prevailing preference for more exclusive cohorts, caution is advised, as exclusions may lead to limited representativeness and potential fairness issues. The study highlights the need for further research to assess the consequences of preprocessing choices on biases within machine learning models and to explore tailored strategies for mitigation of them. Moreover, our future investigations will delve into specific domains where biases are more pronounced and examine the impact of different preprocessing approaches on generalizability, providing valuable insights for enhancing external validity of studies. Ultimately, adopting an informed and careful approach to data preprocessing decisions is crucial for advancing reliability and fairness of machine learning applications across diverse domains.

TABLE I. COHORTS AND THEIR GENDER, RACE, AND ETHNICITY SUBGROUPS

Cohort ID		Gender*			Ethnicity*			Race*					
	Patients count	*Male*	*Female*	*Unknown*	*Hispanic or Latino*	*Not Hispanic or Latino*	*Unknown*	*White*	*Black*	*Asian*	*Native Hawaiian*	*Multiple races*	*Unknown*
A	124,604	64,842 (52.03%)	59,750 (47.95%)	<20 (0.00%)	22,430 (18.00%)	96,314 (77.29%)	5,860 (4.70%)	73,180 (58.73%)	21,942 (17.60%)	4,734 (3.79%)	310 (0.24%)	130 (0.10%)	24,308 (19.50%)
B	303,733	155,624 (51.23%)	148,074 (48.75%)	35 (0.01%)	38,638 (12.72%)	244,658 (80.55%)	20,437 (6.72%)	191,473 (63.03%)	55,195 (18.17%)	8,748 (2.88%)	1,452 (0.47%)	3,276 (1.07%)	43,589 (14.35%)
C	133,604	68,995 (51.64%)	64,597 (48.34%)	<20 (0.00%)	24,432 (18.28%)	102,953 (77.05%)	6,219 (4.65%)	77,992 (58.37%)	23,182 (17.35%)	5,585 (4.18%)	328 (0.24%)	137 (0.10%)	26,380 (19.74%)
D	329,067	167,545 (50.91%)	161,486 (49.07%)	36 (0.01%)	42,456 (12.90%)	264,682 (80.43%)	21,929 (6.66%)	207,817 (63.15%)	58,987 (17.92%)	9,907 (3.01%)	1,508 (0.45%)	3,603 (1.09%)	47,245 (14.35%)
* All P-values < 0.001 using Pearson's Chi-squared test													

The study uses data from August 1, 2020, to December 31, 2021, for analysis. The findings may not capture potential changes in healthcare practices, data reporting, or patient characteristics over time. Not all data preprocessing decisions depicted in the decision tree were investigated yet. Third, the choice of using patients 18+ is arbitrary, as one may argue that should 21+ be used, is as good.

Finally, the current efforts of our team are to investigate the impact of the choices on quality of machine learning-based models induced from the data. While the work presented here clearly indicates that cohorts are different, it is still not clear how much it matters for machine learning.

References

[1] J. J. Heckman, "Selection Bias," in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, Ed., New York: Elsevier, 2005, pp. 463–468. doi: 10.1016/B0-12-369398-5/00115-8.

[2] J. Heckman, "Varieties of Selection Bias," *Am. Econ. Rev.*, vol. 80, no. 2, pp. 313–318, 1990.

[3] H. Lu, S. R. Cole, C. J. Howe, and D. Westreich, "Toward a clearer definition of selection bias when estimating causal effects," *Epidemiol. Camb. Mass*, vol. 33, no. 5, pp. 699–706, Sep. 2022, doi: 10.1097/EDE.0000000000001516.

[4] J. Heckman, H. Ichimura, J. Smith, and P. Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, vol. 66, no. 5, pp. 1017–1098, 1998, doi: 10.2307/2999630.

[5] S. Geneletti, S. Richardson, and N. Best, "Adjusting for selection bias in retrospective, case–control studies," *Biostatistics*, vol. 10, no. 1, pp. 17–31, Jan. 2009, doi: 10.1093/biostatistics/kxn010.

[6] E. A. Nohr and Z. Liew, "How to investigate and adjust for selection bias in cohort studies," *Acta Obstet. Gynecol. Scand.*, vol. 97, no. 4, pp. 407–416, Apr. 2018, doi: 10.1111/aogs.13319.

[7] J. Lambert, "Statistics in Brief: How to Assess Bias in Clinical Studies?," *Clin. Orthop.*, vol. 469, no. 6, pp. 1794–1796, Jun. 2011, doi: 10.1007/s11999-010-1538-7.

[8] R. J. Shaw *et al.*, "Biases arising from linked administrative data for epidemiological research: a conceptual framework from registration to analyses," *Eur. J. Epidemiol.*, vol. 37, no. 12, pp. 1215–1224, 2022, doi: 10.1007/s10654-022-00934-w.

[9] S. Haneuse and M. Daniels, "A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why?," *eGEMs*, vol. 4, no. 1, p. 1203, Aug. 2016, doi: 10.13063/2327-9214.1203.

[10] R. Kundu, X. Shi, J. Morrison, J. Barrett, and B. Mukherjee, "A Framework for Understanding Selection Bias in Real-World Healthcare Data." arXiv, Aug. 17, 2023. doi: 10.48550/arXiv.2304.04652.

[11] J. P. Higgins and S. Green, Eds., *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series*, 1st ed. Wiley, 2008. doi: 10.1002/9780470712184.

[12] K. Nakamura and H. Kawabata, "I Choose, Therefore I Like: Preference for Faces Induced by Arbitrary Choice," *PLoS ONE*, vol. 8, no. 8, p. e72071, Aug. 2013, doi: 10.1371/journal.pone.0072071.

[13] "Quick Safety Issue 23: Implicit bias in health care | The Joint Commission." Accessed: Jan. 16, 2024. [Online]. Available: https://www.jointcommission.org/resources/news-and-multimedia/newsletters/newsletters/quick-safety/quick-safety-issue-23-implicit-bias-in-health-care/

[14] D. P. Gopal, U. Chetty, P. O'Donnell, C. Gajria, and J. Blackadder-Weinstein, "Implicit bias in healthcare: clinical practice, research and decision making," *Future Healthc. J.*, vol. 8, no. 1, pp. 40–48, Mar. 2021, doi: 10.7861/fhj.2020-0233.

[15] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali, "Selection bias and information bias in clinical research," *Nephron Clin. Pract.*, vol. 115, no. 2, pp. c94-99, 2010, doi: 10.1159/000312871.

[16] C. J. Pannucci and E. G. Wilkins, "Identifying and Avoiding Bias in Research," *Plast. Reconstr. Surg.*, vol. 126, no. 2, p. 619, Aug. 2010, doi: 10.1097/PRS.0b013e3181de24bc.

[17] "Types of Bias in Randomized Controlled Trials: A Refresher for Military Mental Health Providers," Military Health System. Accessed: Jan. 17, 2024. [Online]. Available: https://www.health.mil/Military-Health-Topics/Centers-of-Excellence/Psychological-Health-Center-of-Excellence/Clinicians-Corner-Blog/Types-of-Bias-in-Randomized-Controlled-Trials-A-Refresher-for-Military-Mental-Health-Providers

[18] K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim, "Mitigating bias in machine learning for medicine," *Commun. Med.*, vol. 1, no. 1, Art. no. 1, Aug. 2021, doi: 10.1038/s43856-021-00028-w.

[19] J. S. Suri *et al.*, "Understanding the bias in machine learning systems for cardiovascular disease risk assessment: The first of its kind review," *Comput. Biol. Med.*, vol. 142, p. 105204, Mar. 2022, doi: 10.1016/j.compbiomed.2021.105204.

[20] B. Mac Namee, P. Cunningham, S. Byrne, and O. I. Corrigan, "The problem of bias in training data in regression problems in medical decision support," *Artif. Intell. Med.*, vol. 24, no. 1, pp. 51–70, Jan. 2002, doi: 10.1016/s0933-3657(01)00092-6.

[21] B. Fernández-Castilla, L. Declercq, L. Jamshidi, S. N. Beretvas, P. Onghena, and W. Van den Noortgate, "Detecting Selection Bias in Meta-Analyses with Multiple Outcomes: A Simulation Study," *J. Exp. Educ.*, vol. 89, no. 1, pp. 125–144, Jan. 2021, doi: 10.1080/00220973.2019.1582470.

[22] X. Zhu, C. Ambroise, and G. J. McLachlan, "Selection bias in working with the top genes in supervised classification of tissue samples," *Stat. Methodol.*, vol. 3, no. 1, pp. 29–41, Jan. 2006, doi: 10.1016/j.stamet.2005.09.011.

[23] K. Mise and W. Iwasaki, "Environmental Atlas of Prokaryotes Enables Powerful and Intuitive Habitat-Based Analysis of Community Structures," *iScience*, vol. 23, no. 10, p. 101624, Oct. 2020, doi: 10.1016/j.isci.2020.101624.

[24] M. Wich, T. Eder, H. Al Kuwatly, and G. Groh, "Bias and comparison framework for abusive language datasets," *AI Ethics*, vol. 2, no. 1, pp. 79–101, 2022, doi: 10.1007/s43681-021-00081-0.

[25] A. F. Karr *et al.*, "Comparing record linkage software programs and algorithms using real-world data," *PloS One*, vol. 14, no. 9, p. e0221459, 2019, doi: 10.1371/journal.pone.0221459.

[26] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "Capturing Semantic Relationships in Electronic Health Records Using Knowledge Graphs: An Implementation Using MIMIC III Dataset and GraphDB," *Healthc. Basel Switz.*, vol. 11, no. 12, p. 1762, Jun. 2023, doi: 10.3390/healthcare11121762.

[27] M. Yuksel *et al.*, "An Interoperability Platform Enabling Reuse of Electronic Health Records for Signal Verification Studies," *BioMed Res. Int.*, vol. 2016, p. 6741418, 2016, doi: 10.1155/2016/6741418.

[28] N. Pandis, "Sources of bias in clinical trials," *Am. J. Orthod. Dentofacial Orthop.*, vol. 140, no. 4, pp. 595–596, Oct. 2011, doi: 10.1016/j.ajodo.2011.06.013.

[29] "N3C - Home." Accessed: Dec. 21, 2023. [Online]. Available: https://covid.cd2h.org/

[30] P. Leese *et al.*, "Clinical encounter heterogeneity and methods for resolving in networked EHR data: A study from N3C and RECOVER programs," *medRxiv*, p. 2022.10.14.22281106, Oct. 2022, doi: 10.1101/2022.10.14.22281106.

[31] T. D. Bennett *et al.*, "The National COVID Cohort Collaborative: Clinical Characterization and Early Severity Prediction." medRxiv, p. 2021.01.12.21249511, Jan. 13, 2021. doi: 10.1101/2021.01.12.21249511.

ACKNOWLEDGEMENTS

We gratefully acknowledge the following core contributors to N3C:
Adam B. Wilcox, Adam M. Lee, Alexis Graves, Alfred (Jerrod) Anzalone, Amin Manna, Amit Saha, Amy Olex, Andrea Zhou, Andrew E. Williams, Andrew Southerland, Andrew T. Girvin, Anita Walden, Anjali A. Sharathkumar, Benjamin Amor, Benjamin Bates, Brian Hendricks, Brijesh Patel, Caleb Alexander, Carolyn Bramante, Cavin Ward-Caviness, Charisse Madlock-Brown, Christine Suver, Christopher Chute, Christopher Dillon, Chunlei Wu, Clare Schmitt, Cliff Takemoto, Dan Housman, Davera Gabriel, David A. Eichmann, Diego Mazzotti, Don Brown, Eilis Boudreau, Elaine Hill, Elizabeth Zampino, Emily Carlson Marti, Emily R. Pfaff, Evan French, Farrukh M Koraishy, Federico Mariona, Fred Prior, George Sokos, Greg Martin, Harold Lehmann, Heidi Spratt, Hemalkumar Mehta, Hongfang Liu, Hythem Sidky, J.W. Awori Hayanga, Jami Pincavitch, Jaylyn Clark, Jeremy Richard Harper, Jessica Islam, Jin Ge, Joel Gagnier, Joel H. Saltz, Joel Saltz, Johanna Loomba, John Buse, Jomol Mathew, Joni L. Rutter, Julie A. McMurry, Justin Guinney, Justin Starren, Karen Crowley, Katie Rebecca Bradwell, Kellie M. Walters, Ken Wilkins, Kenneth R. Gersing, Kenrick Dwain Cato, Kimberly Murray, Kristin Kostka, Lavance Northington, Lee Allan Pyles, Leonie Misquitta, Lesley Cottrell, Lili Portilla, Mariam Deacy, Mark M. Bissell, Marshall Clark, Mary Emmett, Mary Morrison Saltz, Matvey B. Palchuk, Melissa A. Haendel, Meredith Adams, Meredith

Temple-O'Connor, Michael G. Kurilla, Michele Morris, Nabeel Qureshi, Nasia Safdar, Nicole Garbarini, Noha Sharafeldin, Ofer Sadan, Patricia A. Francis, Penny Wung Burgoon, Peter Robinson, Philip R.O. Payne, Rafael Fuentes, Randeep Jawa, Rebecca Erwin-Cohen, Rena Patel, Richard A. Moffitt, Richard L. Zhu, Rishi Kamaleswaran, Robert Hurley, Robert T. Miller, Saiju Pyarajan, Sam G. Michael, Samuel Bozzette, Sandeep Mallipattu, Satyanarayana Vedula, Scott Chapman, Shawn T. O'Neil, Soko Setoguchi, Stephanie S. Hong, Steve Johnson, Tellen D. Bennett, Tiffany Callahan, Umit Topaloglu, Usman Sheikh, Valery Gordon, Vignesh Subbian, Warren A. Kibbe, Wenndy Hernandez, Will Beasley, Will Cooper, William Hillegass, Xiaohan Tanner Zhang. Details of contributions available at covid.cd2h.org/core-contributors

The following institutions whose data is released or pending:

Available: Advocate Health Care Network — UL1TR002389: The Institute for Translational Medicine (ITM) • Boston University Medical Campus — UL1TR001430: Boston University Clinical and Translational Science Institute • Brown University — U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Carilion Clinic — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • Charleston Area Medical Center — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • Children's Hospital Colorado — UL1TR002535: Colorado Clinical and Translational Sciences Institute • Columbia University Irving Medical Center — UL1TR001873: Irving Institute for Clinical and Translational Research • Duke University — UL1TR002553: Duke Clinical and Translational Science Institute • George Washington Children's Research Institute — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • George Washington University — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Indiana University School of Medicine — UL1TR002529: Indiana Clinical and Translational Science Institute • Johns Hopkins University — UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • Loyola Medicine — Loyola University Medical Center • Loyola University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Maine Medical Center — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Massachusetts General Brigham — UL1TR002541: Harvard Catalyst • Mayo Clinic Rochester — UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Medical University of South Carolina — UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • Montefiore Medical Center — UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Nemours — U54GM104941: Delaware CTR ACCEL Program • NorthShore University HealthSystem — UL1TR002389: The Institute for Translational Medicine (ITM) • Northwestern University at Chicago — UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • OCHIN — INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks • Oregon Health & Science University — UL1TR002369: Oregon Clinical and Translational Research Institute • Penn State Health Milton S. Hershey Medical Center — UL1TR002014: Penn State Clinical and Translational Science Institute • Rush University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Rutgers, The State University of New Jersey — UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Stony Brook University — U24TR002306 • The Ohio State University — UL1TR002733: Center for Clinical and Translational Science • The State University of New York at Buffalo — UL1TR001412: Clinical and Translational Science Institute • The University of Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Iowa — UL1TR002537: Institute for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine — UL1TR002736: University of Miami Clinical and Translational Science Institute • The University of Michigan at Ann Arbor — UL1TR002240: Michigan Institute for Clinical and Health Research • The University of Texas Health Science Center at Houston — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • The University of Texas Medical Branch at Galveston — UL1TR001439: The Institute for Translational Sciences • The University of Utah — UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center — UL1TR002544: Tufts Clinical and Translational Science Institute • Tulane University — UL1TR003096: Center for Clinical and Translational Science • University Medical Center New Orleans — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Alabama at Birmingham — UL1TR003096: Center for Clinical and Translational Science • University of Arkansas for Medical Sciences — UL1TR003107: UAMS Translational Research Institute • University of Cincinnati — UL1TR001425: Center for Clinical and Translational Science and Training • University of Colorado Denver, Anschutz Medical Campus — UL1TR002535: Colorado Clinical and Translational Sciences Institute • University of Illinois at Chicago — UL1TR002003: UIC Center for Clinical and Translational Science • University of Kansas Medical Center — UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • University of Kentucky — UL1TR001998: UK Center for Clinical and Translational Science • University of Massachusetts Medical School Worcester — UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University of Minnesota — UL1TR002494: Clinical and Translational Science Institute • University of Mississippi Medical Center — U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center — U54GM115458: Great Plains IDeA-Clinical & Translational Research • University of North Carolina at Chapel Hill — UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Oklahoma Health Sciences Center — U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • University of Rochester — UL1TR002001: UR Clinical & Translational Science Institute • University of Southern California — UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • University of Vermont — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • University of Virginia — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • University of Washington — UL1TR002319: Institute of Translational Health Sciences • University of Wisconsin-Madison — UL1TR002373: UW Institute for Clinical and Translational Research • Vanderbilt University Medical Center — UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • Virginia Commonwealth University — UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • Wake Forest University Health Sciences — UL1TR001420: Wake Forest Clinical and Translational Science Institute • Washington University in St. Louis — UL1TR002345: Institute of Clinical and Translational Sciences • Weill Medical College of Cornell University — UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • West Virginia University — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI)

Submitted: Icahn School of Medicine at Mount Sinai — UL1TR001433: ConduITS Institute for Translational Sciences • The University of Texas Health Science Center at Tyler — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • University of California, Davis — UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, Irvine — UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, Los Angeles — UL1TR001881: UCLA Clinical Translational Science Institute • University of California, San Diego — UL1TR001442: Altman Clinical and Translational Research Institute • University of California, San Francisco — UL1TR001872: UCSF Clinical and Translational Science Institute

Pending: Arkansas Children's Hospital — UL1TR003107: UAMS Translational Research Institute • Baylor College of Medicine — None (Voluntary) • Children's Hospital of Philadelphia — UL1TR001878: Institute for Translational Medicine and Therapeutics • Cincinnati Children's Hospital Medical Center — UL1TR001425: Center for Clinical and Translational Science and Training • Emory University — UL1TR002378: Georgia Clinical and Translational Science Alliance • HonorHealth — None (Voluntary) • Loyola University Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • Medical College of Wisconsin — UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • MedStar Health Research Institute — UL1TR001409: The Georgetown-Howard Universities Center for Clinical and Translational Science (GHUCCTS) • MetroHealth — None (Voluntary) • Montana State University — U54GM115371: American Indian/Alaska Native CTR • NYU Langone Medical Center — UL1TR001445: Langone Health's Clinical and Translational Science Institute • Ochsner Medical Center — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • Regenstrief Institute — UL1TR002529: Indiana Clinical and Translational Science Institute • Sanford Research — None (Voluntary) • Stanford University — UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • The Rockefeller University — UL1TR001866: Center for Clinical and Translational Science • The Scripps Research Institute — UL1TR002550: Scripps Research Translational Institute • University of Florida — UL1TR001427: UF Clinical and Translational Science Institute • University of New Mexico Health Sciences Center — UL1TR001449: University of New Mexico Clinical and Translational Science Center • University of Texas Health Science Center at San Antonio — UL1TR002645: Institute for Integration of Medicine and Science • Yale New Haven Hospital — UL1TR001863: Yale Center for Clinical Investigation