HHS Public Access

Author manuscript

Stud Health Technol Inform. Author manuscript; available in PMC 2023 October 28.

Published in final edited form as:

Stud Health Technol Inform. 2022 June 29; 295: 316–319. doi:10.3233/SHTI220726.

Assessing Disparities in COVID-19 Testing Using National COVID Cohort Collaborative

Jinyan LYU[a], **Wanting CUI**[a], **Joseph FINKELSTEIN**[a,1]

[a]Icahn School of Medicine at Mount Sinai, New York NY, USA

Abstract

With NCATS National COVID Cohort Collaborative (N3C) dataset, we evaluated 14 billion medical records and identified more than 12 million patients tested for COVID-19 across the US. To assess potential disparities in COVID-19 testing, we chose ten US states and then compared each state's population distribution characteristics with distribution of corresponding characteristics from N3C. Minority racial groups were more prevalent in the N3C dataset as compared to census data. The proportion of Hispanics and Latinos in N3C was slightly lower than in the state census. Patients over 65 years old had higher representation in the N3C dataset and patients under 18 were underrepresented. Proportion of females in the N3C was higher compared with the state data. All ten states in N3C showed a higher representation of urban population versus rural population compared to census data.

Keywords

COVID-19; national COVID cohort collaborative; data integrity

1. Introduction

With the outbreak of the COVID-19 pandemic, to facilitate the COVID-19 research significant efforts were undertaken for aggregating representative clinical data on a national level. National COVID Cohort Collaborative (N3C) dataset provided a comprehensive aggregation of clinical data representing patients tested for COVID-19 across US. N3C is an NIH-funded harmonized electronic health record (EHR) dataset with more than 14.2 billion records, including over 6.8 billion lab test results and 4 million COVID test-positive results patients, and 12 million total patients. Over 71 hospitals or clinical sites in the United States pulled their data into the N3C data repository. N3C represents the nation's largest multicenter cohort of COVID-19 tested cases. From 2020 to 2021, at least 59 papers were published using N3C data including papers used the N3C dataset to predict the clinical severity of COVID-19 [1]. Other papers used survival analysis methods to estimate mortality, hospitalization, and infection rate [2]. The N3C dataset played an essential role in COVID-19 research. However, the cohort of COVID-19 tested patients in the N3C dataset may have a selection bias and which can influence the findings in the future studies.

[1]Corresponding Author, Jinyan Lyu, Icahn School of Medicine at Mount Sinai, 1770 Madison Ave, New York, NY, USA, 10035; jinyan.lyu@mssm.edu.

There are no reports examining the representativeness of the N3C dataset. The goal of this paper was to use US census statistics from ten selected states to compare representation of different racial and other subgroups in N3C and state population data.

2. Methods

Analyses in this study were based on data from National COVID Cohort Collaborative (N3C) Data Enclave. N3C is a nation-wide dataset that contains demographic and clinical characteristics of patients who have been tested for or diagnosed with COVID-19. The N3C's harmonized de-identified dataset includes demographic information, death information, location information, measurement information, observation information, procedure information, condition occurrence information, and visit occurrence information in OMOP format. Due to the regulatory compliance requirements, all analyses are being performed on N3C's official data enclave website. US state's population data used as reference were obtained from the US Census Bureau [3].

N3C included patients who undertook COVID-19 testing. We selected all patients from the 'person' table who had location information (location ID). In the ' location ' table, the variable 'State' was mapped from a 3-digit zip code. Patients' age was defined as the testing date minus the patient's year of birth. Race, ethnicity, and gender were extracted from the patients' table corresponding to 'concept name.' The race was categorized as White, Black and African American, Asian, Native Hawaiian or Other Pacific Islander, others, unknown, and no matching concept. The ethnicity variable included 'Not Hispanic or Latino' and 'Hispanic or Latino,' and no matching concept. The Gender variable was categorized into iFemale, Male, and Unknown. Age was split into four groups under five years old, from 6 to 18 years old, from 19 to 65 years old, and over 65 years old. We further defined a variable 'urbanization' based on the 3-digit zip code. We mapped three-digit zip code based on 2010 Census Urban and Rural Classification and Urban Area Criteria [4]. It was defined as urban for metropolitan and micropolitan areas, rural if it was a small town or rural area. Disparities based on gender, age, ethnicity, race, and urbanization were analyzed. Missing values were coded as 'unknown'.

We grouped US states into five regions: the Northeast, Southwest, West, Southeast, and Midwest, based on their geographical location [5]. In each region, we chose two states. We chose Massachusetts (MA) and New York (NY) in the Northeast. Texas (TX) and Arizona (AZ) were chosen in the Southwest region. We chose California (CA) and Colorado (CO) in the West. Florida (FL) and Virginia (VA) were chosen in the Southeast region, andWisconsin (WI) and Illinois (IL) were chosen in the Midwest region. We further calculated the sample rate by N3C population divided by the corresponding state's census population. Then we performed descriptive statistical analysis on each state's demographic information. Using the Chi-square test, we identified whether the state census population was different from the N3C population.

The analyses described in this publication were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS U24 TR002306. This research

was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource [7].

3. Results

A total of 3,876,194 COVID-19 patients who tested for COVID-19 were included in this study representing ten US states and five regions. Table 1 shows comparison between N3C population and US state's population. Colorado had the highest sample ratio (11.37%) while in contrast, California, Florida, Arizona, and Texas had relatively lower sample ratios, 0.60%, 0.30%, 0.44%, 0.14%, respectively. Overall, the Midwest and Northeast regions had a sufficiently higher number of patients recorded for the COVID-19 testing in N3C sample population. In comparison, the Southwest region had the lowest representation of patients tested for COVID-19.

In N3C, Whites made up the majority of all state's samples even if Whites population in Illinois, Texas, and New York is below 50%. For the minority groups located in WI (N3C: 11.47%; Census: 6.7%), IL (N3C: 21.51%; Census: 16.35%), TX (N3C: 19.23%; Census: 12.19%), and MA (N3C: 18.54%; Census: 9.00%), a significantly higher proportion of Black or African Americans were included in N3C as compared to the the states' census. In Florida, the proportion of Black or African Americans tested for COVID-19 was lower than in the state census (N3C: 8.98%; Census: 16.90%). \ Asians tested for COVID-19 were underrepresented in N3C. The proportion of Asians in N3C were lower than the census in the states such as CA (N3C: 7.67%; Census: 15.50%), TX (N3C: 0.89%; Census: 5.20%), VA (N3C: 1.86%; Census: 6.90%), NY (N3C: 4.90%; Census: 9.00%), and MA (N3C: 4.96%; Census: 7.20%). IL, NY, and TX had the highest proportion of 'no matching concept' of the patients tested for COVID-19, 22.36%, 33.91%, 34.27%, respectively. The "unknown" race was listed in 19% of N3C patients from California. In AZ (N3C: 11.59%; Census: 31.70%), VA (N3C: 4.39%; Census: 9.80%), and FL (N3C: 9.74%; Census: 26.4%), Hispanic or Latino patients tested for COVID-19 were in represented in lower proportion, about twice times less than states' census proportion. For non-Hispanic or Latino patients in Florida, N3C had 77.13% of patients while the state census had 52.3%. In Texas, ethnicity of over 78% of patients tested for COVID-19 had no matching concept.

Generally, there were more females who tested for COVID-19 (over 50%) than males. Also, for most states, the proportion of females in N3C was higher than the state's census. However, about 50.92% of COVID-19 tested patients in Florida were females which was lower than the state census (51.10%). Among ten states, the proportion of N3C patients aged under 18 was much lower than in the state census. Patients aged over 65 years old in N3C were represented in higher proportion than in state census. Moreover, all selected states showed a higher urbanization rate (over 90% of patients located in urban areas) in N3C compared with the state census. We applied chi-square tests to examine the difference between N3C and states census proportions in race and ethnicity groups which demonstrated significant difference with $p<0.05$.

4. Discussion

In the N3C dataset, there was more diversity in the race compared with the state census. The chi-square test showed the significant difference between N3C data and states' data in representation of race, ethnicity, age, and location Minority racial groups were more prevalent in the N3C dataset as compared to census data. The proportion of Hispanics and Latinos in N3C was slightly lower than in the state census. Patients over 65 years old had higher representation in the N3C dataset and patients under 18 were underrepresented. Proportion of females in the N3C was higher compared with the state data. Over 90% of patients were from urban areas, other than 10% from rural areas. This may potentially result in bias in future analyses.

5. Conclusions

Disparity in representation of racial minorities in N3C may potentially bias descriptive studies and predictive models using this dataset. Future studies should account for misalignment between socio-demographic parameters' distribution in N3C and population-based cohorts.

References

[1]. Bennett TD, Moffitt RA, Hajagos JG, Amor B, Anand A, Bissell MM. National COVID Cohort Collaborative (N3C) Consortium. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative. JAMA Netw Open 2021;4(7):e2116901. [PubMed: 34255046]

[2]. Ge J, Pletcher MJ, Lai JC. N3C Consortium. Outcomes of SARS-CoV-2 Infection in Patients with Chronic Liver Disease and Cirrhosis: a N3C Study. medRxiv [Preprint] 2021 Jun;.21258312.

[3]. Bureau, US Census. U.S. Census Bureau Quickfacts: United States [Internet] Census.gov. 2021. Available at: https://www.census.gov/quickfacts/fact/table/US/PST045221. Accessed 2022 Mar 10.

[4]. Bureau, US Census. 2010 Census Urban and Rural Classification and Urban Area Criteria [Internet] Census.gov, 2021 Oct 8. Available at: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html. Accessed 2022 Mar 10.

[5]. National Geographic Society. United States Regions [Internet]. National Geographic Society, 2012 Nov 9. Available at: https://www.nationalgeographic.org/maps/united-states-regions/. Accessed 2022 Mar 10.

[6]. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb) 2013;23(2):143–9. [PubMed: 23894860]

[7]. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, Payne PR, Pfaff ER, Robinson PN, Saltz JH, Spratt H. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association 2021 Mar;28(3):427–43. [PubMed: 32805036]

Table 1.

Descriptive statistics of N3C population and US state's population

Region	State	N3C population	State Census 2020	Sample Ratio in N3C/census
	CO	656510	5773714	11.37%
West	CA	239067	39538223	0.60%
	IL	697437	12812508	5.44%
Midwest	WI	546938	5893718	9.28%
	AZ	31366	7151502	0.44%
Southwest	TX	40045	29145505	0.14%
	FL	65118	21538187	0.30%
Southeast	VA	407701	8631393	4.72%
	MA	507406	7029917	7.22%
Northeast	NY	684606	20201249	3.39%