Big Data Decision-Making and Racial Disparities: A Case Study Among COVID-19 Inpatient Visits

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

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

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

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

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

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

Abstract*— *The COVID-19 pandemic has had a disproportionate impact on certain racial and ethnic groups, resulting in significant health outcome disparities. The National COVID Cohort Collaborative (N3C) provides a valuable resource for exploring these disparities through big data analytics. This study belongs to a broader work that examines decisions made during data procrssing and their impact on the analyses performed. Central to our analysis is the introduction of the Continuous Inpatient Encounter (CIE) concept—a novel method we propose for aggregating inpatient visits. By utilizing big data analytics, we aim to identify potential disparities in CIE rates among different racial groups. The results of this study are critical for enhancing the equity of data-driven decision-making in healthcare and for addressing the racial disparities observed in COVID-19 outcomes.

Keywords*— *Big Data Analytics, Racial Disparities, Decision-Making, Macrovisits, Inpatient Encounter, Data Processing, National COVID Cohort Collaborative (N3C)

I. INTRODUCTION

The COVID-19 pandemic has highlighted and exacerbated existing health disparities across various populations. Evidence suggests that racial and ethnic minorities have experienced disproportionately higher rates of infection, hospitalization, and mortality due to COVID-19 compared to their White counterparts [1]. Several factors contribute to these disparities, including socioeconomic status, access to healthcare, pre-existing health conditions, and social determinants of health [2]. Research has demonstrated that social determinants of health, such as poverty, housing conditions, and occupational exposure, play a significant role in shaping these disparities [3]. For instance, individuals from racial and ethnic minority groups are more likely to work in essential jobs that increase their risk of exposure to the virus and have less access to healthcare services [4].

The National COVID Cohort Collaborative (N3C) provides a comprehensive dataset that aggregates Electronic Health Records (EHR) from multiple healthcare institutions across the United States. This resource enables researchers to conduct large-scale analyses of COVID-19 outcomes across diverse populations. The N3C data encompasses detailed information about patient demographics, clinical characteristics, and health outcomes, making it an invaluable tool for examining disparities in COVID-19 [5]. Previous studies have indicated that Black, Hispanic, and Indigenous populations are at higher risk for severe outcomes from COVID-19, including hospitalization and death [6], [7]. Furthermore, data has shown that these groups often face barriers to timely and adequate healthcare, which exacerbates the impact of COVID-19. For example, underlying health conditions such as hypertension, diabetes, and obesity, which are more prevalent in minority communities, increase the risk of severe COVID-19 outcomes [8]. Research also suggests that systemic racism and implicit biases in healthcare delivery contribute to these disparities. Minority patients may receive lower-quality care, face longer wait times, and have less access to advanced treatments [9]. The pandemic has further strained healthcare resources, potentially worsening these inequities [10].

[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.

Big data analytics has emerged as a powerful tool in healthcare decision-making, offering unprecedented opportunities to improve patient outcomes, optimize resource allocation, and enhance overall healthcare delivery. In the healthcare sector, big data analytics integrates and analyzes vast amounts of structured and unstructured data from various sources, including electronic health records, medical imaging, wearable devices, and genomic sequencing [11]. This comprehensive approach allows healthcare organizations to gain deeper insights into patient populations, disease patterns, and treatment efficacy.

One of the key applications of big data analytics in healthcare decision-making is in outcome prediction and risk stratification. By analyzing large datasets, healthcare providers can identify patients at high risk of developing certain conditions or experiencing complications, enabling proactive interventions and personalized care plans. For instance, predictive analytics can be used to forecast patient admissions trends, optimize staffing levels, and improve resource allocation in hospitals [12].

Additionally, big data analytics plays a crucial role in enhancing the quality of care and patient outcomes. It enables healthcare organizations to identify best practices, develop evidence-based treatment protocols, and monitor the effectiveness of interventions in real-time [13]. However, the implementation of big data analytics in healthcare also faces challenges, including data privacy concerns, the need for robust data governance frameworks, and the requirement for skilled professionals who can effectively interpret and apply analytical insights in clinical settings [14].

Moreover, the use of big data analytics in decision-making processes carries the risk of reinforcing existing biases and inequities if not carefully managed. For instance, in healthcare, algorithms designed to predict patient outcomes or allocate resources may inadvertently discriminate against certain racial or ethnic groups if the training data does not accurately reflect their specific health needs and risks [15]. As Obermeyer et al. [16] illustrate, these algorithms, when trained on biased datasets, can amplify disparities by mirroring and magnifying historical inequalities in access and treatment.

This paper presents a case study on Continuous Inpatient Encounters (CIE) for COVID-19 patients, examining how Big Data-driven decisions impact racial disparities in hospitalization outcomes. The study critically analyzes the processes of data collection, analysis, and application in the context of CIE and investigates how these processes might vary across different racial groups.

By focusing on CIE, this study aims to contribute to the discourse on healthcare equity and the ethical use of Big Data in making critical medical decisions. Understanding these disparities is crucial for developing targeted interventions and policies that can mitigate the adverse effects of COVID-19 on vulnerable populations, thereby promoting more equitable health care practices.

II. Methods

The National COVID Cohort Collaborative (N3C) (covid.cd2h.org) serves as an essential tool for COVID-19 research, offering an extensive dataset for analysis through its specialized platform, the N3C Data Enclave [17]. The N3C is a partnership aimed at aggregating and harmonizing EHR data across clinical organizations in the United States. This initiative includes the Clinical and Translational Science Awards (CTSA) Program hubs, the National Center for Advancing Translational Science (NCATS), the Center for Data to Health (CD2H), and the community [18]. The N3C faces the challenge of working with heterogeneous EHR encounter data, particularly hospitalizations, which are complex due to their longer temporal span and the variety of services they include [5]. To address this, the N3C has developed a generalizable method for resolving encounter heterogeneity by combining related atomic encounters into composite *Macrovisits* [19].

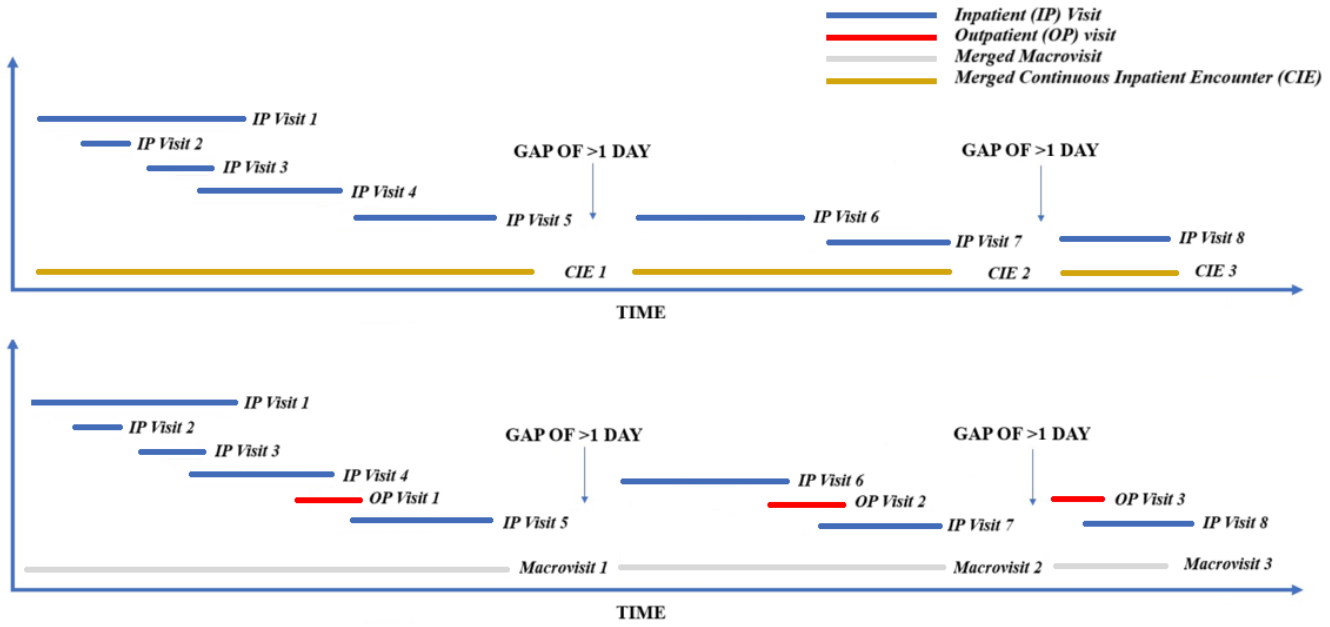

Fig. 1. A demonstration of how N3C defines the concepts of Macrovisit and our definition of Continuous Inpatient Encounters (CIE), incorporating multiple small visits. The delineation criterion is a gap of at least one day between sequential visits. Reproduced from [20] [21].

Macrovisits are created by aggregating discrete EHR encounters into longitudinal clinical experiences that more accurately reflect the patient's clinical journey. This process, termed *Macrovisit* [20] aggregation, involves merging overlapping inpatient and other longitudinal facility visits and then adding any other types of visits that occur during the merged interval [19] (Fig. 1). The goal is to group small visits (microvisits) together to create what the patient and physician might recognize as a single hospitalization. The *Macrovisit* logic is designed to identify 'hospitalizations' more broadly than just inpatient visits, including observation stays and multi-day facility stays following outpatient hospital procedures.

The *Macrovisits* aggregation algorithm is applied to encounter datasets composed of mixed local definitions, resulting in a consistently defined set of longitudinal, multi-encounter experiences for use in further analyses [21]. The algorithm selects visits to form the *Macrovisit* scaffold based on specific criteria, such as having non-null start and end dates, a non-negative length-of-stay, and being one of the specified visit types. It then merges intervals and joins other visits to the scaffold, generating a microvisit_to_macrovisit table [21]. Additional metadata, such as 'covid_dx', 'all_icu', and 'likely_hospitalization', is calculated and added to the table to make *Macrovisits* more research-ready [21].

N3C Macrovisit Aggregation

In this study, we harness the N3C data to examine racial disparities among patients hospitalized for COVID-19, with a particular focus on their CIE. The N3C's *Macrovisit* aggregation algorithm plays a pivotal role in this analysis, as it synthesizes related atomic encounters into *Macrovisits*, thereby enabling a more precise and thorough examination of patient care experiences.

This innovative method for *Macrovisit* aggregation, not only enhances our understanding of the inpatient visits but also represents a significant advancement in the harmonization of EHR data for clinical research, ultimately aiding in the investigation of the disproportionate impact of COVID-19 on different racial and ethnic groups.

Introduce Continuous Inpatient Encounter (CIE)

We introduce the concept of **Continuous Inpatient Encounter (CIE)** - our proposed method for aggregating inpatient visits - as the core of our analysis (Fig. 1). Unlike the N3C *Macrovisit* approach, which includes various types of care encounters (inpatient, outpatient, emergency, and specialty services), our method focuses exclusively on continuous inpatient care episodes.

This distinction is critical for a more focused examination of patient care during hospitalizations related to COVID-19.

It is a common practice that a patient's record indicates hospital discharge immegiatelly followed by admission. Often this is related to billing practices or the need to re-classify patient. Such records are part of the same hospital stay, but indiicated by multiple records. The presented work assumes that if a discharge and consequent admission happen on the same day, thay are groupe dtogether.

Cohort Construction and Analysis

First, and one of the most important steps of any data analysis is cohort construction. To construct the cohorts, a series of decisions were applied to about 22 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. Detailed descriptions of these decisions are available in [22].

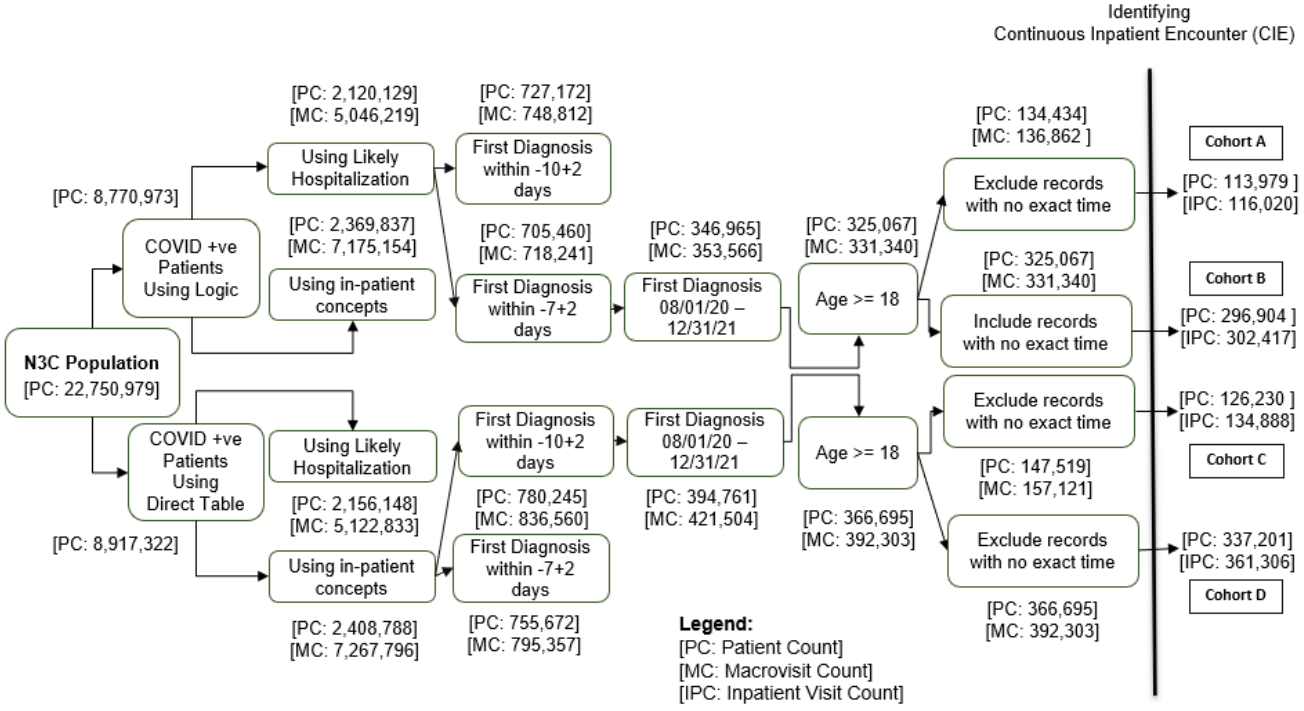

Fig. 2. Partial inclusion/exclusion tree outlining extreme choices for the four decisions. PC indicates patient counts, MC indicates macrovisit counts, IPC indicates inpatient visit count. Repreduced from [22].

Macrovisits vs. CIP

The Fig. 2 also includes counts of *Macrovisit* (CC), which may seem like a good indicator for several aspects related to healthcare utilization and patient care, especially when analyzing data from the N3C. The seemingly arbitrary choices can result in 16 potential datasets of different sizes and properties. The presented work focuses on analyzing the cohort A but can be generalized to all potential datasets.

According to definitions in the N3C data repository [19], a *Macrovisit* includes various types of microvisits, categorized into several groups. These groups include Inpatient services, Outpatient services, Emergency services, and Specialty services.

Inpatient services encompass settings such as inpatient visit, inpatient hospital, inpatient critical care facility, comprehensive inpatient rehabilitation facility, inpatient hospice, and inpatient psychiatric facility. Outpatient services encompass a broad spectrum, including outpatient visits, outpatient hospitals, and various ambulatory centers such as Infusion Therapy, Surgical, Oncology, Dental, MRI, Oncological Radiation, Endoscopy, Mammography, and Rehabilitation. Emergency services include emergency room visit and ambulance visit. Specialty services comprise telehealth, laboratory visit, pharmacy visit, case management visit, home visit, and health examination. Each of these categories reflects different aspects of patient care as outlined in the N3C data framework.

In contrast, our Continuous Inpatient Encounter (CIE) focuses exclusively on inpatient services, encompassing settings such as "Emergency Room and Inpatient Visit" (concept id: 262), "Inpatient Hospital" (concept id: 8717), ("Inpatient Visit") (concept id: 9201), "Intensive Care" (concept id: 32037), and "Inpatient Critical Care Facility" (concept id: 581379) [19].

The Fig. 2 presents the counts of inpatient visits (IPC). By filtering for these specific identifiers, we can isolate records that pertain to various kinds of inpatient visits.

The question addressed in this study is if the definition of hospitalization, *Macrovisit* vs. CIE, affects conclutions of analysis of racial disparities.

III. Results & Discussions

The simplest comparison between the two datasets based on “Macrovisit” and “Continuous Inpatient Encounter,” CIE, is to test differences in cohort demographics. Such differences may indicate possible changes in analyzed disparities. This analysis provides insights into the distribution and frequency of healthcare services utilized by patients, helping us understand the prevalence and accessibility of various healthcare services among different racial groups.

Not surisingly the comparison reveals that "Inpatient Visits" account for the highest percentage among the visit types, with 63.03% of patients in CIE, and very high, but significantly lower 53.44% in *Macrovisits*. Slightly less frequent encounter type "Emergency Room and Inpatient Visit" also shows a notable difference, with 26.45% for CIE compared to *Macrovisits* 22.42% (Fig. 3).

It is also not surprising that “Outpatient Visits” and “Emergency Room Visits” are highly prelevant in *Macrovisits*, at about 20% each, while excluded by definition from CIE.

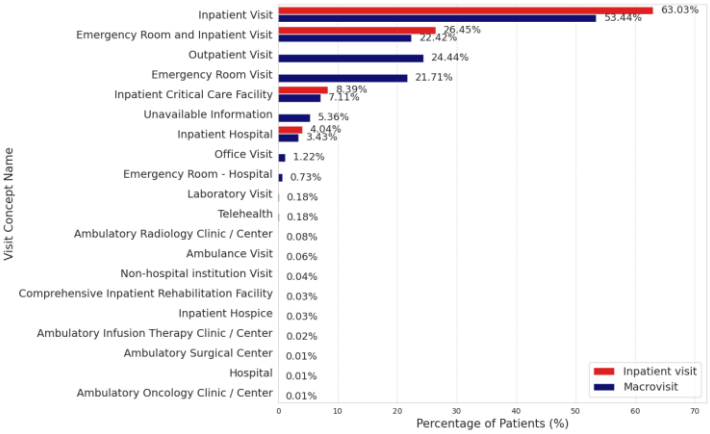

Fig. 3. Comparison of *Macrovisit* and inpatient visit percentages

The Table I presents demographic data related to COVID-19 *Macrovisits* and *CIE* for Cohort A. It includes information on gender, race, and ethnicity distributions among the patients. This analysis helps address the racial disparities observed in COVID-19 outcomes

The Table II serves as a demonstration of the impact of COVID-19 across various demographic segments of Cohort A. It dissects the cohort by gender, racial, and ethnic backgrounds, offering a granular view of the distribution of inpatient visits, including those to hospital emergency rooms and critical care facilities.

This detailed breakdown is instrumental in pinpointing disparities that may exist in healthcare utilization among different racial and ethnic groups affected by COVID-19. By shedding light on these disparities, the table contributes to the overarching goal of the study: to foster more equitable healthcare practices through data-driven decision-making and to address the imbalances in health outcomes that have been exacerbated by the pandemic.

TABLE I. DEMOGRAPHIC ANALYSIS OF COVID-19 MACROVISITS AND INPATIENT VISITS IN COHORT A

		Macrovisit	**Inpatient Visits**
	Total	134,430	113,980
Gender	*Female*	49 % (65,540)	49 % (55,930)
	Male	51 % (68,890)	51 % (58,040)
	Unknown	0.01 % (<20)	0.01 % (<20)
Race	*Multiple races*	0.01 % (130)	0.01 % (<20)
	Nativ Hawian	0.01 % (360)	0.3 % (330)
	Asian	3 % (4,470)	3 % (3,760)
	Black	22 % (29,040)	22 % (25,600)
	Unknown	18 % (23,740)	20 % (22,660)
	White	57 % (76,700)	54 % (61,600)
Ethnicity	*Hispanic or Latino*	18 % (24,310)	19 % (21,480)
	Not Hispanic or Latino	78 % (104,590)	77 % (87,300)
	Unknown	4 % (5,540)	5 % (5,200)

Fig 4, 5, and 6 collectively provide a comprehensive analysis of visit durations, utilizing logarithmic frequency scales to effectively present the data's wide range. Fig. 5 highlights the distribution of *Macrovisit* durations, where shorter stays under 500 days are notably more frequent, while durations beyond 2000 days are rare, illustrating the exceptional nature of prolonged visits.

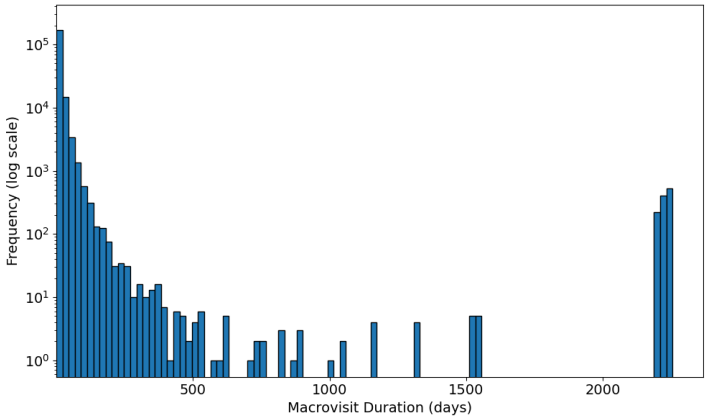

Fig. 5. Histogram of *Macrovisit* durations with logarithmic scale

Similarly, Fig. 6 delves into inpatient visit durations, showing a similar trend where shorter stays are prevalent, and those exceeding 2000 days are uncommon. This pattern underscores the typical brevity of inpatient visits.

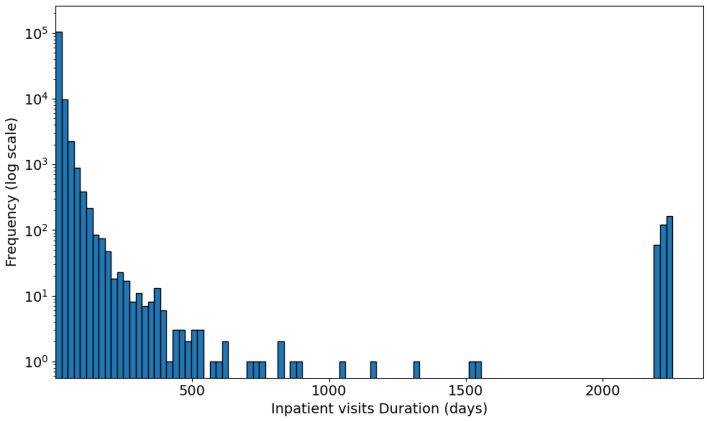

Fig. 6. Histogram of CIE durations with logarithmic scale

Fig. 4 focuses on inpatient visits shorter than 200 days, revealing that durations under 25 days are the most common. As durations increase, their frequency decreases, emphasizing the rarity of extended stays. Collectively, these figures highlight a consistent trend: shorter durations are typical across different types of visits, while extended stays are exceptional occurrences.

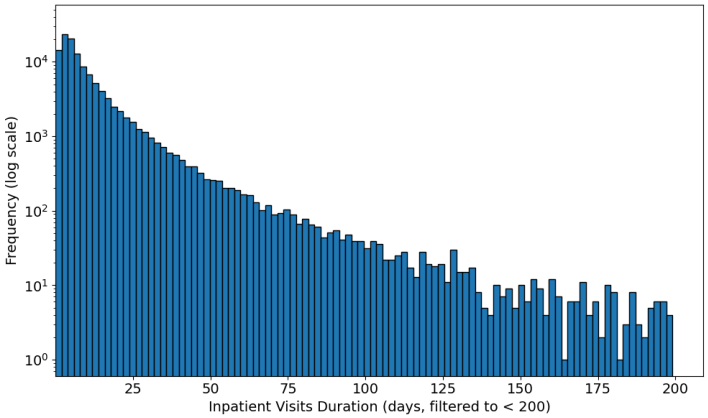

Fig. 4. Histogram of CIE durations (Filtered to < 200 days) with logarithmic scale

The analysis of racial disparities among patients hospitalized for COVID-19, using the N3C data, has provided valuable insights into the differential impact of the pandemic on various demographic groups. The study's focus on first inpatient visits offers a unique perspective on the initial healthcare encounters of COVID-19 patients, which is crucial for understanding the early dynamics of the disease's spread and its effects on different populations.

IV. Conclusions

This study provides a comprehensive analysis of how data-driven decisions can significantly influence the identification and analysis of racial disparities in healthcare, particularly during the COVID-19 pandemic.

By utilizing the N3C dataset, the study underscores the importance of careful data processing and cohort construction to ensure that the analysis accurately reflects the health needs and risks of diverse racial and ethnic groups. This approach is crucial in avoiding the amplification of existing biases and ensuring equitable healthcare outcomes.

TABLE II. DEMOGRAPHIC CHARACTERISTICS FOR COHORT A: INPATIENT VISITS

		Gender			***Ethnicity***			***Race***					
Visit concept name	***Patients count (Total)***	***Male***	***Female***	***Unknown***	***Hispanic or Latino***	***Not Hispanic or Latino***	***Unknown***	***White***	***Black***	***Asian***	***Native Hawaiian***	***Multiple races***	***Unknown***
Inpatient Visit	63% (71,840)	31% (35,300)	32% (36,530)	0% (<20)	12% (13,530)	48% (54,660)	3% (3,650)	31% (35,760)	16% (18,200)	2% (2,500)	0% (210)	0% (<20)	13% (15,150)
Inpatient Hospital	4% (4,610)	2% (2,430)	2% (2,180)	-	0% (80)	4% (4,400)	0% (130)	3% (3,375)	1% (1,030)	0% (60)	-	-	0% (140)
Emergency Room and Inpatient Visit	26% (30,150)	14% (16,070)	12% (14,070)	0% (<20)	7% (7,600)	19% (21,460)	1% (1,080)	5% (6,000)	2% (2,330)	0% (150)	0% (<20)	-	1% (1,080)
Inpatient Critical Care Facility	8% (9,560)	5% (5,480)	4% (4,080)	0% (<20)	1% (881)	7% (8,310)	0% (370)	16% (17,680)	4% (4,260)	1% (1,210)	0% (150)	-	7% (6,840)

A key methodological advancement discussed in the document is the aggregation of Electronic Health Records (EHR) into *Macrovisits*. This process involves combining discrete healthcare encounters into longitudinal clinical experiences, providing a more holistic view of a patient's healthcare journey.

The *Macrovisit* aggregation algorithm allows for a more precise and comprehensive analysis of patient care experiences, which is essential for understanding the broader context of healthcare utilization and outcomes.

Inpatient visits play a significant role in analyzing healthcare utilization and racial disparities during the COVID-19 pandemic. The study highlights that CIE account for the highest percentage of healthcare service utilization, indicating a critical dependency on these services for managing severe COVID-19 cases. By examining the distribution of CIE among different racial and ethnic groups, the research sheds light on the disparities in healthcare access and outcomes, which have been exacerbated by the pandemic.

Moreover, shorter visit durations are prevalent across various contexts, whether analyzing *Macrovisits* or inpatient stays. The application of logarithmic scales effectively emphasizes the high frequency of shorter durations in contrast to the rarity of extended stays. This pattern underscores a common characteristic in healthcare and visit scenarios: brief visits are typical, whereas extended durations are relatively rare. This suggests efficient turnover and resource utilization within these settings.

Therefore, the study emphasizes the transformative potential of big data analytics in addressing racial disparities in healthcare. By leveraging comprehensive datasets like the N3C and employing innovative methods such as *Macrovisit* aggregation, researchers can gain valuable insights into the differential impact of COVID-19 on various demographic groups.

This research not only highlights the importance of equitable data-driven decision-making but also underscores the need for targeted interventions and policies to mitigate the adverse effects of the pandemic on vulnerable populations, ultimately fostering more equitable healthcare practices.

V. LIMITATIONS

The use of large-scale datasets and big data analytics to study complex issues like racial disparities in healthcare presents several limitations and challenges. While the study focuses on a specific cohort (Cohort A) within the N3C dataset created as oart oif a larger project, the findings may not be directly applicable to other populations or settings.

The study's conclusions are based on the specific characteristics and demographics of the cohort analyzed, which may limit the broader applicability of the results. Additionally, the aggregation of EHR from multiple healthcare institutions leads to heterogeneity in data definitions and recording practices. This variability can complicate the analysis and interpretation of results, as different institutions may have different standards for recording patient encounters and outcomes.

Moreover, the study's reliance on data from August 1, 2020, to December 31, 2021, for analysis may not capture the most recent trends and changes in healthcare practices, data reporting, or patient characteristics. The rapidly evolving nature of the COVID-19 pandemic, including the emergence of new variants and changes in treatment protocols, suggests that findings based on historical data may not fully reflect the current situation.

The decision to include patients aged 18 and above is another potential limitation. While this age threshold is commonly used in medical research, it is somewhat arbitrary and could introduce bias. Different age thresholds (e.g., 21+) might yield different results, reflecting varying levels of maturity, independence, and health status among young adults.

Furthermore, the study's findings are limited by the quality and completeness of the data available from the N3C. Missing data, inaccuracies, and variations in how data is recorded across different healthcare systems can affect the reliability of the analysis.

These challenges highlight the need for careful consideration and management of data quality and methodological approaches when using big data analytics to address complex healthcare disparities.

References

[1] M. Webb Hooper, A. M. Nápoles, and E. J. Pérez-Stable, "COVID-19 and Racial/Ethnic Disparities," *JAMA*, vol. 323, no. 24, pp. 2466–2467, Jun. 2020, doi: 10.1001/jama.2020.8598.

[2] K. Mackey *et al.*, "Racial and Ethnic Disparities in COVID-19-Related Infections, Hospitalizations, and Deaths : A Systematic Review," *Ann. Intern. Med.*, vol. 174, no. 3, pp. 362–373, Mar. 2021, doi: 10.7326/M20-6306.

[3] S. Magesh *et al.*, "Disparities in COVID-19 Outcomes by Race, Ethnicity, and Socioeconomic Status," *JAMA Netw. Open*, vol. 4, no. 11, p. e2134147, Nov. 2021, doi: 10.1001/jamanetworkopen.2021.34147.

[4] M. Webb Hooper, A. M. Nápoles, and E. J. Pérez-Stable, "COVID-19 and Racial/Ethnic Disparities," *JAMA*, vol. 323, no. 24, pp. 2466–2467, Jun. 2020, doi: 10.1001/jama.2020.8598.

[5] M. A. Haendel *et al.*, "The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment," *J. Am. Med. Inform. Assoc.*, vol. 28, no. 3, pp. 427–443, Mar. 2021, doi: 10.1093/jamia/ocaa196.

[6] L. Rubin-Miller, C. Alban, S. Artiga, and S. Sullivan, "COVID-19 racial disparities in testing, infection, hospitalization, and death: analysis of epic patient data," *Kais. Fam. Found.*, vol. 2020916, 2020, Accessed: Jul. 13, 2024. [Online]. Available: https://madison365.com/wp-content/uploads/2020/09/KFF-Epic-Analysis_COVID19-and-racial-disparities.pdf

[7] E. G. Price-Haywood, J. Burton, D. Fort, and L. Seoane, "Hospitalization and Mortality among Black Patients and White Patients with Covid-19," *N. Engl. J. Med.*, vol. 382, no. 26, pp. 2534–2543, Jun. 2020, doi: 10.1056/NEJMsa2011686.

[8] C. W. Yancy, "COVID-19 and African Americans," *JAMA*, vol. 323, no. 19, pp. 1891–1892, May 2020, doi: 10.1001/jama.2020.6548.

[9] L. E. Egede and R. J. Walker, "Structural Racism, Social Risk Factors, and Covid-19 — A Dangerous Convergence for Black Americans," *N. Engl. J. Med.*, vol. 383, no. 12, p. e77, Sep. 2020, doi: 10.1056/NEJMp2023616.

[10] K. Bibbins-Domingo, "This Time Must Be Different: Disparities During the COVID-19 Pandemic," *Ann. Intern. Med.*, vol. 173, no. 3, pp. 233–234, Aug. 2020, doi: 10.7326/M20-2247.

[11] K. Batko and A. Ślęzak, "The use of Big Data Analytics in healthcare," *J. Big Data*, vol. 9, no. 1, p. 3, 2022, doi: 10.1186/s40537-021-00553-4.

[12] S. Nijjer, K. Saurabh, and S. Raj, "Predictive big data analytics in healthcare," in *Big Data Analytics and Intelligence: A Perspective for Health Care*, Emerald Publishing Limited, 2020, pp. 75–91. Accessed: Sep. 03, 2024. [Online]. Available:

https://www.emerald.com/insight/content/doi/10.1108/978-1-83909-099-820201009/full/html
[13] B. J. Awrahman, C. Aziz Fatah, and M. Y. Hamaamin, "A Review of the Role and Challenges of Big Data in Healthcare Informatics and Analytics," *Comput. Intell. Neurosci.*, vol. 2022, p. 5317760, Sep. 2022, doi: 10.1155/2022/5317760.
[14] S. Fanelli, L. Pratici, F. P. Salvatore, C. C. Donelli, and A. Zangrandi, "Big data analysis for decision-making processes: challenges and opportunities for the management of health-care organizations," *Manag. Res. Rev.*, vol. 46, no. 3, pp. 369–389, Jan. 2023, doi: 10.1108/MRR-09-2021-0648.
[15] D. A. Vyas, L. G. Eisenstein, and D. S. Jones, "Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms," *N. Engl. J. Med.*, vol. 383, no. 9, pp. 874–882, Aug. 2020, doi: 10.1056/NEJMms2004740.
[16] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.
[17] "N3C - Home." Accessed: Dec. 21, 2023. [Online]. Available: https://covid.cd2h.org/
[18] "Clinical and Translational Science Awards (CTSA) Program | National Center for Advancing Translational Sciences." Accessed: Jul. 02, 2024. [Online]. Available: https://ncats.nih.gov/research/research-activities/ctsa
[19] P. Leese *et al.*, "Clinical encounter heterogeneity and methods for resolving in networked EHR data: a study from N3C and RECOVER programs," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 30, no. 6, pp. 1125–1136, May 2023, doi: 10.1093/jamia/ocad057.
[20] 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.
[21] "Introducing Macrovisits (converted from report on 2024-03-08T15:03:58-05:00)." Accessed: Jul. 02, 2024. [Online]. Available: https://unite.nih.gov/workspace/notepad/view/ri.notepad.main.notepad.06307420-56e3-4012-9a70-f2edc7cde9d6
[22] A. Haghighathoseini *et al.*, "Selection Bias from Data Processing in N3C," in *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, Jun. 2024, pp. 234–241. doi: 10.1109/ICHI61247.2024.00038.

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