

Data quality and social determinants of health: Methods for health equity research

by Lily A. Cook

Dissertation

Presented to the Department of Medical Informatics and Clinical Epidemiology

Oregon Health & Science University

School of Medicine

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

June 17th, 2022

CERTIFICATE OF APPROVAL

This is to certify that the PhD Dissertation of

Lily A. Cook

has been approved.

Mentor/Advisor: Nicole G. Weiskopf, PhD

Chair: Michelle Hribar, PhD

Member: Charisse Madlock-Brown, PhD

Member: Dian Chase, PhD

Member: Miguel Marino, PhD

TABLE OF CONTENTS

ABSTRACT	VI
CHAPTER 1	1
1.1 INTRODUCTION	1
1.2 BACKGROUND	4
1.3 SPECIFIC AIMS.....	13
1.4 REFERENCES	18
CHAPTER 2. THE QUALITY OF SOCIAL DETERMINANTS DATA IN THE ELECTRONIC HEALTH RECORD: A SYSTEMATIC REVIEW	20
2.1 INTRODUCTION	21
2.2 BACKGROUND	21
2.3 METHODS.....	23
2.4 RESULTS	26
2.5 DISCUSSION	34
2.6 CONCLUSIONS	36
2.7 REFERENCES	36
CHAPTER 3. EVALUATING PATIENT RACE AND ETHNICITY DATA FOR SECONDARY USE: VARIABILITY IN EHR DATA ABOUT RACE AND ETHNICITY IN THE NATIONAL COVID COHORT COLLABORATIVE DATA ENCLAVE.....	41
3.1 INTRODUCTION	42
3.2 BACKGROUND	43
3.3 METHODS.....	45
3.4 RESULTS	47
3.5 DISCUSSION	53
3.6 CONCLUSION	57
3.7 REFERENCES	59
CHAPTER 4. THE CARE AND FEEDING OF INTEGRATED DATA: THE IMPACT OF DATA PREPARATION METHODS ON THE RESULTS OF OBSERVATIONAL STUDIES	62
4.1 INTRODUCTION	63
4.2 BACKGROUND	64
4.3 METHODS.....	67
4.4 RESULTS	72
4.5 DISCUSSION	75
4.6 CONCLUSION	78
4.7 REFERENCES	78
APPENDIX 4A. PYTHON CODE FOR THE ALGORITHM DESCRIBED IN CHAPTER 4	81
APPENDIX 4B. FOREST PLOT OF ODDS RATIOS AND CONFIDENCE INTERVALS FOR ALL NODES ...	88
CHAPTER 5. CONCLUSION	89
5.1 CONTRIBUTIONS	90
5.2 LIMITATIONS	91
5.3 CONCLUSION	92

Dedication

For my son, Otis Charlie Maynard,

whose curiosity about the world inspires and motivates me.

Of all the things I have done or accomplished, I am most proud of YOU!

Love ya, kid.

(To answer your question, Otis: No, you do not have to call me Dr. Mommy now.

You can call me Dr. Cook just like everyone else.)

Acknowledgements

I could not have created this without the support of many, many people. Thank you to all my coauthors for your help and many pleasant hours of brainstorming. I'd also like to thank the wonderful staff and faculty of OHSU, especially at DMICE and in the library – I'm looking at you, Diane Doctor, Andrea Ilg, Christine Alpi, Robin Champieux, and Andrew Hamilton! Thank you to my fellow students for your encouragement and comradery, especially: the other attendees of RAD (the incomparable Rose Goueth, Wei-Chun Lin, and Da Jin!); my fellow GRUmlins (too numerous to name here); and those in my qualifying exam cohort (Meena Mishra, Wei-Chun Lin – again! – and Chris Hoekstra). Dr. Karen Eden is the wonderful person who gave me this opportunity and gave me the pep talks and connections I needed in order to succeed. My mentor, Dr. Nicole Weiskopf, constantly amazes me with her intelligence, and everything I have accomplished has been made so much better as a result of her feedback, input, and impressive editing skills. And finally, I am very grateful to the members of my committee for volunteering their time to guide and educate me. Thank you all!

Definitions

Data cleaning: “Data cleaning deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data.”(1) Also known as *data cleansing, wrangling, preprocessing, munging, or scrubbing*.

Data integration: A multistep process of combining datasets, one step of which is data linkage. It is a more comprehensive process than linkage, however, involving additional steps such as data cleaning and removing duplicates. Data integration is sometimes referred to as *data harmonization*.

Data linkage: The step of the *data integration* process where information from different sources is joined. Also known as *record linkage* or simply as a *join*.

Data quality (DQ): “Data that are fit for use by data consumers... a ‘data quality dimension’ is a set of data quality attributes that represent a single aspect or construct of data quality.”(2) Also known as *data integrity, data accuracy, or data validation*.

Geocoding: Roughly, this is the process of reconciling an address to the correct neighborhood information.

Health equity: “The commitment to reducing disparities in health and in its determinants, including social determinants.”(3)

Secondary use: The use of electronic health record data for observational medical research.

Social determinants of health (SDoH): The non-clinical factors that affect health. The generally accepted variables that compose these factors include income level, educational attainment, race/ethnicity, transportation access, housing quality, health insurance status, access to healthy food, built environment, exposure to pollutants, and more. Sometimes also referred to as *social-behavioral determinants of health* or *SBDH*.

Social informatics: “The application of information technologies to capture and apply social data in conjunction with health data to improve clinical care and advance individual and population health. Social informatics uses SDOH-relevant data from informatics resources, such as EHRs, claims data, and mHealth data, to inform research, enhance patient care, and facilitate rapidly growing activities at the intersection of SDOH and medical care.”(4)

Dictionary references

1. Rahm E, Do HH. Data Cleaning: Problems and Current Approaches. Bulletin of the Technical Committee on Data Engineering. 2000;23(4):3 - 13.
2. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996;12:5-33.
3. Khairat S, Haithcoat T, Liu S, Zaman T, Edson B, Gianforcaro R, et al. Advancing health equity and access using telemedicine: a geospatial assessment. J Am Med Inform Assoc. 2019;26(8-9):796-805.
4. Pantell MS, Adler-Milstein J, Wang MD, Prather AA, Adler NE, Gottlieb LM. A call for social informatics. J Am Med Inform Assoc. 2020.

Abstract

Background: Interest among clinicians and researchers in social determinants of health has increased in recent years, driven by a recognition of its role as a major contributor to health outcomes and by interest in improving health equity. However, little is known about the quality of social determinants of health data sourced from real-world data sources such as the electronic health record. Garbage In, Garbage Out, or GIGO, has long been a tenet of information science. This cautionary adage has been used since the 1950s to warn data users that poor quality data leads to correspondingly poor-quality results. Without attention to the quality of social determinants of health data, therefore, it cannot be guaranteed that any analysis performed with it is providing valid or useful insights.

Objectives: The overall objective of this work is to improve the quality of the social determinants of health data used in health equity research. Better data quality will increase the reliability and validity of the resulting research product – a necessary step towards understanding how to improve the health of both individuals and the population. To meet this objective, I pursued the following specific aims:

- 1) Identify and categorize data quality issues specific to social determinants of health variables,
- 2) Evaluate and describe how variations in the ways healthcare systems collect and conceptualize information about social determinants of health can affect the quality of the data,
- 3) Demonstrate that the choices made to improve the quality of social determinants data prior to analysis have a direct and measurable impact on study results.

Results and Discussion: My first study, a systematic literature review, had three major findings: first, the quality of data about race and ethnicity is an important topic with a large body of literature; second, problems with the quality of SDoH data not distributed randomly, with data from Hispanic patients being particularly prone to bias; and third, there are methods that may improve the quality of “real-world” data. These initial findings inspired my subsequent studies. My next study, an analysis of the

data about race and ethnicity in the National Covid Cohort Collaborative Data Enclave, demonstrated the relevance of the first aim by showing how those findings generalize across healthcare systems. Finally, my third study applied the recommendations identified in Aim 1 to the data preparation process and demonstrated how the choices made during this process can have a substantial impact on the results.

Conclusion: This dissertation furthers the development of methods to improve the quality of social determinants of health variables used in health equity research. Using a data-driven approach to the application of data quality improvement methods, I have shown that consideration of data quality and thoughtfully applied quality improvement methods may help prevent bias, thus improving the validity of research conducted with social determinants of health data.

Chapter 1

1.1 Introduction

Interest among clinicians and researchers in social determinants of health has increased in recent years, driven by a recognition of its role as a major contributor to health outcomes, and by interest in improving health equity. However, little is known about the quality of social determinants of health data. It is well-established that poor data leads to poor results – “Garbage in, garbage out,” or GIGO, has been a tenet of information science since the 1950s.(1) Without attention to the quality of social determinants of health data, it cannot be guaranteed that any research performed with it is providing valid or useful insights. **My work intends to close that gap by looking carefully at the data quality issues specific to social determinants of health data, and by exploring how those problems are identified and addressed by data users.**

Weiskopf states: “DQ is task-dependent. Different studies have different data needs, which have different DQ requirements. These different DQ requirements, in turn, call for different DQ assessment methods.”(2) She goes on to state that she intends her work “to allow researchers to match their task-specific data needs to the appropriate DQ requirement and DQ assessment methods.” For my first two aims, **I intend to examine the task-specific data needs of social determinants of health research by working to identify the data quality problems, requirements, and assessment methods specific to that field of study.** Below I will demonstrate the gaps in information which form the rationale for my work.

Gap 1: What are the data quality issues specific to social determinants research?

There are two potential sources for social determinants of health data:

1) *The electronic health record.* Social determinants data can be found either in demographic information collected during the patient onboarding process (e.g., race or educational attainment), or in structured assessments. For these studies, the data quality problems are not that different from

those encountered in other kinds of secondary use research: missing data, incorrect data, or data sequestered in unstructured text fields. These kinds of EHR data quality problems have been identified and explored elsewhere (2-4).

2) Data linkage. Many social determinants of health researchers find it too limiting to use the EHR as an exclusive source of social determinants data because many important factors (e.g. income, exposure to environmental pollutants) are either sparsely collected in the health record or are simply not collected at all. For this reason, SDoH researchers often choose to integrate biomedical data from the EHR with data from external, non-clinical sources, usually data from the community or population level. For example, social determinants of health researchers have connected EHR data with air quality data from the Environmental Protection Agency (EPA), and with neighborhood data from the US Census. **Linking data allows us to incorporate novel variables into clinical research,(5) deepening our understanding of the many factors that affect patient health and aiding in the identification of disparities that create health inequity.** Integrating health data from the EHR with community-level data is a crucial step for identifying all the factors that affect health, be they environmental, economic, or social. However, these integrated datasets compound the secondary use data quality problems mentioned earlier (i.e., missing data, etc), displaying these same issues while also introducing an additional potential source of error: the linkage process itself (6, 7). Thus, while linking data can add a great deal of richness to the dataset by introducing dozens - if not hundreds - of novel variables, it also has the potential to multiply data quality problems, because error and bias can be introduced as a result of incorrect record matching, faulty data from the patient record, faulty data from the external source, or possibly all three.

Gap 2: How do we improve the quality of social determinants of health data?

In recent years, a body of informatics literature has sought to define and create standards for EHR data quality in order to facilitate the secondary use of EHR data for clinical research.(3, 4, 8, 9) Expert informaticians have delineated the terminology and constructed the theoretical frameworks

necessary to identify and categorize the quality issues specific to EHR data. This current body of data quality research has focused on defining the elements of data quality and developing a shared data quality vocabulary. However, a shared definition of data quality is only the beginning of the journey.(10, 11) In order to continue this work, we need to build upon this foundation by considering how to applications for the data quality dimensions we have defined. This work intends to answer questions such as: how can data users implement these insights in order to improve the quality of datasets? Can these data quality frameworks be used to assess real-world datasets? If so, once data quality has been assessed, what comes next?

Most often, data quality assessment is preparatory to data cleaning and, for studies that incorporate linked information, data integration. Decisions about data cleaning and integration are based on any issues identified while evaluating data quality. Data cleaning, also referred to as data munging, data wrangling, or preprocessing, is intended to improve data quality prior to analysis. **As researchers inevitably encounter data quality issues when conducting research – whether this is missing or inaccurate data - they make choices about how to handle these issues.** These choices are most frequently made during the data cleaning process.

Although the data cleaning and integration steps are often very time-consuming – perhaps even the longest (and most tedious) steps in most research projects - there is very little guidance on how they should be conducted. Education and training in research methodology focuses on study design and statistical analytic techniques, not on data preparation.(12) Most often, researchers are taught statistical analysis by practicing on pre-constructed, well-known datasets such as the Iris dataset, the Titanic passenger dataset, etc. These datasets have already been cleaned during some prior, unseen step. Although the body of informatics literature exhaustively focuses on just about every other step of the research cycle,(13) producing manuscripts on study protocols, study design and methodology, participant recruitment, database structures, and of course, data analysis, it is curiously silent on the topics of data cleaning and on data integration. Perhaps as a result of the inadequate

literature on the topic, **there are no standards for resolving data quality problems - and given the fact that the needs of individual users tend to vary greatly, perhaps there shouldn't be.** I wish to emphasize that further research into data cleaning, data integration, and data quality improvement does not imply that a standard is necessary or desirable. **Instead, my purpose is to further the dialogue about the importance of these steps to the research process by documenting how they are conducted in a particular field.**

1.2 Background

My research intersects the following fields:

(a) *Clinical Research Informatics*, specifically the subfield focused on the quality of secondary use data.

(b) *Social informatics*.⁽¹⁴⁾ I am particularly interested in social determinants of health data that is utilized for research on health equity.

I will begin by presenting an overview of the theoretical concepts underpinning the field of data quality. Next, I will tie these concepts back to social determinants of health by briefly reviewing some key data quality issues specific to social determinants of health research.

Key Concepts in Data Quality

Approaches to Data Quality. Beyond Accuracy,⁽¹⁵⁾ a classic of data quality literature by Wang and Strong, was an early introduction to the concepts of data quality. *Beyond Accuracy* distinguishes itself from other data quality literature not only by its early entry into the field, but also in that it emphasizes the importance of the viewpoint from which data quality is approached. Wang and Strong argue that data quality can be approached three possible ways: *intuitively*, *theoretically*, or *empirically*. Some of the current body of work on EHR data quality could be characterized by what Wang and Strong might label a *theoretical* approach, intended to “provide a comprehensive set of data quality attributes that are intrinsic.” However, Wang and Strong point out that both the intuitive and the

theoretical approach suffer from a “focus on the product in terms of development characteristics instead of use characteristics.”

In contrast, data quality pioneers Wang and Strong recommend an *empirical* approach. The empirical approach is informed by direct experience, and as such could be thought of as fundamentally case-driven.(16-18) It is a way of conceptualizing data quality from the point at which it is *used*, rather than the point at which it is *gathered*. The empirical approach to data quality relies on the data user to determine the characteristics necessary “to assess whether data are fit for use in their tasks.”(15)

My work differs from existing research in that I intend to adopt an empirical approach to data quality, in the sense that I will view data quality through the lens of a specific data user: researchers using patient health data for social determinants of health research. Beyond Accuracy presents the empirical approach to data quality as focused on “fitness for use;” my work intends to evaluate how datasets are made or determined to be ‘fit for use’ in social determinants research. I feel it is important to base the work of clinical research informatics firmly within the practice of research itself, and I intend my work to contribute to the growing body of informatics data quality literature by doing just that. Likewise, my work intends to emphasize that it is important to connect the theories about data quality to the practicalities of cleaning and processing data. My work is informed by the understanding that clinical research informatics is necessarily connected to the conduct of research.

To that end, my second aim will focus specifically on how researchers clean and prepare datasets prior to analysis. I will rely on existing EHR data quality frameworks to contextualize data quality in social determinants of health research, and to characterize the choices social determinants researchers make about their datasets during the cleaning and preparation process. **Data cleaning and preparation is, after all, the ‘action’ step for data quality, where decisions about the data are made by researchers with the intention of improving the quality of the data preparatory to analysis.**

A student and co-author of Wang’s, Yang Lee studied these data quality activities in her work,

‘Context-Reflective Data Quality Problem Solving.’(19) Lee focused on how data quality problems are identified and resolved by what she terms ‘practitioners’ (in the context of social determinants of health research, *researchers* could be considered practitioners). Lee’s findings echo Wang and Strong’s emphasis on user and context: “professionals solve problems by crafting rules to integrate business process and data process.” Placing Lee’s findings into the context of health research, the same conclusion might read “*researchers* solve data quality problems by crafting rules to integrate *research* process and data process.”

Rahm and Do point out that data quality problems can be addressed with data cleaning, and that data cleaning is particularly important with integrating multiple data sources.(20) They conceptualize data quality in terms of the source of the error, and describe the data cleaning process as a series of phased steps:

- 1) *Data analysis*, the step where errors are identified by analyzing the data. Errors can be identified with data profiling (examining data type, length, frequency, variance, etc) and summary statistics.
- 2) *Definition of transformation workflow and mapping rules*, a planning step where the structure of the data is conceptualized and a draft plan is created for transforming and integrating the data. At this step, the logic rules are defined to clean and map the data. This usually requires some knowledge of programming, although it can also be conducted with commercial software.
- 3) *Verification*, the testing step where the workflow of the previous step is iterated and refined.
- 4) *Transformation*, where the final version of the transformation and matching step is completed (the “T” of ETL).
- 5) *Backflow of cleaned data*, where the cleaned data is swapped out for the original dirty dataset.

Rahm and Do recommend documenting the data cleaning process and recording all changes so that transformed data can be traced back to the original source if necessary.

Data Quality Research. In Madnick et al's 'Overview and Framework for Data and Information Quality Research,'(18) the authors divide data quality research into four main topic areas:

1) The organizational impact of data quality. This research area examines the implications that data quality has on organizational performance and policy.

2) Database-related technical solutions for data quality. Research that focuses on database technologies and database design, such as enhancing quality within a data warehouse, enterprise architecture, etc.

3) Data quality in computer science and IT. The authors state that "research in this area develops technologies and methods... to manage, ensure, and enhance data quality."(18) The concerns of this area of research are similar to those in topic number 2, but are specific to a particular setting rather than generalizable to database systems. This topic includes the research areas most relevant to my project, such as assessing data quality, identifying data quality problems, and the development of solutions to manage data quality.

4) Data quality in curation. This field originated in library science and concerns the preservation of data.

The authors also suggest fourteen methodologies appropriate for the conduct of data quality research. I will not list all fourteen here, but I will mention the methods suggested by Madnick et al which I plan to utilize for this research project: case study research, and empirical approaches. These approaches are described in more detail in Chapter 3, Aims and Approach.

Data Quality Terminology in Informatics. Although data quality was initially conceptualized for broad application across information systems regardless of context, there is a growing body of literature devoted to the topic of EHR data quality. **As the field of data science has shifted from**

collecting data toward processing the massive amounts of readily available data already collected, informatics researchers have been working on the problem of secondary use – how to make patient data intended as clinical documentation usable as a research source. In order to transform clinical documentation into a research dataset, it is first necessary to agree on how to define data quality. How do we know if health record data is ready to be used for research purposes? What are the qualities of, well, quality itself?

In her construction of EHR data quality dimensions, Weiskopf's definition echoed the user- and context-specific approach of data quality pioneers Wang and Strong. She, too, emphasizes the importance of 'fitness for use' – the idea that data quality is defined by "the needs of a given user pursuing specific goals."⁽⁸⁾ She identified five dimensions of data quality: completeness, correctness, concordance, plausibility, and currency. Of these, she recognized three as essential, irreducible elements:

- 1) *Correctness*, which is defined by whether or not the element present in the EHR is true.
- 2) *Currency*, which is defined as whether or not an element in the EHR is relevant to the patient at a fixed point in time.
- 3) *Completeness*, which is defined as whether or not the truth about a patient is present in the EHR.

Weiskopf has also stressed the importance of context when defining data completeness: "data completeness does not have a simple, objective definition," she states, "but is instead task-dependent."⁽²¹⁾

The Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data, published in 2016, collected the above-mentioned work of Weiskopf,⁽⁸⁾ Wang,⁽¹⁵⁾ and many others into a single, coordinated conceptual framework.⁽³⁾ The Harmonized Model is not merely the combined the work of many experts, however – it represents nearly four years of workshops, consultation, webinars, and iterative refinement. The result is a

framework that provides three major categories of data quality, within which reside an additional six subcategories:

1) Conformance: Conformance speaks to whether the dataset's reported values meet structural standards and formats. Conformance is further broken into three subcategories:

1.a) Value Conformance: Typically outlined in a data dictionary, value conformance may include expectations regarding the data type of each value (example: "Entries in the 'ZIP CODE' column should be integers"), value length (example: "Entries in the 'ZIP CODE' column should be five digits long"), or value range (example: "Entries in the 'ZIP CODE' column should all be from Oregon and southwest Washington, and should therefore fall between 97000 and 99500").

1.b) Relational Conformance: Relational conformance evaluates whether the data elements comply with the expectations inherent to the structure of the database. Example: "Missing values in the 'ZIP CODE' column will be represented with the number zero (0)," or "'ZIP CODE' can be used as a foreign key."

1.c) Computational Conformance: Computational conformance looks at the accuracy of the instruments used to calculate values. Example: "The patient age mentioned in the unstructured clinical documentation should equal to the date of the document minus the patient's date of birth").

2) Completeness: Completeness looks at whether the data is actually present. Example: "Does this patient's record include their zip code?"

3) Plausibility: Plausibility asks if the data values are believable, and conceptually it can be broken into three additional subcategories:

3.a) Uniqueness plausibility: Is a single object represented a single way? Example: "Patients can only have one valid residential address at a given time, and therefore should only have at

most one zip code listed for each episode of care.”

3.b) Atemporal plausibility: Does the data make sense and seem trustworthy? Example: “Summary statistics say that two million of our patients live in the 97002 zip code, although the US Census says that 97002 only has 253 residents.”

3.c) Temporal plausibility: Do values change over time as expected? Example: “It seems unlikely that a patient’s home zip code would change four times over the space of a single week.”

The article presenting the Harmonized Framework concludes with a statement that “future research on the DQ terminology should aim to verify its generalizability and utility.” The secondary purpose of my project is to work towards this goal – to verify the generalizability and utility of the Harmonized Data Quality Framework by applying it to the field of social determinants of health research.

From Terminology to Methodology. Weiskopf is known most widely for her work on data quality terminology, but she has also discussed how researchers identify these data quality issues in the first place. This aspect of her work tends to be overlooked in the body of informatics literature, which thus far has tended to focus on creating agreement around definitions. In addition to the five data quality dimensions mentioned elsewhere, Weiskopf and Weng’s 2013 review also identified seven data quality assessment methods:(8)

1) The gold standard is the most commonly used method for assessing data quality. As the name implies, it involves comparing the dataset to one drawn from another source (or sources). It is used primarily to evaluate completeness and correctness.

2) The data element agreement method is similar to the gold standard in that it also involves a comparison, but in this case the comparison data comes entirely from within the EHR.

3) Element presence simply determines whether or not an element is present in the health

record. This method's only application is for assessing data completeness, although it should be noted that it is not the ONLY method for examining completeness.

4) Data source agreement compares EHR data to data from another, non-EHR, source.

5) Distribution comparison is a statistical method that compares the distribution of EHR values to expected distributions.

6) Validity check, similar to plausibility, looks at whether the values in the health record make sense.

7) Log review examines the data entry practices. This method is primarily used to assess *currency* but can also be used for evaluating *correctness*.

While the body of informatics data quality literature provides practical tools to *assess* data quality, there is very little work currently available on how to *address* these issues once these assessment tools have identified them (see Table 1 below). My research intends to fill this critical need – my project is a response to a dearth of studies on the data cleaning process in research.(11, 20)

Table 1: Pubmed search results for clinical informatics literature relevant to data quality (left column) and data cleaning (right). Search performed in PubMed on 7/1/2020. All numbers indicate results in English.

Data quality search results	Data cleaning search results
(data quality[MeSH Terms]) AND (clinical informatics[MeSH Terms]): 637 results	Not applicable: “data cleaning” is not yet available as a MeSH Term.
("data quality") AND (clinical informatics[MeSH Terms]): 1,742 results	("data cleaning") AND (clinical informatics[MeSH Terms]): 73 results
("data quality"[Other Term]) AND (clinical informatics[MeSH Terms]): 127 results	("data cleaning"[Other Term]) AND (clinical informatics[MeSH Terms]): 9 results
“assessing data quality”: 70	“addressing data quality”: 7

Another significant data quality problem encountered in social determinants research is that

much of the individual-level social determinants data in the EHR is inaccurate or incomplete. Although it is difficult to determine exactly which social determinants variables are the most important, when examining the suspected sources of health inequity the working theory is that they can be distilled down to the effects of poverty and racial segregation, and therefore that the most integral variables are income and race/ethnicity.(22) A significant unresolved problem with social determinants of health data is that these particular variables often are not stored in the individual patient record. Household income is very rarely collected at the point of care – one recent study found that this information was present in only 0.07% of records, and that using NLP to comb through the free-text portion of the patient record could only improve this to 1-3% of records.(23)

In addition, race and ethnicity data is often incomplete, with the exception of federally-qualified health centers which are required by the Affordable Care Act to collect it. OCTRI told me that approximately 2/3 of the 3.5 million records in OHSU's data warehouse "have something other than 'unknown' or 'declined.'" In other words, at least 1/3 of the records in OHSU's data warehouse are missing racial/ethnic data. And this data is *not* missing at random: Sholle found that "underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation."(24) This finding confirms other research which has found that filtering EHRs for patients with complete data carries with it a number of biases.(25, 26) This is significant because it is a common for researchers to discard observations with any missing data elements.

Prior Work in This Area. Although most researchers adopt an ad hoc approach to addressing data quality issues, that is not to say that all of them do. While there are no guides to resolving data quality issues specific to social determinants of health, some have tried to address those specific to clinical data used for other purposes.(27-34) Data quality case studies have been conducted on patient data collected from emergency rooms,(30) registries, (31) and dental records.(27) Several of these case studies have also been developed into recommendations and published as methods articles. These works serve as a precedent for my own research, particularly Aims 2 and 3 (see Table 2).

Table 2: Articles influencing the approach for Specific Aims 2 & 3

Context of Data	Title of data quality article	Notes
Dentistry	“Electronic primary dental care records in research: A case study of validation and quality assurance strategies”(27)	Case study. Model article for Aim 2.
Hospital (single-center)	“Initializing a hospital-wide data quality program: the AP-HP experience”(28)	Case study. Model article for Aim 2.
Primary Care, Family Medicine (national database)	“Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data”(29)	Case study. Model article for Aim 2.
Emergency Department (single center)	“Using a Data Quality Framework to Clean Data Extracted from the Electronic Health Record: A Case Study”(30)	Case study. Model article for Aim 2.
Registry of Premature Births	“Quality of EHR data extractions for studies of preterm birth in a tertiary care center: guidelines for obtaining reliable data”(31)	Case study and recommendations. Model article for Aims 2 and 3.
Outpatient care	“Enhancing medical data quality through data curation: a case study in primary Sjögren’s syndrome”(32)	Case study and recommendations. Model article for Aims 2 and 3.
General practice (single center)	“An eight-step method for assessing diagnostic data quality in practice: chronic obstructive pulmonary disease as an exemplar”(33)	Case study and recommendations. Model article for Aims 2 and 3.
(1) prescription data; and (2) distance glucose monitoring	“TAQIH, a tool for tabular data quality assessment and improvement in the context of health data”(34)	Tool/method. Model article for Aim 3.

1.3 Specific Aims

Interest among clinicians and researchers in social determinants of health has increased in recent years, driven by a recognition of its role as a major contributor to health outcomes and by interest in improving health equity. However, little is known about the quality of social determinants of health data. Because it is well-established that poor quality data leads to poor results, GIGO – Garbage In, Garbage Out – has long been a tenet of information science. Without attention to the quality of social determinants of health data, it cannot be guaranteed that any research performed with it is providing valid or useful insights. My work intends to fill this *critical need* by looking carefully at the data quality issues specific to social determinants research, and by exploring how those problems are

identified and addressed. In order to accomplish this I have focused on data quality improvement practices for real-world data, particularly on the data preparation process and the data pipeline.

My long-term goal is to improve the quality of the data used in health equity research. Better data quality will increase the reliability and validity of the resulting research product – a necessary step towards understanding how to improve the health of both individuals and the population. My overall objective is to find ways for researchers to minimize error and increase the reproducibility of their work. This project's rationale is that there are few protocols or agreed upon methods available to guide researchers, so most rely upon ad hoc approaches to improve data quality. This grab bag of methodologies leads to uneven research results that are difficult to compare across studies, and it also discourages reproducibility. In addition, the time and effort devoted to improving data quality tends to be steep, thus serving as a rate limiting factor for many social determinants research projects.

To meet these goals and objectives, I have pursued the following specific aims:

1. ***Identify and categorize data quality issues specific to social determinants of health variables.*** I conducted a systematic literature review to collect and synthesize evidence regarding data quality problems encountered when working with variables related to social determinants of health. The objective of this review was to better understand the data quality problems encountered when working with social determinants of health variables sourced from the EHR. I drew upon an existing EHR data quality framework, the Harmonized model by Kahn et al,(3) to identify and describe the data quality issues identified in the literature.
2. ***Evaluate and describe how variations in the ways healthcare systems collect and conceptualize information about social determinants of health can affect the quality of the data in aggregated clinical databases.*** This second aim demonstrated the relevance of the first aim by showing how the findings of the literature review generalize across healthcare systems. Aim 2 utilized the unique opportunity provided by the National Covid Cohort Collaborative in order to better understand how different healthcare systems across the nation

collect and conceptualize information about their patients' race and ethnicity. This second aim also examined how these data were integrated when they were aggregated into a large clinical database.

- 3. *Demonstrate that choices made to improve the quality of social determinants data prior to analysis have a direct and measurable impact on study results.*** Drawing upon the data quality improvement recommendations identified in the first aim, I showed that applying these recommendations to the data preparation process effects the analysis results.

The expected outcome of these aims is a body of work that will further the development of methods to enhance the data quality of clinical social determinants of health variables. The positive impact of this work will be improved validity and reliability of social determinants research, less error, greater reproducibility, and faster research turnaround.

These three aims work together to further the development of methods that will guide users of clinical social determinants of health data during the process of improving data quality for secondary use research. These methods will be the product of three studies. The first study was literature-based, and characterized the data quality needs specific to SDoH research (Specific Aim 1).

The second study demonstrates the relevance of the first aim by showing how Aim 1's findings generalize across healthcare systems. This second study draws upon the unique opportunity provided by the National Covid Cohort Collaborative to better understand how different healthcare systems across the nation collect and conceptualize information about their patients' race and ethnicity. I also examined how these data were integrated when they were aggregated into a large clinical database in order to make the data usable for researchers (see Specific Aim 2).

Finally, my last study shows that choices made to improve the quality of social determinants data prior to analysis have a direct and measurable impact on study results. Drawing upon the data quality improvement recommendations identified in the first aim, I showed that applying these recommendations to the data preparation process effects the analysis results.

Aim 1

Goal: To conduct a systematic literature review to collect and synthesize evidence regarding data quality problems encountered when working with variables related to social determinants of health.

Objectives:

- a) From the existing literature, extract information on common data quality issues encountered during the research process. Use an existing EHR DQ framework, the Harmonized Model, to categorize and identify the problems encountered by clinical social determinants of health researchers when working with patient health data.(3)
- b) Identify whether the data quality issues in social determinants of health research are similar to those found in other secondary use research.

Research Questions:

- a) What kinds of data quality issues are typically encountered while conducting social determinants of health research? What types of missingness are most commonly assessed in social determinants of health research?
- b) Does social determinants of health research suffer from data quality problems similar to other secondary use research? Weiskopf and Weng, for example, found that most of the articles in their review assessed completeness (64%) and/or correctness (60%); (8) does the social determinants literature reflect similar proportions? When data completeness is an issue, is data missing completely at random, or is it not missing at random (and therefore more prone to bias)?
- c) What are specific examples of how the issues described in the data quality frameworks play out in social determinants research?

Aim 2

Objective: Evaluate and describe variations in how healthcare systems collect and conceptualize

information about the race and ethnicity of their patients, and how race and ethnicity data from multiple healthcare institutions are integrated in an aggregated database.

Research Questions:

- a) Do differences in how race and ethnicity data is conceptualized and encoded by healthcare institutions affect the quality of the data in aggregated clinical databases?
- b) How do the problems with data about race and ethnicity data identified in Aim 1 play out in a national, aggregated clinical database?
- c) What differences are there in the ways that healthcare institutions record information about race and ethnicity, and what effect do these differences have on what insights we can reliably gather from the data?

Aim 3

Goal: To further the development of methods for improving the quality of clinical social determinants datasets, thus reducing error and increasing study validity and reliability.

Objective: To demonstrate quantitatively that applying the recommendations identified in Aim 1 to the data preparation process have a substantial impact on study results.

Research Questions:

- a) What “best practices” will help researchers limit error when they are integrating health data with external datasets by linking on address data?
- b) What data preparation practices are most important for analyses that rely on data integration?
- c) How should data users maximize quality when transforming real-world social determinants of health data?
- d) Do the methods used to address data quality problems have an impact on study results?

Hypothesis: Applying the recommendations identified in Aim 1 to choices made during the data

preparation process will have a substantial impact on study results.

1.4 References

1. Work with new electronic 'brains' opens field for Army math experts. The Hammond Times. 1957, Nov 10.
2. Weiskopf NG. Supporting Reuse of Electronic Health Record Data for Research Through Actionable Guidelines for Data Quality Assessment [Dissertation Proposal]: Columbia University; 2012.
3. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC). 2016;4(1):1244.
4. Zozus MN, Kahn MG, Weiskopf NG. Data Quality in Clinical Research. Clinical Research Informatics. Health Informatics.2019. p. 213-48.
5. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA. 2014;311(24):2479-80.
6. Aalsma MC, Schwartz K, Haight KA, Jarjoura GR, Dir AL. Applying an Electronic Health Records Data Quality Framework Across Service Sectors: A Case Study of Juvenile Justice System Data. EGEMS (Wash DC). 2019;7(1):26.
7. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. BMC Health Serv Res. 2010;10:346.
8. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-51.
9. Marchionini G, Lee CA, Bowden H, Lesk M. Curating for quality: ensuring data quality to enable new science. National Science Foundation; 2012.
10. Weng C. Clinical data quality: a data life cycle perspective. Biostat Epidemiol. 2020;4(1):6-14.
11. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2005;2(10):e267.
12. Zhu Y, Hernandez LM, Mueller P, Dong Y, Forman MR. Data Acquisition and Preprocessing in Studies on Humans: What Is Not Taught in Statistics Classes? Am Stat. 2013;67(4):235-41.
13. Johnson SB. Clinical Research Informatics: Supporting the Research Study Lifecycle. Yearb Med Inform. 2017;26(1):193-200.
14. Pantell MS, Adler-Milstein J, Wang MD, Prather AA, Adler NE, Gottlieb LM. A call for social informatics. J Am Med Inform Assoc. 2020.
15. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996;12:5-33.
16. Yin RK. Case Study Research: Design and Methods. 3rd ed: Sage Publications; 2002. 179 p.
17. Ellinger AD, McWhorter R. Qualitative Case Study Research as Empirical Inquiry. International Journal of Adult Vocational Education and Technology. 2016;7(3):1-13.
18. Madnick SE, Wang RY, Lee YW, Zhu H. Overview and Framework for Data and Information Quality Research. ACM Journal of Data and Information Quality. 2009;1(1):1-22.
19. Lee YW. Crafting Rules: Context-Reflective Data Quality Problem Solving. Journal of Management Information Systems. 2004;20(3):93 - 119.
20. Rahm E, Do HH. Data Cleaning: Problems and Current Approaches. Bulletin of the Technical Committee on Data Engineering. 2000;23(4):3 - 13.
21. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform. 2013;46(5):830-6.
22. Chetty R, Stepner M, Abraham S, Lin S, Scuderi B, Turner N, et al. The Association Between Income and Life Expectancy in the United States, 2001-2014. JAMA. 2016;315(16):1750-66.
23. Hatef E, Rouhizadeh M, Tia I, Lasser E, Hill-Briggs F, Marsteller J, et al. Assessing the Availability of Data on Social and Behavioral Determinants in Structured and Unstructured Electronic Health Records: A Retrospective Analysis of a Multilevel Health Care System. JMIR Med Inform. 2019;7(3):e13802.

24. Sholle ET, Pinheiro LC, Adekkanattu P, Davila MA, Johnson SB, Pathak J, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J Am Med Inform Assoc.* 2019;26(8-9):722-9.
25. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, et al. Biases introduced by filtering electronic health records for patients with "complete data". *J Am Med Inform Assoc.* 2017;24(6):1134-41.
26. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak.* 2014;14(1):51.
27. Wanyonyi KL, Radford DR, Gallagher JE. Electronic primary dental care records in research: A case study of validation and quality assurance strategies. *Int J Med Inform.* 2019;127:88-94.
28. Daniel C, Serre P, Orlova N, Breant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. *Comput Methods Programs Biomed.* 2019;181:104804.
29. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. *Health Inf Manag.* 2019;1833358319887743.
30. Dziadkowiec O, Callahan T, Ozkaynak M, Reeder B, Welton J. Using a Data Quality Framework to Clean Data Extracted from the Electronic Health Record: A Case Study. *EGEMS (Wash DC).* 2016;4(1):1201.
31. Knake LA, Ahuja M, McDonald EL, Ryckman KK, Weathers N, Burstain T, et al. Quality of EHR data extractions for studies of preterm birth in a tertiary care center: guidelines for obtaining reliable data. *BMC Pediatr.* 2016;16:59.
32. Pezoulas VC, Kourou KD, Kalatzis F, Exarchos TP, Venetsanopoulou A, Zampeli E, et al. Enhancing medical data quality through data curation: a case study in primary Sjögren's syndrome. *Comput Biol Med.* 2019;107:270-83.
33. Faulconer ER, Lusignan Sd. An eight-step method for assessing diagnostic data quality in practice- chronic obstructive pulmonary disease as an exemplar. *Inform Prim Care.* 2004;12:243-53.
34. Alvarez Sanchez R, Beristain Iraola A, Epelde Unanue G, Carlin P. TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. *Comput Methods Programs Biomed.* 2019;181:104824.

Chapter 2. The quality of social determinants data in the electronic health record: a systematic review

Note: The following was published in the Journal of the American Medical Informatics Association. Citation: Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. JAMIA. 2022 Jan 1;29(1):187–96.

Abstract

Objective: The aim of this study was to collect and synthesize evidence regarding data quality problems encountered when working with variables related to social determinants of health (SDoH).

Materials and Methods: We conducted a systematic review of the literature on social determinants research and data quality and then iteratively identified themes in the literature using a content analysis process.

Results: The most commonly represented quality issue associated with SDoH data is plausibility (n=31, 41%). Factors related to race and ethnicity have the largest body of literature (n=40, 53%). The first theme, noted in 62% (n=47) of articles, is that bias or validity issues often result from data quality problems. The most frequently identified validity issue is misclassification bias (n=23, 30%). The second theme is that many of the articles suggest methods for mitigating the issues resulting from poor social determinants data quality. We grouped these into 5 suggestions: avoid complete case analysis, impute data, rely on multiple sources, use validated software tools, and select addresses thoughtfully.

Discussion: The type of data quality problem varies depending on the variable, and each problem is associated with particular forms of analytical error. Problems encountered with the quality of SDoH data are rarely distributed randomly. Data from Hispanic patients are more prone to issues with plausibility and misclassification than data from other racial/ethnic groups.

Conclusion: Consideration of data quality and evidence-based quality improvement methods may help prevent bias and improve the validity of research conducted with SDoH data.

2.1 Introduction

Interest among clinicians and researchers in social determinants of health (SDoH) has increased in recent years, driven both by a recognition of their role as major contributors to health outcomes and by interest in improving health equity. There are substantial and justifiable concerns, however, regarding the quality of SDoH in clinical data. It is a long-established tenet of information science that poor quality data leads to poor quality results.⁽¹⁾ Without attention to the quality of social determinants of health data, researchers cannot guarantee that results provide valid or useful insights.

Our objective was to conduct a review of the literature on SDoH data quality in order to characterize the relevant issues. Specifically, our goal was to collect and synthesize available evidence regarding the kinds of data quality issues typically encountered while conducting social determinants of health research, the biases these issues can create during analysis, and any identified methodological solutions to these issues that can be used by researchers to improve social determinants data quality prior to analysis.

We were unable to find any prior work that both gathers information about how researchers can improve the quality of clinical social determinants of health data, and also summarizes how specific issues of bias and validity are introduced into it. To the best of our knowledge this review is the first to bring together information about a variety of social determinants variables in order to examine the issues inherent to the field of social informatics.

2.2 Background

Although data quality - also known as data integrity, data accuracy, or data validation - was initially conceptualized for broad application across information systems regardless of context, there is a growing body of literature devoted to the topic of EHR data quality. As the field of data science has shifted from collecting data toward processing the massive amounts of data already collected, informatics researchers are addressing the problem of secondary use – how to make clinical documentation usable as a research source.

Published in 2016, the Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data (2) provides three major concepts that describe the quality of EHR data used for research:

1) Conformance speaks to whether the dataset's reported values meet structural standards and formats. Conformance is further broken into three subcategories: value conformance, relational conformance, and computational conformance.

2) Completeness looks at whether or not the data is actually present.

3) Plausibility asks if the data values are believable and accurate. It can be broken into three additional subcategories: uniqueness plausibility, atemporal plausibility, and temporal plausibility.

Social determinants of health are, by their very nature, non-medical. As such, few SDoH data elements are routinely stored in the medical record. Race/ethnicity and health insurance status (which is used as a proxy to represent socioeconomic status or access to care) are commonly collected during patient onboarding and are usually available as demographic or administrative data. However, these two elements alone represent only a small portion of all the non-clinical factors known to influence health, and although there has been a widespread effort to collect more SDoH from patients these data remain sparse. Indeed, some concepts, such as income, are not available in the EHR at all. This leaves researchers interested in social determinants other than race/ethnicity or health insurance status with several options:

1) Diagnostic codes (e.g., ICD Z codes) indicating social determinants of health such as "Problems relating to housing and economic circumstances" have the benefit of being standardized across systems, but in practice are rarely used by clinicians.(3-5)

2) Geocoding patient addresses allows researchers to integrate biomedical data from the EHR with community-level data sources such as the US Census. In 2014, the Institute of Medicine (IOM) suggested using "neighborhood and community composition" as a proxy for individual level

indicators that cannot be directly collected from patients.(6)

3) Structured assessments, such as PRAPARE, collect a wider variety of social determinants of health information from patients than can be found in administrative information. However, they are only collected by a portion of clinics, and because there is no ubiquitous standard there is a wide variety of assessments being administered.

Each of these methods has benefits and drawbacks that must be considered when selecting a research dataset. However, little information is available to assist researchers making these choices, and there are no agreed-upon best practices for working with social determinants of health data.

2.3 Methods

Search Strategy and Screening. We conducted an iterative, deductive, and systematic literature review. Specifics of the workflow are detailed in Figure 1. First, PubMed and Ovid MEDLINE databases were searched for articles focused on social determinants research and data quality, accuracy, validity or the introduction of bias. Although no date range was specified for articles, the initial results were limited by the date range of databases themselves. The search terms were constructed in keeping with the definition of social determinants of health put forward by major organizations such as Gravity Project and the World Health Organization.(7, 8) Social and economic factors such as race/ethnicity, socioeconomic status, education level, environmental health (proximity to healthy food, walkability, exposure to environmental toxins, etc.), and health insurance status were all considered social determinants of health for the purpose of this review. However, behaviors such as smoking and exercise were not. Because data linkage is commonly used to enrich clinical social determinants datasets with community-level information, articles about the quality of geocoded patient address data were included if they discussed research focused on linking clinical data to exterior datasets for clinical research purposes.

The medical subheading (MeSH) database was used to identify appropriate keywords. The initial search was conducted in PubMed on December 14th, 2020 using the following keywords:

((data accuracy [MeSH Terms]) OR (data quality[Other Term]) OR (data accuracy[Other Term]) OR (data reliability[Other Term]) OR (data validity[Other Term]) OR (data consistency[Other Term]) OR (data completeness[Other Term]) OR (data errors[Other Term]) OR (data error[Other Term]) OR (data bias[Other Term]) OR (information bias[Other Term]) OR (missingness[Other Term]))

AND ((social determinants of health [MeSH Terms]) OR (health equity[MeSH Terms]) OR (social class[MeSH Terms]) OR (income[MeSH Terms]) OR (socioeconomic factors[MeSH Terms]) OR (psychosocial deprivation[MeSH Terms]) OR (educational status[MeSH Terms]) OR (healthcare disparities[MeSH Terms]) OR (ethnic groups[MeSH Terms]) OR (stereotyping[MeSH Terms]) OR (Health Status Disparities[MeSH Terms]) OR (geographic information systems[MeSH Terms]) OR (social support[MeSH Terms]))

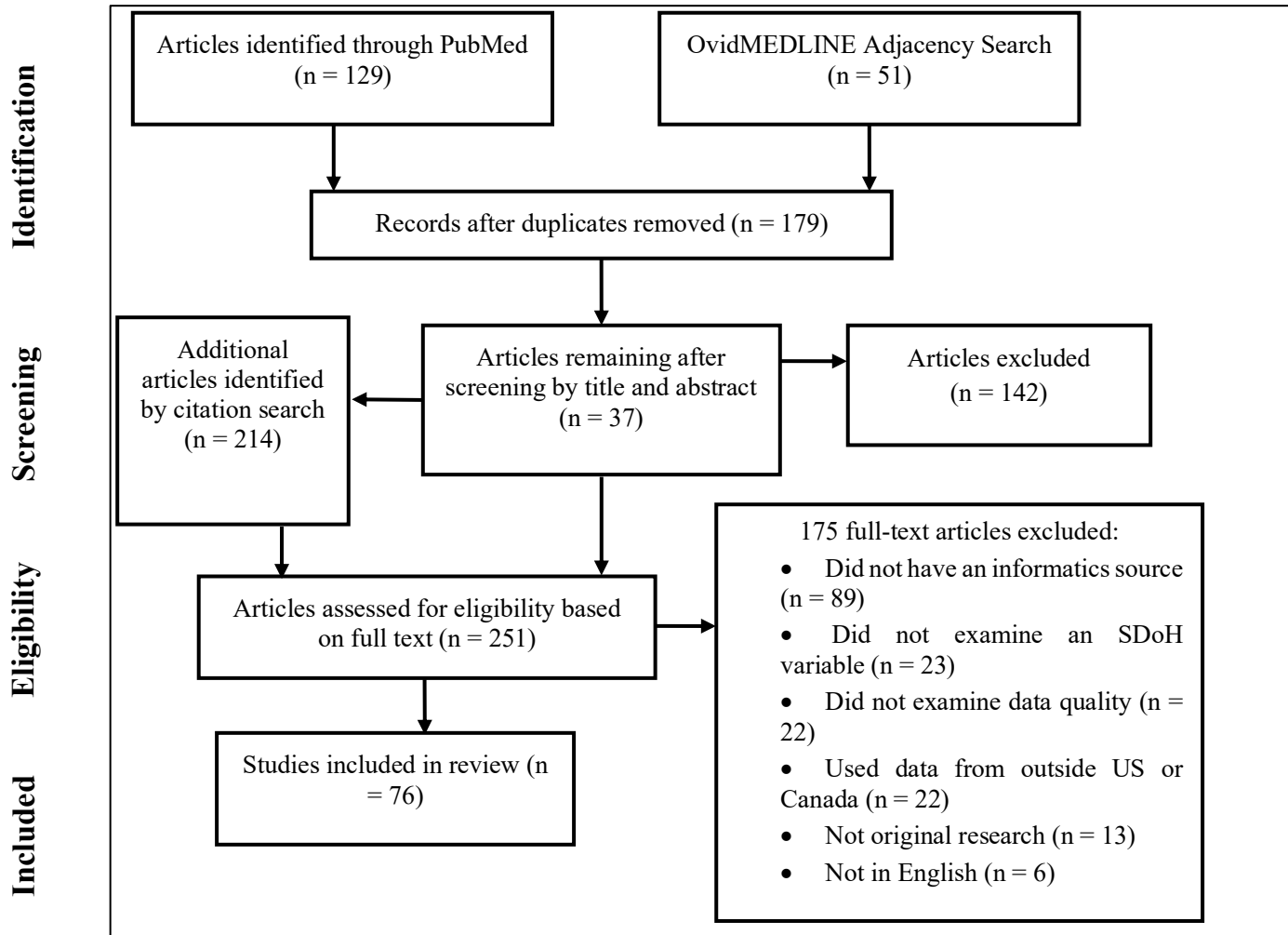


Figure 1: PRISMA flow diagram

Using the following keywords, an adjacency search of OvidMEDLINE was conducted to find articles not indexed in PubMed:

(data adj3 (accura* or consisten* or valid* or collect*) adj10 ((social* adj3 determin* adj3 health*)

or ((socioeconom or economi*) adj3 (dispar* or inequal* or unequal*))) .mp.*

Eligibility. The results of the two database searches were compared in order to identify and remove duplicate articles, and then the remaining articles were screened based on title and abstract. After screening, the first author then manually reviewed the remaining articles to determine whether they met the eligibility criteria described below and summarized in Table 1.

Table 1. Eligibility criteria for articles

	✓ Included	✗ Excluded:
Topic/Focus	Original, peer-reviewed research focused on the quality of social determinants of health data.	Reviews; opinion pieces; research that has not been peer reviewed.
Social Determinants of Health Factors	Race/ethnicity, language preference, health insurance status, country of origin, occupation, socioeconomic status, education level, environmental health (proximity to healthy food, walkability, exposure to environmental toxins, etc.), geocoded patient address data (only included if the article primarily focused on linking clinical data to external datasets for research on social determinants)	Behaviors (e.g., smoking and exercise)
Sources of Health Data	Clinical sources within the US and Canada: EHR, medical registries, administrative databases compiled from EHR data, observational studies using clinical data pulled directly from the medical records of participants	Non-clinical sources: population-level data, mHealth sources, genomic datasets, vital records (i.e., birth or death certificate data); Clinical sources outside the US or Canada
Language	Articles written in English.	Articles in languages other than English.

Inclusion criteria. Articles were eligible for the review if they were in English, used data from healthcare systems in the United States and Canada, and were original research. The research described in the articles must use patient health data originating from clinical sources (i.e., data sourced from electronic health records, disease registries, etc). Studies examining information from registries were included because this information is often abstracted directly from medical records.(9) Also included were several large databases compiled from EHR data, such as the Biomedical Translational Research Information System repository,(10) and the Healthcare Cost and Utilization Project's State Inpatient Database.(3, 11-13) Research using datasets from large cohort studies such as the National Birth Defect Prevention Study were included if the study's dataset was drawn directly from the participant's medical records.(14, 15)

Exclusion criteria. Articles were ineligible for this review if they described research that used patient health data from non-clinical sources (e.g., population-level data, patient-generated data such as mHealth sources, or data from genomic datasets not originating from clinical sources). Vital records such as birth and death certificate data were not included because they often differ significantly in structure and content from other medical records. To identify additional works missed by the initial query, a snowball technique was applied to the citations in the eligible articles.

Data extraction and thematic analysis. Using a deductive approach, content analysis was performed on eligible articles. Each manuscript was categorized by (a) the specific social determinant examined, and (b) the primary data quality issue.⁽¹⁶⁾ This information was then abstracted and tabulated.⁽¹⁷⁾ Data abstraction was performed alongside a closer reading of the selected articles, which informed the thematic analysis.⁽¹⁸⁾ An iterative, inductive approach was taken to identify themes in the literature, with a specific focus on themes that could be useful to researchers using health records for social determinants research.⁽¹⁹⁾ Once themes were identified, the articles were reviewed a final time to abstract and tabulate information about the themes.

2.4 Results

A total of 76 articles were included in this review. Throughout the literature, the most commonly represented quality issue associated with social determinants data was plausibility, i.e., accuracy ($n = 31$, 41%). Thirty-eight percent ($n = 29$) of manuscripts focused primarily on the completeness of social determinants data in the medical record – whether or not data was missing. The remaining 21% looked largely at conformance – whether data was compatible ($n = 16$). A tabulated breakdown of all the articles by data quality issue and social determinants category is available in Table 2.

Articles about race, ethnicity, and country of origin were grouped into a single category, and had the largest body of data quality literature (40 of 76 articles; 53%). Sixteen articles (21%) addressed the quality of geocoded patient address data, which is frequently used in clinical social determinants research to link individual, patient-level data to community-level datasets in order to incorporate variables not available in the medical record. Also represented in the literature were occupation (9%), environmental

Table 2. Characteristics of studies included in this review

	Primary Social Determinant of Health:					
Primary data quality issue:	Race, ethnicity, and/or country of origin, (n, %)	Insurance status, (n, %)	Occupation, (n, %)	General community-level, (n, %)	Environmental, (n, %)	Non-specific (n, %)
Completeness (missing data), n = 29	15 (37.5%)	1 (100%)	6 (86%)	2 (12.5%)	1 (20%)	6 (86%)
Conformance (incompatible data), n = 16	0	0	1 (14%)	10 (62.5%)	3 (60%)	0
Plausibility (inaccurate data), n = 31	25 (62.5%)	0	0	4 (25%)	1 (20%)	1 (14%)
TOTAL, n = 76	40	1	7	16	5	7
Typical article title:	"Accuracy of Race, Ethnicity, and Language Preference in an Electronic Health Record"	"Primary Payer at DX: Issues with Collection and Assessment of Data Quality"	"Availability and accuracy of occupation in cancer registry data among Florida firefighters"	"Match Rate and Positional Accuracy of Two Geocoding Methods for Epidemiologic Research"	"Residential mobility in early childhood and the impact on misclassification in pesticide exposures"	"Utilization of Social Determinants of Health ICD-10 Z-Codes Among Hospitalized Patients in the United States, 2016-2017"
	Usual source for these variables within the patient record:					
	administrative or demographic sources	patient address is geocoded to link community-level data		diagnosis codes		

factors (7%), and insurance status (1%). Seven articles (9%) addressed social determinants data generally without focusing on any specific variable. Three of these non-specific articles discussed the use of ICD codes (a.k.a., Z-codes) as a source of social determinants information in the medical record.

Bias. The first theme identified in the thematic analysis was that bias or validity problems were likely to result from data quality problems. A majority of articles (46 of 76; 61%) either found bias when running test analyses on their datasets, or they noted that data quality was differentially poor for certain groups and thus there was a high potential for bias. Twenty-five articles (32%) did not evaluate their datasets for bias, and five articles (7%) tested for bias but were unable to find any. Fifteen articles (20%) observed that data is differentially incomplete; 23 (30%) noted misclassification bias. Results about bias associated with specific social determinants is presented below and summarized in Table 3.

Table 3. Findings about bias and differential data quality

BIAS?	Social determinant	Bias type	% of literature reporting finding
Yes	Race/ethnicity	misclassification	25.0%
		Differentially incomplete (MNAR)	11.8%
		Differentially implausible (INAR)	2.6%
		other	1.3%
	Insurance	MNAR	1.3%
	Occupation	Differentially incomplete (MNAR)	3.9%
	Community Level	Rural data is problematic	3.9%
		other	3.9%
	Environmental	misclassification	3.9%
	Nonspecific	Differentially incomplete (MNAR)	2.6%
No	Did not evaluate for bias		32.9%
	Evaluated for bias and found none		6.6%

Key: usual source for these variables within the patient record

administrative or demographic sources	patient address is geocoded to link community-level data	diagnosis codes
---------------------------------------	--	-----------------

Race/ethnicity/country-of-origin variables

Plausibility. Articles that discussed race, ethnicity, or country-of-origin data were most concerned with plausibility (i.e., accuracy), which was discussed in 26 of the 40 articles (65%). Eighty-five percent of these articles (22 of 26) noted the potential for implausible data to cause error or bias in research, most commonly with misclassification. Three studies (12%) looked for bias but did not find any (20-22).

Misclassification bias, i.e., incorrect assignment, was noted as a problem or a potential problem in 18 of the 26 (69%) articles about the plausibility of race/ethnicity data (10, 12, 23-38). Further, several studies reported that implausible data and misclassification errors were more likely for certain groups. Fourteen studies reported that Hispanic patients were more likely to be misclassified;(10, 23-25, 29, 30, 32, 33, 36-41) four of these also found this to be true for Asian patients.(24, 30, 33, 36) Six studies found disproportionately high rates of misclassification for American Indians in comparison to other racial/ethnic

groups.(27, 28, 30, 32, 41, 42)

Completeness. The remaining fifteen studies that looked at race/ethnicity data primarily examined its completeness (38%). Ten of these (66%) identified that the incomplete data led to validity issues, most commonly that this data was not missing at random and had the potential to introduced bias (11, 34, 43-50).

Geocoded patient address data used for linkage to community-level variables. Articles about geocoded patient address data, on the other hand, were largely concerned with relational conformance (i.e., linkage match rates for geocoding), which was examined in 10 of the 16 articles (63%). Plausibility, i.e. accuracy, was the primary concern of four of the articles about geocoded patient address data.

Overall, 43% (n = 7) of the studies about geocoded patient address data acknowledged or established a potential for bias in their datasets. The remainder looked at data quality but did not evaluate their datasets for any validity problems that may result. Although race/ethnicity data was mostly plagued with a single type of error (i.e., misclassification), geocoded address data linked with community-level data was associated with multiple forms, including cartographic confounding and Type II error.(51)

Another study characterized the issues encountered with geocoded patient data as the distinction between individual and community-level variables. They found that “the accuracy of the community-level data for identifying patients with and without social risks was 48.0%.” The authors noted that use of this data for patient risk stratification “may heighten the risk of ecologic fallacy, wherein incorrect assumptions are made about an individual based on aggregate-level information about a group.”(52)

It should also be noted that the quality problems found with geocoded patient address data were not randomly distributed; for example, relational conformance (i.e., match rates) tend to be poorer for rural areas and certain parts of the country.(53-55)

Environmental health variables. Because information about exposure to toxins is rarely recorded

in the medical record, geocoded patient addresses are used to link health information to data about the environment. We found five studies discussing the use of patient addresses for exposure assessment. As with the geocoded community-level variables discussed above, relational conformance was represented in the majority of the articles about the quality of the datasets used for environmental health (60%).

Four (80%) of the included articles explored the impact of residential mobility on exposure assessment; i.e., whether patients' moving impacted the results of studies looking at environmental outcomes.(14, 15, 56, 57) In all of these articles, bias was characterized as misclassification of exposure to contaminants, an issue that has been noted elsewhere to be a source of Type II error in environmental health research.(58)

Non-specific social determinants. Six of the seven articles which addressed social determinants data as a broad, general category assessed the completeness of this information (86%); one addressed plausibility. Two of the articles mentioned the potential for bias, in both cases due to data missing nonrandomly.(5, 59) Three of the articles looked at the completeness of ICD Z- or V-codes, diagnostic codes which can be used by clinicians to collect SDoH data from patients.(3-5) All three articles concluded that clinicians were utilizing ICD codes to represent social determinants of health around 2% of the time.

Occupation. Six of the seven studies (86%) which looked at the quality of occupational data primarily examined completeness, all noting that occupational information is frequently missing from the health record.(9, 60-63) Three studies found that data was not missing at random, and that male patients were more likely to have occupational information in their record.(60, 62, 63)

Recommendations from the literature. The second theme we identified in our analysis is that there are solutions researchers can use to mitigate the issues caused by data quality problems. Forty-seven of the articles (62%) made at least one evidence-based recommendation for researchers seeking to improve the quality of social determinants data after it has been collected. We grouped these recommendations into six suggestions, which are detailed below and briefly summarized in Table 4.

Table 4. Summary of recommendations found in the articles

Recommendation	Sources
1. Avoid complete case analysis	23,50, 65
2. Impute data	14,22, 29, 30, 36, 27, 45, 46, 50, 61, 65-71
3. Rely on multiple sources	13,26,28,49,60,62,64,72
4. Use validated software tools	10,12,54,60,62,73–77
5. Select addresses thoughtfully	15,16,56,57

Avoid complete case analysis. It is a common practice to exclude incomplete (i.e., missing) data from the analysis, a method also known as casewise deletion or complete case analysis. However, three studies in our review found that casewise deletion decreased the quality of race/ethnicity data.(23, 50, 64) Grundmeier found that using only complete cases “produced highly biased results,” and in fact reversed the odds ratio for the Black subjects in their dataset.(64) Brown et al. found that their “race and ethnicity coefficient estimates are often biased downwards either toward zero or more negative when data with missing race and ethnicity is dropped.”(50) In all three studies, imputation was recommended as preferable to casewise deletion of missing data.

In a study on using patient addresses to determine pesticide exposures, Ling et al noticed that there were significant differences between patients with complete address information and those missing information.(57) In particular, Hispanic women born in Mexico and people living in poor neighborhoods were more likely to have missing addresses. For studies that rely on patient address to determine exposure, this means that these groups are more likely to be excluded from the analysis, potentially biasing the results.

One additional study about geocoding of patient addresses did not evaluate for bias, but did note that “unmatched addresses tend to be unevenly distributed – more likely to occur in rural areas and newly developed suburban areas, and less likely to occur in inner-city areas.”(53) In other words, complete case analysis would likely exclude a disproportionate number of rural patients.

Impute data. Several studies looked at the use of imputation, also referred to as indirect estimation,

as a method for increasing the completeness of missing data.(13, 21, 30, 45, 46, 50, 61, 64-68) Most of these studies used imputation on race/ethnicity data, although several looked at imputing geocoded patient address location for linking data,(65, 67, 68) and one looked at imputing occupational data.(61)

The most widely researched imputation method was Bayesian Improved Surname Geocoding (BISG), used to supplement missing race/ethnicity data.(21, 45, 46, 50, 64) In their study on the use of BISG, Dembosky et al found that imputation “did not substantially alter the estimated overall racial/ethnic distribution, but it did modestly increase sample size and statistical power.”(46)

Imputation was recommended not only as a solution for missing data, but also to improve the accuracy of implausible data.(22, 29, 30, 36, 37, 69) Methods involving Spanish surname coding were investigated in several studies to increase the reliability and consistency of data from Hispanic patients.(22, 29, 36, 37) Two articles validated similar, surname-based methods for data from Asian/Pacific Islander patients.(30, 69) All of these studies confirmed that these techniques reduced misclassification errors and enhanced data quality for their respective populations.

Finally, two additional studies looked at imputing race/ethnicity data with anonymized clinical datasets (13, 66), a situation where patient identifying information has been removed, and therefore it is not possible to use imputation methods that rely on patient surname or geocoded address. Ma et al compared four imputation methods and found that conditional multiple imputation “substantially improved statistical inferences for racial health disparities research.”(13)

Rely on multiple sources. This method compares and links data across an individual’s record in order to fill in or correct missing data fields, and was recommended by several studies as a way to either to increase completeness (26, 28, 49, 60, 62, 70) and/or check the accuracy of implausible data.(12, 26, 28) For example, Smith et al. recommended supplementing race/ethnicity data from the EHR with birth certificate data to not only increase completeness and plausibility.(26) Another study used NLP to “improve the identification of race and ethnicity in EHR data.”(49) Those researchers used NLP to comb through the unstructured text fields in clinical notes, then used the results to augment race/ethnicity data missing from

structured fields.

Use validated software tools. Several articles evaluated a specific software tool, usually one developed by the authors, for either assessing or increasing the quality of the data. Articles about geocoding patient address data, particularly, evaluated the ability of specific geocoding tools, such as ArcGIS, to increase data quality by improving match rates or positional accuracy.(54, 71-73) For researchers seeking to increase the quality of occupation data prior to analysis, The National Institute of Occupational Safety and Health has created a free, web-based coding system called NIOCCS (74) that was evaluated and recommended by several studies.(9, 60, 62, 75)

Some studies tested data quality assessment tools. For example, the Data Quality Assessment Tool was created to evaluate patient records at Community Health Centers.(76)The Data Completeness Analysis Package, as the name implies, evaluates completeness, and although it is not specific to social determinants data it has been used to appraise the completeness of SDoH variables.(11)

Select addresses thoughtfully. Patients move over addresses over time, which means that researchers often have decisions to make about which address to use. When confronted with data from patients who may have moved several times over the study period, three studies concluded that a single patient address was sufficient;(14, 56, 57) one study recommended using the most recent address,(56) while the other two concluded that address at birth was adequate for research on the effect of early exposures.(14, 57) However, a study by Brokamp et al compared the effect of using the most recent patient address, birth address, and an average across various addresses.(15) They found that using the most recent address or address at birth for a mostly urban population monitored over a seven year period could create bias towards the null.

Be transparent. One article suggested that researchers be honest about the amount and type of missing data.(77) Similarly, another recommended making an effort to quantify and report error in the dataset.(78)

2.5 Discussion

The data quality issues represented in our review found in similar proportions to those in Weiskopf and Weng's review examining the quality of EHR data for secondary use.(79) This both validates our findings and suggests that social determinants of health data suffers from quality problems similar to other data in the medical record.

Our review found that the category of data quality problem varies depending on the variable. This means that the kind of error created by these data quality problems also varies based on the social determinant factor in question. Most notably, problems encountered with the quality of social determinants of health data do not occur randomly. Although many researchers are aware that data "missing not at random," commonly abbreviated to MNAR, can cause bias during analysis, fewer are aware of the problems associated with other kinds of "data quality not at random." However, problems with plausibility not at random – e.g., the accuracy of data for Hispanic patients being lower than the accuracy of data from white patients in the same dataset – has similarly profound implications.

Several articles documented that race/ethnicity/country of origin data tends to be recorded inconsistently across a patient's record, especially for Hispanic patients. Why is data quality so poor for this group? Thirteen studies speculated that this may be due to the "fluid, debatable, and problematic" (34) nature of the definition of race and ethnicity.(10, 22, 24, 25, 27, 34, 37, 40, 41, 45, 46, 80, 81) Race is, after all, not a biological category but a social and political one,(82) and thus its terminology shifts over time. As terminology changes, so does the way we use that terminology to describe ourselves and others. Pellegrin et al noted that the fluidity of these definitions leads patients to respond inconsistently to questions about their race/ethnicity, thus causing problems with data reliability, and the fact that these categories are so broad and poorly defined leads to difficulties with data validity.(40)

At the institutional level, several studies speculate that the quality of data about ethnicity has been impaired by variation in how healthcare systems record and handle this information. As one study noted, "inconsistent classification of Hispanics is likely attributable, in part, to differences in the definition of being

Hispanic across classification schemes.”(27) Because the gold standard for race/ethnicity is widely considered to be patient self-report,(12, 24, 32, 34, 37, 46, 80, 83) it is possible that the increasing use of dynamic patient-facing data entry tools may allow people to inform and correct their own demographic information, thus helping to improve the quality of race, ethnicity, and country-of-origin data in the future. (80)

Limitations. It should be noted that the data quality concepts used here are not completely orthogonal or distinct. For example, several studies found that the plausibility of the patient race/ethnicity information in their dataset was questionable because the data was incomplete; in other words, plausibility was found to be low because a patient’s race was reported correctly in one area, but as “other” in another.(20, 84, 85)

The snowball sampling approach we used may have caused some research areas to be under-represented. Citation searching is inherently exponential; if our initial search turned up few articles within a certain domain, then that domain may appear to have a smaller body of literature than is in fact present. In addition, publication bias may have affected our findings if authors did not report negative findings when evaluating their datasets for bias.

Although some social determinants variables may have more thoroughly documented data quality issues, this does not mean that those variables are of poor quality. A larger body of research may indicate simply that these variables are more accessible to researchers and therefore easier to study. For example, our finding that articles about race/ethnicity data were such a large proportion of the literature may reflect that this information is more readily available in the structured fields of the health record than other variables.

Another limitation is that our search for solutions to these data quality problems is focused on the needs of researchers using observational databases. Because researchers require ex post facto methods for improving secondary use data, we ignored any recommendations from in the literature about improving data collection practices at the point of care.

2.6 Conclusions

The types of quality problems found with social determinants of health data varies depending on the variable; race/ethnicity data from the health record can be implausible or incomplete, while linked community-level data is prone to problems with non-conformance as well as plausibility. Similarly, data quality problems can lead to corresponding issues of validity and reliability; race/ethnicity data that are implausible or missing at random may lead to misclassification bias, while problems with geocoding can lead to misclassification, confounding, or ecologic fallacy. Several studies have documented that data quality from Hispanic patients can be particularly implausible and is especially prone to misclassification bias when compared to data from other racial/ethnic groups.

Fortunately, evidence-based solutions are available for researchers who want to improve the quality of social determinants data ex post facto. While complete case analysis has the potential for bias, imputation techniques can avoid these shortcomings. Consideration of data quality by researchers prior to analysis, along with thoughtfully applied quality improvement methods, may help prevent bias and improve the validity of research conducted with social determinants of health data.

2.7 References

1. Work with new electronic 'brains' opens field for Army math experts. *The Times*. 1957, Nov 10.
2. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)*. 2016;4(1):1244.
3. Torres JM, Lawlor J, Colvin JD, Sills MR, Bettenhausen JL, Davidson A, et al. ICD Social Codes: An Underutilized Resource for Tracking Social Needs. *Med Care*. 2017;55(9):810-6.
4. Guo Y, Chen Z, Xu K, George TJ, Wu Y, Hogan W, et al. International Classification of Diseases, Tenth Revision, Clinical Modification social determinants of health codes are poorly used in electronic health records. *Medicine*. 2020;99(52):e23818.
5. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of Social Determinants of Health ICD-10 Z-Codes Among Hospitalized Patients in the United States, 2016–2017. *Med Care*. 2020;58(12):1037-43.
6. IOM (Institute of Medicine). Capturing social and behavioral domains and measures in electronic health records: Phase 2. Washington, DC: The National Academies Press; 2014.
7. World Health Organization (WHO). Social determinants of health 2021 [Available from: https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1].
8. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. *JAMIA Open*. 2018;2(1):81-8.

9. Freeman MB, Pollack LA, Rees JR, Johnson CJ, Rycroft RK, Rousseau DL, et al. Capture and coding of industry and occupation measures: Findings from eight National Program of Cancer Registries states. *Am J Ind Med.* 2017;60(8):689-95.
10. Magana Lopez M, Bevans M, Wehrle L, Yang L, Wallen GR. Discrepancies in Race and Ethnicity Documentation: a Potential Barrier in Identifying Racial and Ethnic Disparities. *J Racial Ethn Health Disparities.* 2016.
11. Nasir A, Liu X, Gurupur V, Qureshi Z. Disparities in patient record completeness with respect to the health care utilization project. *Health Informatics Journal.* 2017;25(2):401 - 16.
12. Zingmond DS, Parikh P, Louie R, Lichtensztajn DY, Ponce N, Hasnain-Wynia R, et al. Improving Hospital Reporting of Patient Race and Ethnicity—Approaches to Data Auditing. *Health Serv Res.* 2015;50(S1):1372-89.
13. Ma Y, Zhang W, Lyman S, Huang Y. The HCUP SID Imputation Project: Improving Statistical Inferences for Health Disparities Research by Imputing Missing Race Data. *Health Serv Res.* 2018;53(3):1870-89.
14. Chen L, Bell EM, Caton AR, Druschel CM, Lin S. Residential mobility during pregnancy and the potential for ambient air pollution exposure misclassification. *Environ Res.* 2010;110(2):162-8.
15. Brokamp C, LeMasters GK, Ryan PH. Residential mobility impacts exposure assessment and community socioeconomic characteristics in longitudinal epidemiology studies. *J Expo Sci Environ Epidemiol.* 2016;26(4):428-34.
16. Bryman A. *Social Research Methods.* 5th ed: Oxford University Press; 2016. 314 p.
17. Huberman AM, Miles MB. *Qualitative data analysis : an expanded sourcebook.* 2nd ed. ed. Huberman AM, editor. Thousand Oaks: Thousand Oaks : Sage Publications; 1994.
18. Ezzy D, Liamputtong P. *Qualitative research methods : a health focus.* Ezzy D, editor. South Melbourne, Vic.: South Melbourne, Vic. : Oxford University Press; 1999.
19. Boyatzis RE. *Transforming qualitative information: thematic analysis and code development.* Thousand Oaks, CA: Thousand Oaks, CA : Sage Publications; 1998.
20. Hamilton NS, Edelman D, Weinberger M, Jackson GL. Concordance Between Self-Reported Race Ethnicity and that Recorded in a Veteran Affairs Electronic Medical Record. *N C Med J.* 2009;70(4):296-300.
21. Derose SF, Contreras R, Coleman KJ, Koebnick C, Jacobsen SJ. Race and Ethnicity Data Quality and Imputation Using U.S. Census Data in an Integrated Health System: The Kaiser Permanente Southern California Experience. *Med Care Res Rev.* 2013;70(3):330-45.
22. Wei II, Virnig BA, John DA, Morgan RO. Using a Spanish Surname Match to Improve Identification of Hispanic Women in Medicare Administrative Data. *Health Serv Res.* 2006;41(4p1):1469-81.
23. Lee SJ, Grobe JE, Tiro JA. Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals. *J Am Med Inform Assoc.* 2016;23(3):627-34.
24. Grafova IB, Jarrin OF. Beyond Black and White: Mapping Misclassification of Medicare Beneficiaries Race and Ethnicity. *Med Care Res Rev.* 2020;1077558720935733.
25. Webster PS, Fulton JP, Sampangi S. Conflicting Race/Ethnicity Reports: Lessons for Improvement in Data Quality. *Journal of Registry Management.* 2013;40(3):122 - 6.
26. Smith N, Iyer RL, Langer-Gould A, Getahun DT, Strickland D, Jacobsen SJ, et al. Health plan administrative records versus birth certificate records: quality of race and ethnicity information in children. *BMC Health Serv Res.* 2010;10:316.
27. Gomez SL, Glaser SL. Misclassification of race/ethnicity in a population-based cancer registry (United States). *Cancer Causes Control.* 2006;17(6):771-81.
28. Bigback KM, Hoopes M, Dankovchik J, Knaster E, Warren-Mears V, Joshi S, et al. Using Record Linkage to Improve Race Data Quality for American Indians and Alaska Natives in Two Pacific Northwest State Hospital Discharge Databases. *Health Serv Res.* 2015;50 Suppl 1:1390-402.
29. Pinheiro PS, Sherman R, Fleming LE, Gomez-Marin O, Huang Y, Lee DJ, et al. Validation of Ethnicity in Cancer Data: Which Hispanics Are We Misclassifying? *Journal of Registry Management.* 2009;36(2):42 - 6.

30. Eicheldinger C, Bonito A. More Accurate Racial and Ethnic Codes for Medicare Administrative Data. *Health Care Financ Rev.* 2008;29(3):27-42.
31. Fiscella K, Fremont AM. Use of geocoding and surname analysis to estimate race and ethnicity. *Health Serv Res.* 2006;41(4 Pt 1):1482-500.
32. West CN, Geiger AM, Greene SM, Harris EL, Liu IL, Barton MB, et al. Race and ethnicity: comparing medical records to self-reports. *J Natl Cancer Inst Monogr.* 2005(35):72-4.
33. Waldo DR. Accuracy and Bias of Race Ethnicity Codes in the Medicare Enrollment Database. *Health Care Financ Rev.* 2004;26(2).
34. Klinger EV, Carlini SV, Gonzalez I, Hubert SS, Linder JA, Rigotti NA, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. *J Gen Intern Med.* 2015;30(6):719-23.
35. Swallen KC, Glaser SL, Stewart SL, West DW, Jenkins CNH, McPhee SJ. Accuracy of racial classification of Vietnamese patients in a population-based cancer registry. *Ethn Dis.* 1998;8(2):218 - 27.
36. Polednak AP. Agreement in race-ethnicity coding between a hospital discharge database and another database. *Ethn Dis.* 2001;11(1):24-9.
37. Morgan RO, Wei II, Virnig BA. Improving Identification of Hispanic Males in Medicare: Use of Surname Matching. *Med Care.* 2004;42(8):810-6.
38. Lauderdale D, Goldberg J. The expanded racial and ethnic codes in the Medicare data files: their completeness of coverage and accuracy. *Am J Public Health.* 1996;86(5):712-6.
39. Lee W-C, Veeranki SP, Serag H, Eschbach K, Smith KD. Improving the Collection of Race, Ethnicity, and Language Data to Reduce Healthcare Disparities: A Case Study from an Academic Medical Center. *Perspectives in Health Information Management.* 2016;13.
40. Pellegrin KL, Miyamura JB, Ma C, Taniguchi R. Improving Accuracy and Relevance of Race/Ethnicity Data: Results of a Statewide Collaboration in Hawaii. *J Healthc Qual.* 2016;38(5):314-21.
41. Gomez SL, Kelsey JL, Glaser SL, Lee MM, Sidney S. Inconsistencies between self-reported ethnicity and ethnicity recorded in a health maintenance organization. *Ann Epidemiol.* 2005;15(1):71-9.
42. Fiscella K, Meldrum S. Race and ethnicity coding agreement between hospitals and between hospital and death data. *Med Sci Monit.* 2008;14(3):SR9 - 13.
43. Gomez SL, Glaser SL, Kelsey JL, Lee MM. Bias in Completeness of Birthplace Data for Asian Groups in a Population-Based Cancer Registry (United States). *Cancer Causes Control.* 2004;15(3):243-53.
44. Lin SS, O'Malley CD, Lui SW. Factors associated with missing birthplace information in a population-based cancer registry. *Ethn Dis.* 2001;11(4):598-605.
45. Haas A, Elliott MN, Dembosky JW, Adams JL, Wilson-Frederick SM, Mallett JS, et al. Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. *Health Serv Res.* 2019;54(1):13-23.
46. Dembosky JW, Haviland AM, Haas A, Hambarsoomian K, Weech-Maldonado R, Wilson-Frederick SM, et al. Indirect estimation of race ethnicity for survey respondents who do not report race/ethnicity. *Med Care.* 2019;57(5):e28-e33.
47. Gomez SL, Glaser SL. Quality of cancer registry birthplace data for Hispanics living in the United States. *Cancer Causes Control.* 2005;16(6):713-23.
48. Chen Y, Lin HY, Tseng TS, Wen H, DeVivo MJ. Racial Differences in Data Quality and Completeness: Spinal Cord Injury Model Systems' Experiences. *Top Spinal Cord Inj Rehabil.* 2018;24(2):110-20.
49. Sholle ET, Pinheiro LC, Adekkanattu P, Davila MA, Johnson SB, Pathak J, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J Am Med Inform Assoc.* 2019;26(8-9):722-9.
50. Brown DP, Knapp C, Baker K, Kaufmann M. Using Bayesian Imputation to Assess Racial and Ethnic Disparities in Pediatric Performance Measures. *Health Serv Res.* 2016;51(3):1095-108.
51. Sherman RL. Address at Diagnosis: Place Matters. *Journal of Registry Management.* 2017;44(2):76-7.
52. Cottrell EK, Hendricks M, Dambrun K, Cowburn S, Pantell M, Gold R, et al. Comparison of Community-Level and Patient-Level Social Risk Data in a Network of Community Health Centers. *JAMA Network Open.* 2020;3(10):e2016852-e.

53. Lin G, Gray J, Qu M. Improving geocoding outcomes for the Nebraska Cancer Registry: learning from proven practices. *Journal of Registry Management*. 2010;37(2):49 - 56.
54. Dilekli N, Janitz A, Campbell J. Improved Geocoding of Cancer Registry Addresses in Urban and Rural Oklahoma. *Journal Registry Management*. 2020;47(1):13-20.
55. Wilkins R. Use of postal codes and addresses in the analysis of health data. *Health Rep*. 1993;5(2):157-77.
56. Hughes AE, Pruitt SL. The utility of EMR address histories for assessing neighborhood exposures. *Ann Epidemiol*. 2017;27(1):20-6.
57. Ling C, Heck JE, Cockburn M, Liew Z, Marcotte E, Ritz B. Residential mobility in early childhood and the impact on misclassification in pesticide exposures. *Environ Res*. 2019;173:212-20.
58. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med*. 1998;55:651-6.
59. Cottrell EK, Dambrun K, Cowburn S, Mossman N, Bunce AE, Marino M, et al. Variation in Electronic Health Record Documentation of Social Determinants of Health Across a National Network of Community Health Centers. *Am J Prev Med*. 2019;57(6, Supplement 1):S65-S73.
60. McClure LA, Koru-Sengul T, Hernandez MN, Mackinnon JA, Schaefer Solle N, Caban-Martinez AJ, et al. Availability and accuracy of occupation in cancer registry data among Florida firefighters. *PLoS One*. 2019;14(4):e0215867.
61. Scott E, Hirabayashi L, Graham J, Krupa N, Jenkins P. Using hospitalization data for injury surveillance in agriculture, forestry and fishing: a crosswalk between ICD10CM external cause of injury coding and The Occupational Injury and Illness Classification System. *Injury Epidemiology*. 2021;8(1):6.
62. Silver SR, Tsai RJ, Morris CR, Boiano JM, Ju J, Scocozza MS, et al. Codability of industry and occupation information from cancer registry records: Differences by patient demographics, casefinding source, payor, and cancer type. *Am J Ind Med*. 2018;61(6):524-32.
63. Armenti KR, Celaya MO, Cherala S, Riddle B, Schumacher PK, Rees JR. Improving the quality of industry and occupation data at a central cancer registry. *Am J Ind Med*. 2010;53(10):995-1001.
64. Grundmeier RW, Song L, Ramos MJ, Fiks AG, Elliott MN, Fremont A, et al. Imputing Missing Race/Ethnicity in Pediatric Electronic Health Records: Reducing Bias with Use of U.S. Census Location and Surname Data. *Health Serv Res*. 2015;50(4):946-60.
65. Curriero FC, Kulldorff M, Boscoe FP, Klassen AC. Using imputation to provide location information for nongeocoded addresses. *PLoS One*. 2010;5(2):e8998.
66. Kim J-S, Gao X, Rzhetsky A. RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning. *PLoS Comput Biol*. 2018;14(4):e1006106.
67. Henry KA, Boscoe FP. Estimating the accuracy of geographical imputation. *Int J Health Geogr*. 2008;7:3.
68. Hibbert JD, Liese AD, Lawson A, Porter DE, Puett RC, Standiford D, et al. Evaluating geographic imputation approaches for zip code level data: an application to a study of pediatric diabetes. *Int J Health Geogr*. 2009;8:54.
69. Hsieh M-C, Pareti LA, Chien VW. Using NAPIIA to improve the accuracy of Asian race codes in registry data. *Journal of Registry Management*. 2011;38(4):190 - 5.
70. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P. Post Office Box Addresses: A Challenge for Geographic Information System-Based Studies. *Epidemiology*. 2003;14(4).
71. Kumar S, Liu M, Hwang S-A. A multifaceted comparison of ArcGIS and MapMarker for automated geocoding. *Geospatial health*. 2012;7(1):145 - 51.
72. Goldberg DW, Wilson JP, Knoblock CA, Ritz B, Cockburn MG. An effective and efficient approach for manually improving geocoded data. *International Journal of Health Geographics*. 2008;7(1):60.
73. Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match Rate and Positional Accuracy of Two Geocoding Methods for Epidemiologic Research. *Ann Epidemiol*. 2006;16(11):842-9.
74. National Institute for Occupational Safety and Health (NIOSH). NIOSH Industry and Occupation Computerized Coding System (NIOCCS): Centers for Disease Control and Prevention (CDC); [Available from: <https://csams.cdc.gov/nioccs/>].

75. Weiss NS, Cooper SP, Socias C, Weiss RA, Chen VW. Coding of Central Cancer Registry Industry and Occupation Information: The Texas and Louisiana Experiences. *J Registry Manag.* 2015;42(3):103-10.
76. Laberge M, Shachak A. Developing a tool to assess the quality of socio-demographic data in community health centres. *Appl Clin Inform.* 2013;4(1):1-11.
77. Kinnee EJ, Tripathy S, Schinasi L, Shmool JLC, Sheffield PE, Holguin F, et al. Geocoding Error, Spatial Uncertainty, and Implications for Exposure Assessment and Environmental Epidemiology. *Int J Environ Res Public Health.* 2020;17(16).
78. Strickland MJ, Siffel C, Gardner BR, Berzen AK, Correa A. Quantifying geocode location error using GIS methods. *Environmental Health.* 2007;6(1):10.
79. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc.* 2013;20(1):144-51.
80. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, et al. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc.* 2019;26(8-9):730-6.
81. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. *BMC Med Inform Decis Mak.* 2018;18(1):139.
82. HL7. v3 Code System Race. Terminology Value Sets: HL7 International; 2018. Available from: <https://www.hl7.org/fhir/v3/Race/cs.html>.
83. Pinto AD, Glattstein-Young G, Mohamed A, Bloch G, Leung F-H, Glazier RH. Building a Foundation to Reduce Health Inequities: Routine Collection of Sociodemographic Data in Primary Care. *The Journal of the American Board of Family Medicine.* 2016;29(3):348-55.
84. Kressin NR, Chang B-H, Hendricks A, Kazis LE. Agreement Between Administrative Data and Patients' Self-Reports of Race/Ethnicity. *Am J Public Health.* 2003;93(10):1734-9.
85. Maizlish N, Herrera L. Race/ethnicity in medical charts and administrative databases of patients served by community health centers. *Ethn Dis.* 2006;16:483-7.

Chapter 3. Evaluating patient race and ethnicity data for secondary use: variability in EHR data about race and ethnicity in the National Covid Cohort Collaborative Data Enclave

Note: A two-page version was accepted as a podium abstract at the 2022 AMIA Informatics Summit in Chicago. The full manuscript is currently in the process with JMIR: Medical Informatics, and is available as a pre-print.

Presentation citation:

Cook LA, Espinoza J, Weiskopf NG, Mathews N, Dorr D, Gonzalez KL, et al. Variability in EHR data about race and ethnicity as observed in the National COVID Cohort Collaborative Data Enclave. Podium Abstract presented at: AMIA Informatics Summit; 2022 Mar 22; Chicago, IL.

Preprint citation:

Cook L, Espinoza J, Weiskopf NG, Mathews N, Dorr DA, Gonzales KL, Wilcox A, Madlock-Brown C, on behalf of the N3C Consortium. Variability in EHR Data About Race and Ethnicity As Observed in the National COVID Cohort Collaborative Data Enclave. JMIR Preprints. 09/05/2022:39235

Abstract

Background: A significant technical challenge related to integrating race and ethnicity data across EHR systems is the lack of consistency in how data about race and ethnicity is collected and structured by healthcare organizations.

Objective: To evaluate and describe variations in how healthcare systems collect and report information about the race and ethnicity of their patients, and how these data are integrated when it is aggregated into a large clinical database.

Methods: At the time of our analysis, the National COVID Cohort Collaborative (N3C) Data Enclave contained records from 6.5 million patients contributed by 56 healthcare institutions. We assessed the quality of race and ethnicity data by analyzing its conformance to federal standards, then drilled into the non-conforming data.

Results: “No matching category” was the second largest harmonized racial group in the N3C. 20.7% of the race data did not conform to the federal standard; the largest category was data that were missing. Hispanic or Latino patients were over-represented in the non-conforming racial data, and data from American Indian or Alaska Native patients were obscured. Although only a small proportion of the source data had not been mapped to the correct concepts (0.6%), Black or African-American and Hispanic/Latino patients were over-represented in this category.

Conclusions: The impact of data quality issues was not equal across all races and ethnicities, which has the potential to introduce bias in analyses and conclusions drawn from these data. The adverse impact of COVID-19 on marginalized and under-resourced communities of color has highlighted the need for accurate, comprehensive race and ethnicity data. Differences in how race and ethnicity data is conceptualized and encoded by healthcare institutions can affect the quality of the data in aggregated clinical databases. Transparency about how data has been transformed can help users make accurate analyses and inferences, and eventually better guide clinical care and public policy.

3.1 Introduction

The United States has had more COVID-19 cases and deaths than any other country.(1) Black or African American, Hispanic or Latino, and American Indian or Alaska Native communities have experienced disproportionate morbidity and mortality from COVID-19.(2 - 5) Compared to the non-Hispanic white population, the Black or African American population has a higher prevalence of COVID-19, as well as higher mortality and hospitalization rates from the virus.(2) The CDC reported that between February and May 2020, Hispanic or Latino and nonwhite individuals under 65 were two to three times more likely to die from COVID-19 than their non-Hispanic white counterparts.(4) COVID-19 incidence for American Indian or Alaska Native persons is estimated to be 3.5 times higher than for non-Hispanic white persons.(5) The full consideration of the social, economic, and health impacts of COVID-19 on these communities relies on datasets structured to answer such questions.

Resources have been created with the intention of tracking, quantifying, and analyzing the impact of COVID-19 within and across populations.(6-8) One such resource is the National COVID Cohort Collaborative (N3C), an NIH-funded collaboration between the National Center for Advancing Translational Sciences and the Center for Data to Health.(8) N3C's Data Enclave is one of the largest collections of COVID-19 patient-level data globally,(7) providing harmonized electronic health record (EHR) data from 56 healthcare institutions and networks across the country. Currently, 1,615

researchers representing 186 research institutions have been granted access to the Enclave to work on 215 research projects.(9)

Large, federated datasets like N3C, however, face a substantial challenge in the form of data heterogeneity, stemming from varying data collection, documentation, and coding practices.(10) These upstream processes may result in data quality problems and other artifacts that can lead to misleading signals in the data. EHR data on race and ethnicity are known to suffer from such data quality issues, including high rates of incompleteness (11) and low accuracy for certain racial groups.(12,13) Moreover, healthcare organizations increasingly are allowing patients to select more granular race and ethnicity subcategories to describe themselves;(14 - 18) however, the availability of these subcategories to researchers is largely unknown and varies across sites.

Data standardization and harmonization is one of the best tools for combating heterogeneity and ensuring that observed signals are genuine. The N3C provides a unique opportunity to assess how different healthcare systems in various locations collect and conceptualize information about their patients' race and ethnicity, and to examine efforts to integrate these categories across different data models. A note on language: In this paper, we discuss race and ethnicity from the perspective of data standards and database harmonization. In order to maintain clarity and consistency we have used the terms for race and ethnicity defined by the Office of Management and Budget and listed in **Table 1**.(19)

3.2 Background

The encodings used to represent race and ethnicity in the health record vary across institutions and data models, and require specialized harmonization.(20) Indeed, a significant technical challenge related to integrating race and ethnicity data across EHR systems is the lack of consistency in how data about race and ethnicity is collected and structured by healthcare organizations. The Institute of Medicine's landmark report on racial and ethnic disparities in healthcare, *Unequal Treatment: Confronting Racial and Ethnic Disparities in Healthcare*, highlighted the need for standardized collection and reporting of race and ethnicity data.(21)

The most commonly used standard healthcare systems use to collect and organize data about race and ethnicity was initially created for the 2000 US Census. The Office of Management and Budget (OMB) released this standard in 1997,(19) and shortly afterward the Center for Disease Control added encodings to the OMB Standard (see **Table 1**) (22). The OMB classification system was then adopted with minor changes by HL7, the creator of the standard most widely used by healthcare systems to transmit and receive health records;(23) any references to “the healthcare standard” in our paper refer to how this information is currently structured in HL7 v3.

Table 1. OMB race and ethnicity category definitions

OMB Category (HL7 Code)	Category Definition
Race: American Indian or Alaska Native (1002-5)	A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment
Race: Asian (2028-9)	A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam
Race: Black or African American (2054-5)	A person having origins in any of the black racial groups of Africa. Terms such as "Haitian"; or "Negro"; can be used in addition to "Black or African American"
Race: Native Hawaiian or Other Pacific Islander (2076-8)	A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands
Race: White (2106-3)	A person having origins in any of the original peoples of Europe, the Middle East, or North Africa
Ethnicity: Hispanic or Latino (2135-2)	A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race. Ethnicity is considered a distinct category from race

The current healthcare standard uses terminology in a manner different from how they are used colloquially. For the purposes of collecting and organizing patient demographic data, race and ethnicity are considered distinct concepts, and ethnicity refers only to Hispanic or Latino origin. Thus ethnicity has two minimum codes: patients can either be Hispanic or Latino or non-Hispanic or Latino. However, this category is intended to be hierarchical, and “granular” ethnicity refers to the 41 subcategories (Panamanian, Venezuelan, etc) that are required to roll up into Hispanic or Latino.

The 2009 IOM Subcommittee on Standardized Collection of Race/Ethnicity Data for

Healthcare Quality Improvement report provided direction for healthcare systems on how to implement the standard.(24) Because the healthcare standard treats race and ethnicity as separate concepts, it is recommended that the question about Hispanic or Latino origin be presented first when gathering demographic information from patients. The standard has five minimum categories for race: (a) American Indian or Alaska Native, (b) Asian, (c) Black or African American, (d) Native Hawaiian or Other Pacific Islander, and (e) white. The standard for race data is hierarchical, with almost 900 different subcategories that could be used to describe more granular race categories, all of which are required to collapse (or “roll up”) into one of the five major categories. “Other” race is deprecated within HL7, although “unknown” and “asked but not answered” are permissible.(23) For patients who identify as multiracial, OMB and the IOM Subcommittee both recommend allowing for the selection of more than one race rather than offering a single “multiracial” category.(21,24) However, the OMB acknowledged that allowing for multiple selections creates complications during tabulation and analysis, and the IOM noted that “some health information technology systems are unable to support the collection and reporting of data in a ‘Select one or more’ manner” (p. 73).

The federal standard only recommends a structure for how information about patient race and ethnicity should be stored; in practice, there are wide variations in how healthcare systems collect this information. Studies have documented that it is frequently missing from the patient record, and when it is collected it is often of poor quality.(12,13,25,26)

3.3 Methods

Data Source. Although the size of the N3C Enclave has continued to grow, at the time of our analysis (July and August of 2021) the N3C Enclave contained health records from 6.5 million patients tested for COVID-19, including 2.1 million who had tested positive. The data in the Enclave are updated twice weekly with new information. To keep our numbers consistent we used Release-v40-2021-07-30 to conduct analyses whenever possible; small numerical inconsistencies may appear as the result of occasions when different release versions were used.

Significant technical and regulatory hurdles were addressed to make the N3C Data Enclave available to researchers seeking insight into COVID-19. The clinical data is stored in the OMOP Common Data Model; institutions using ACT, PCORnet, and TriNetX CDMs have their data mapped OMOP, while those already using OMOP have their data ingested directly. N3C worked with the CDM Harmonization project to provide syntactic mapping with conversion logic and semantic mapping to the OMOP vocabulary. N3C met with subject matter experts from source CDMs and the OHSDI community to finalize these mappings, which are available to the public on GitHub.(19)

Defining non-conformance. We defined race and ethnicity data as “conforming” if they had been mapped to one of the five minimum categories for race congruent with the federal standard (21,23,24) (see **Table 1**). Ethnicity data were conforming if they were harmonized into one of the two standard categories for ethnicity: “Hispanic or Latino” or “Not Hispanic or Latino.” All other data - including missing data - were deemed “non-conforming.” Categories of non-conforming data (e.g., missing or misclassified) were induced from the data that had not been mapped directly to a standard category. For data models where source data were recorded in a separate field (i.e., non-OMOP models), we compared these data to the final harmonized data.

Analyses. In order to quantify the variability in how healthcare institutions are reporting data about patient race and ethnicity to N3C, we used a multi-step process that included data processing, terminology harmonization, and descriptive analyses. We first sorted the values in the harmonized, mapped race field into primary race categories. We then aggregated each category by source values and tabulated these results, enumerating how many patients had been reported with each encoded value and how many distinct ways each category was reported. We then cross-tabulated the resulting categories by three factors: patient ethnicity, the number of data partners using each code, and which data models utilized those particular encodings. For the non-conforming data, we used an inductive approach to sort the resulting values into categories. These analyses were conducted within the N3C Data Enclave using the software tools available within the platform during July and August 2021. Additional descriptive statistics were done with Excel. Figures were developed using Lucidchart, Excel,

and Keynote.

Comparison to other data sources and repositories. To assess the external validity of the N3C’s race and ethnicity data, we compared race and ethnicity distributions across multiple data sources, including the 2019 American Communities Survey (ACS) demographic and housing estimates and Cerner HealthFacts (CHF).(26) ACS data are compiled by the US Census Bureau and provide yearly updates and estimates to key demographic, economic, housing and social data. Cerner HealthFacts is a data warehouse that includes almost 70 million patients treated at hospitals and clinics throughout the United States between 2001 and 2017 using the Cerner EHR platform. Total unadjusted COVID-19 cases and deaths from the CDC are also included for comparison.(28,29)

Table 2. Race data reporting schema by contributing sites

Reporting Schema	Data Partners, n (%)	American Indian or Alaska Native	Asian	Black or African American	Native Hawaiian or Other Pacific Islander	White	Black	Hispanic	Asian Indian	Chinese	Filipino	Japanese	Korean	Other Pacific Islander	Polynesian	Vietnamese	
A	41 (73.2)		✓	✓	✓	✓											Standards Only
B	2 (3.6)		✓	✓	✓	✓								✓			Standards with Additional Categories
C	1 (1.8)		✓	✓	✓	✓		✓									
D	1 (1.8)		✓	✓	✓	✓			✓								
E	1 (1.8)		✓	✓	✓	✓			✓	✓	✓	✓	✓		✓	✓	Missing Standards
F	5 (8.9)		✓	✓		✓											
G	2 (3.6)			✓		✓											
H	1 (1.8)			✓	✓	✓											Missing Standards, Additional Categories
I	1 (1.8)		✓			✓	✓										
J	1 (1.8)		✓	✓		✓								✓			
Total Data Partners, n		0	53	55	49	56	1	1	2	1	1	1	1	3	1	1	
		Standard OMB Categories					Other Race Categories										

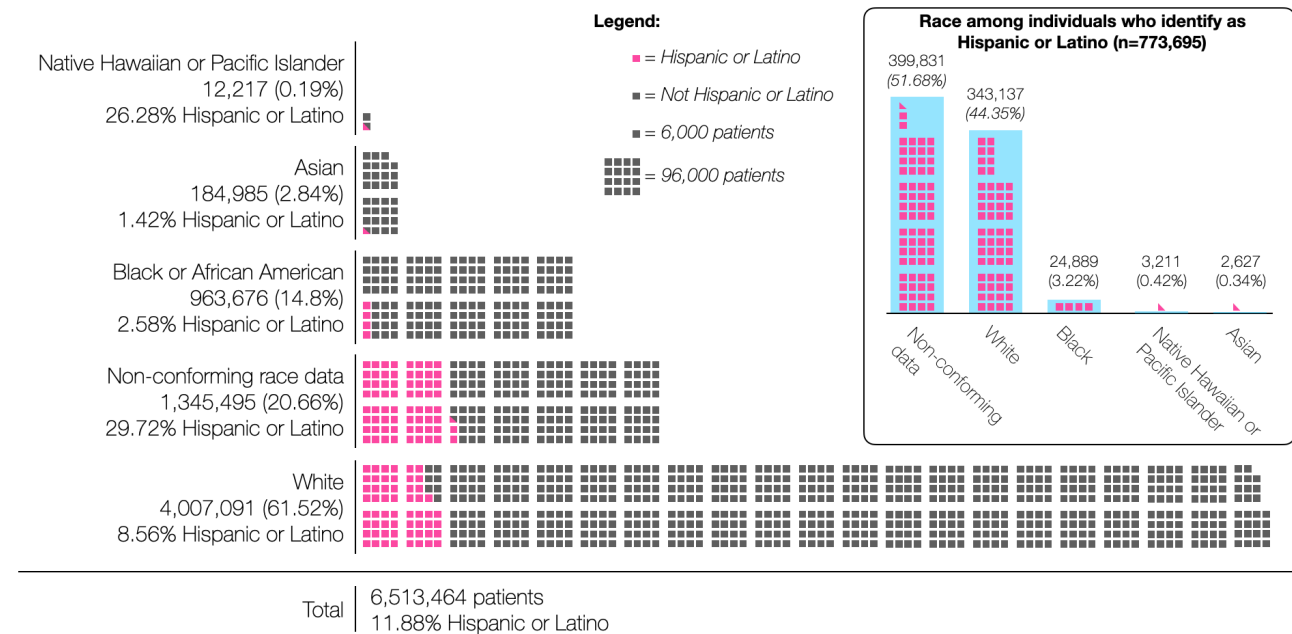
3.4 Results

There are a total of 25 harmonized categories for race available in the N3C Data Enclave, representing mapped data contributed by 56 healthcare institutions. The top three categories-- white, “No matching concept,” and Black or African American-- account for 97.0% of the data (**Figure 1**). No patients with race of American Indian or Alaska Native were found; we learned that the health records of American Indian or Alaska Native patients were deliberately obscured during the ingestion at the request of the NIH (see Discussion for an explanation).

Following ingestion and mapping to standard OMOP race categories, we identified 10 different reporting schema (combinations of reported race categories) among contributing institutions (**Table 2**). Forty-one (73.2%) data partners had harmonized patient race data that adhered to the standard OMB race categories other than American Indian or Alaska Native, as noted above. Five data partners had harmonized data that included all of the standard race categories, plus some additional categories such as Filipino or Korean that had not been correctly rolled up. Data from ten contributing institutions (17.9%) omitted at least one of the standard race categories other than American Indian or Alaska Native). The only OMB race category present for all data partners was white. Ethnicity as a separate field was present for 51 (91.1%) of the data partners.

Conforming Data. Examining the data that had been harmonized into categories that conformed with the OMB standard, **Figure 1** illustrates the racial and ethnic makeup of patients in N3C. White race was reported by healthcare institutions a total of 21 different ways, most commonly using the PCORnet code 05 (n = 1,442,961, 36.0% of all white patients). “Jewish” was the only granular category available for patients whose race had been mapped to white, and 141 of patients whose race had been mapped to white had Jewish ancestry recorded in the source data. The most common code used to report Black or African American patients was PCORnet’s simplified encoding 03 (n = 411,537; 43.9%). Although there were 24 different encodings for this group, there were no granular sub-categories or Black or African American available in either the source or the mapped

Figure 1. Reported race and ethnicity data in the N3C data enclave after harmonization



data.

Asian race was reported a total of 22 distinct ways, most commonly using the PCORnet code 02 (n = 77,426, 41.9% of all Asian patients). Source data revealed six more granular race categories had been rolled up into Asian; these more granular data represented 546 patients. As with the non-conforming data reported below, the most common sub-category available for Asian race was Asian Indian (n = 388). This is a smaller number of Asian Indian patients, however, than we found in the non-conforming race data.

Native Hawaiian or Other Pacific Islanders are the smallest of the primary racial groups found in the Enclave, and their data was reported using 24 different encodings. We found two subcategories of Native Hawaiian or Other Pacific Islander: Guamanian/Chamorro, and Polynesian. This group contained the largest percentage of people who also identified as Hispanic or Latino (see **Figure 1**).

Overall, 83.8% of the data about ethnicity conformed to the standard. There were a total of 26 different ways to represent Hispanic or Latino ethnicity in the source data, including seven granular subcategories such as Puerto Rican, Mexican, and South American.

Non-conforming Race Data. As shown in **Figure 2**, non-conforming data could be divided

into seven categories:

Missing: Incompleteness was the most common reason for race data to be non-conforming, and 11.3% of all patients in the N3C Data Enclave were marked as missing race data. Thirty-one of the contributing healthcare institutions reported missing data in 29 distinct ways; most often, a zero was recorded to indicate that the data were incomplete (n = 348,057). 13.6% of the patients missing data about race were noted as being Hispanic or Latino in the ethnicity column.

Other: The second largest category of non-conforming race data was patients labelled by healthcare systems as “Other” race. The majority of these patients (61.0%, n = 217,476) were recorded as being of Hispanic or Latino ethnicity. Note that all American Indian or Alaska Native patient data submitted to N3C is obscured and aggregated in the “Other” category (30).

Refused: Patients who declined to answer questions about race represented 8.0% (n = 107,445) of the non-conforming data. However, examining the data about ethnicity showed that 21.1% of these patients were Hispanic or Latino.

Multiracial: Multiracial patients represented 7.4% of the non-conforming data and 1.5% of all the patients in the N3C Data Enclave. Of the 257 different codes used by systems to represent race, 119 of them were distinct codes used to represent multiracial patients. Much of the variety was due to the fact that some systems allowed patients to select multiple races, and then reported all of the selections in a single column. Despite the fact that only 3.8% of all multiracial patients actually had multiple races recorded, this group represented 101 different combinations of codes. The most

common of these were combinations of white and Black or African American (n=1,563).

Misclassified (i.e., incorrectly mapped): Examining the source data revealed that 3.1% of the non-conforming race data, 0.6% of all data in the N3C, appears to have been incorrectly mapped during the harmonization process. Although only 14.8% of the patients in the Enclave are Black or

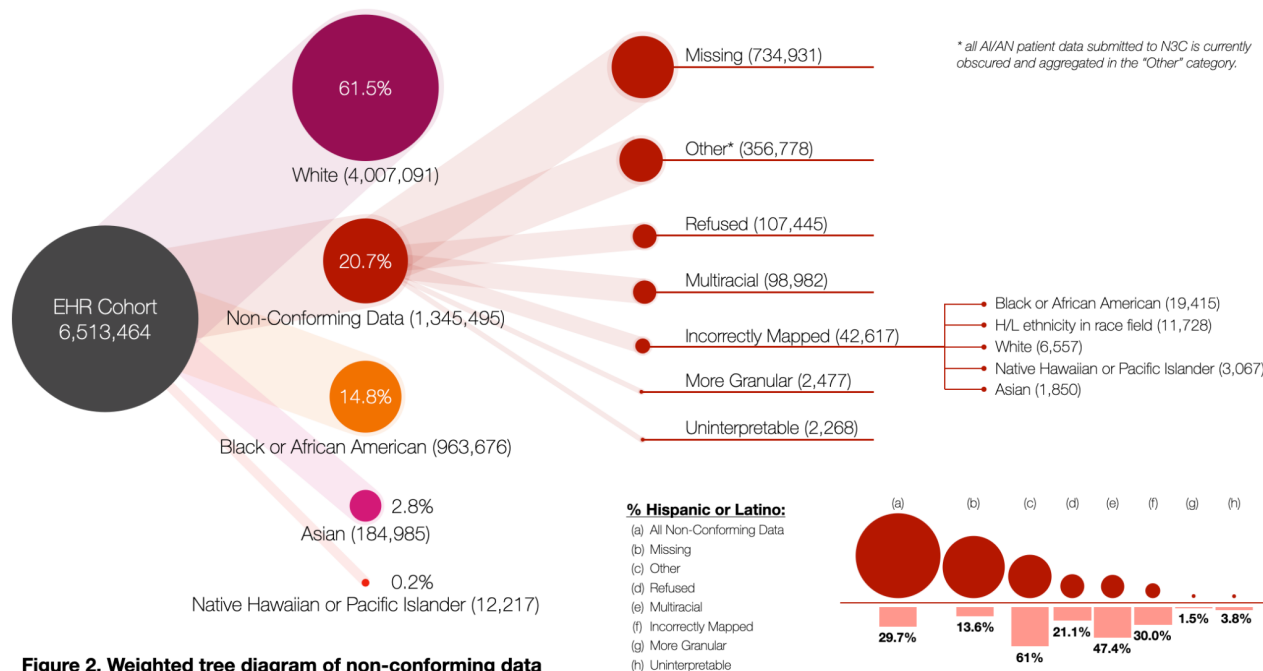


Figure 2. Weighted tree diagram of non-conforming data

African American, source data showed that 47.1% of these misclassified patients should have had their race mapped to Black or African American. The next largest group of misclassified patients were those whose source institutions had recorded Hispanic or Latino ethnicity in the race field (n = 11,728). Of note, 19.3% (n = 2,258) of the patients whose race was reported as Hispanic or Latino were labeled as Not Hispanic or Latino in the ethnicity field. Patients identified as white represented 15.9% of the misclassified non-conforming data; 7.4% of these misclassified patients were Native Hawaiian or Other Pacific Islander, and 4.5% were Asian. Because all American Indian or Alaska Native patients had their race transformed to "Other," (30) from a purely technical standpoint 100% of this group was misclassified. However, it is important to note that this was not the result of error.

Uninterpretable: For 2,268 patients, the source institution had provided a code such as "@" that did not conform to those recognized by any of the known data models. There were two encodings

we were unable to decipher, both of which came from institutions using the TriNetX Common Data Model.

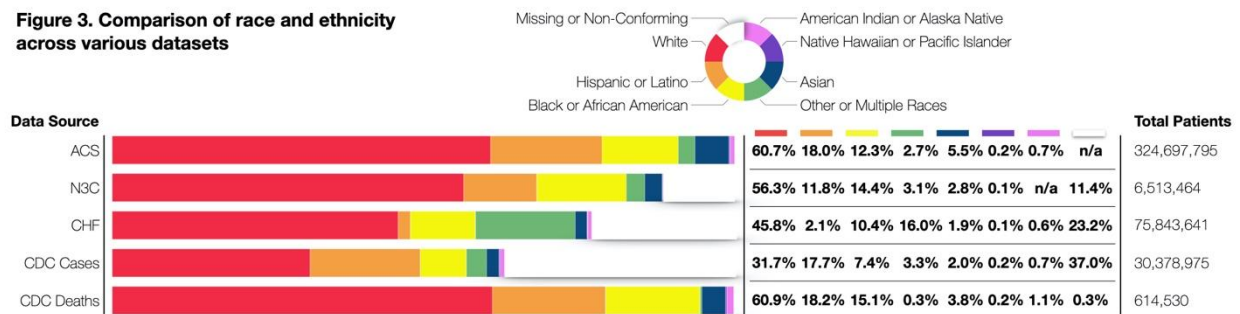
More granular. Finally, 2,477 patients did not map to one of the five categories because they had been labeled with a granular racial subcategory that had not been rolled up into one of the five main race categories. Nine racial subgroups are available in the non-conforming data in the N3C; six of these (Asian Indian, Filipino, Chinese, Korean, Vietnamese, and Japanese) should have been rolled up into the larger category of “Asian.” The most widely reported subcategory we found in the non-conforming data was “Asian Indian” (n = 1,534). These granular data all came from healthcare systems using the OMOP Common Data Model.

For patients with Hispanic or Latino ethnicity, “non-conforming” was the single largest racial category (51.7%; n = 399,831). One data partner mapped 2,533 patients whose race had originally been recorded as Hispanic or Latino to a racial category that was subsequently labelled only as “Non-white.”

Variations by Common Data Model. Four CDMs are used by the healthcare institutions contributing data to the N3C Enclave: OMOP, PCORnet, TriNetX, and ACT. Some OMOP sites also included data in the PEDSnet CDM, which is an extension of OMOP that includes pediatric specific data fields and standards such as age-normalized anthropometrics (31). The Enclave itself uses the OMOP model, and all contributing institutions pre-process their data so it can be harmonized OMOP. When stratifying the patient data by the data model used by their healthcare institution, we found that data about patient race from TriNetX had the best conformance; 86.2% of the TriNetX data conformed to one of the five main categories. Only 66.2% of the race data from OMOP PEDSnet, on the other hand, achieved conformance. We found that, depending on the data model, the conformance of data about ethnicity varied more widely than the race data; while 93.1% of data from healthcare institutions using PCORnet’s data model conforms to the standard for reporting patient ethnicity, only 50.8% of ethnicity data from institutions using the ACT model were adherent to the standard.

Comparison to other data sources. Figure 3 shows how the distribution of race and ethnicity data in

Figure 3. Comparison of race and ethnicity across various datasets



N3C compares to the US overall (ACS) and one other EHR-based data repository, Cerner Health Facts (CHF). N3C, similar to Cerner Health Facts, has fewer Hispanic or Latino and Asian patients than ACS, but comparable rates for other groups. This is likely related to both the types of institutions that contribute data (and the patients they serve) as well as the large amount of missing or non-conforming data in both datasets.

3.5 Discussion

Our analyses of the N3C Enclave revealed a number of facts that are important for researchers to consider when drawing conclusions based on these data, or data from similar large-scale data resources. First, a substantial portion of the records (20.7%) were in some way non-conforming, and 11.7% of all records were missing race and/or ethnicity data (missing data was considered a subcategory of non-conformance in this study). Second, there are opportunities to improve the quality of these data at various points in the data pipeline, not only when it is collected at the point of care. Finally, the impact of these data quality issues was not equal across all races and ethnicities. The magnitude and type of non-conformance varied across race and ethnicity, with patients of color and vulnerable communities over-represented in the misclassified data and the non-conforming data. The fact that the data was not randomly non-conforming means there is a potential to introduce bias in analyses and conclusions drawn from these data.

By delving into the data ingestion process, we were able to locate points where such errors might be introduced. For example, standard scripts are used to prepare data for harmonization, and

sometimes these scripts do not anticipate the use of deprecated codes that may have been used at the point of care to record race or ethnicity. When data preparation is automated using such scripts, patients who have been assigned deprecated codes may be harmonized to “No matching category.” Misclassification of patients might also occur if multiple values were entered into a single field, perhaps because patients were allowed to select more than one race, or because race and ethnicity were both recorded in the same column. We hope to utilize the results of this analysis to add additional coding to these scripts to prevent these errors, and to identify ways to correct the race and ethnicity data post-ingestion.

‘Hispanic or Latino,’ race, and ethnicity. Our finding that Hispanic or Latino patients are overrepresented in the non-conforming race data reflects that the distinction between the terms “race” and “ethnicity” continues to be a source of variability and confusion. Although the standard recommends that race and ethnicity be captured separately, some healthcare systems combine them and offer “Hispanic or Latino” as a possible selection under Race. The current PCORnet CDM specification recommends mapping data from patients whose race has been recorded as Hispanic or Latino to “Other,” which explains our finding that 61% of the data harmonized to the “Other” category comes from Hispanic or Latino patients.(32) However, Hispanic or Latino patients were also overrepresented in other categories of non-conforming data, such as “Refused” (21.1% Hispanic or Latino) and “Multiracial” (47.4% Hispanic or Latino). During the 2010 Census, the US Census Bureau tested a combined race-ethnicity question and found that including Hispanic origin as a racial category dramatically reduced both the item nonresponse rate and the selection of “some other race.” Their report noted a “lack of agreement on the terms race and ethnicity, and errors in reporting resulting from misinterpretations.”(33)

Obscured data from American Indian or Alaska Native patients. The lack of accessible data on American Indian or Alaska Native populations is a significant problem of the dataset. The Urban Indian Health Institute has stated that, “current standard data collection practices by many federal, state, and local entities effectively omit or misclassify American Indian or Alaska Native

populations, both urban and rural. This is particularly concerning in the midst of the COVID-19 pandemic as these current standards of practice are resulting in a gross undercount of the impact COVID-19 has on Native people.”(34) A number of federal laws, treaties, and executive orders have established the sovereignty of Tribes and Tribal Nations over their data and the power to regulate research, although the gap between *recognition* of those rights, and the *assertion* of those rights remains wide.(35-37) A legacy of harm, medical maltreatment, and research misconduct has engendered mistrust between the Tribal and clinical research communities.(38) To begin to address these issues, Tribal leaders, scholars and advocates have established protocols and institutions to ensure human protections for research involving the American Indian or Alaska Native community. In 2010, the US Department of Health and Human Services established a formal Tribal Consultation Policy to create a mechanism for collaboration at the federal level.(39)

In September of 2020, the NIH released the results of the Tribal Consultation they had initiated in 2019 in order to develop a policy on the management and sharing of biomedical data.(38) In December of 2021, NCATS Director Joni Rutter formally initiated a Tribal Consultation about the N3C Data Enclave, writing “Ideally, NIH would have sought Tribal Consultation before the start of this program. However, given other COVID-related Consultations and urgency of the pandemic, NCATS decided to obscure AI/AN data until Consultation could occur.”(40) The N3C Tribal Consultation took place on February 11th, 2022, and as of this writing the testimony of the Tribal Leadership is being collected. NCATS expects to implement their recommendations by summer 2022.

Value of granular data. A small number of data partners in the N3C enclave have submitted granular data about racial and ethnic subgroups. Currently, the healthcare system understands little about the complex and diverse health needs of these subgroups,(41) and making these granular data more widely available to researchers within white dominant institutions could help alleviate these gaps in knowledge. The adoption and consistent use of updated data standards such as those recommended by the Department of Health and Human Services Office of Minority Health (OMH) might help improve the availability of data about subgroups.(42)

Implications for COVID-19 research. Compared with the CDC data on cases, the N3C system has significantly less missingness at 11% compared to 24% in the CDC COVID-19 cases dataset. Seventy-nine percent of the data conforms to CDC standard racial categories (higher than the 64% in the CHF dataset), making this repository useful for COVID-19 health disparities research. Moreover, because the system gives access to race and ethnicity source values, we were able to assess misclassification and can update the racial and ethnic categorization of patients for research purposes. Though granular data represents a small portion of the overall data, it can be used for small-scale projects analyzing differences within racial categories. Given that this dataset has representation from all regions in the U.S.,⁽⁸⁾ it can be used to validate against the CDC COVID positivity rates by race.

Limitations. Our analysis is not intended as an endorsement of the standard developed by OMB and implemented by federal agencies, but rather a description of real world, contemporary practices by healthcare institutions around the US. **Our perspective is that race and ethnicity are not biological categories, but rather social constructs that are highly context-dependent and tied to existing power dynamics. As variables in clinical research, race and ethnicity function as indirect and highly imperfect proxies for complex and systemic factors (e.g. racism, colonialism, socioeconomic barriers to healthcare delivery systems) that drive and perpetuate inequities.** The OMB itself stresses this point, stating that, “the racial and ethnic categories set forth in the standard should not be interpreted as being scientific or anthropological in nature.”⁽⁴³⁾ The US Census Bureau has written that racial categories “generally reflect a social definition of race recognized in this country and not an attempt to define race biologically, anthropologically, or genetically.”⁽⁴⁴⁾

Finally, it should be noted that data partners, as the contributing healthcare institutions are referred to in the N3C, were provided with anonymity as a consideration for contributing data. This means that the provenance of these data is limited, so we do not know how they were initially collected. Finally, because most of the contributing healthcare institutions are recipients of a Center for

Translational Science Award with an established relationship with the National Center for Advancing Translational Sciences, it is likely that academic medical centers are overrepresented.

3.6 Conclusion

Twenty-eight years after Congress mandated the inclusion of racial and ethnic minority groups in federally funded clinical research with the NIH Revitalization Act (45), the ongoing lack of racially and ethnically diverse cohorts remains a challenge to improving equity in research and healthcare (46). Because the COVID-19 pandemic disproportionately impacts communities of color by exacerbating existing health inequities, the accurate identification of these cohorts within N3C is crucial to identifying, understanding, and ultimately addressing these disparities. Data problems arise for many reasons, but primary among them is the discrepancy between how institutions conceptualize race and ethnicity, and the far more varied ways people identify themselves.(47) The complex history of racial identification in the United States has resulted in shifting concepts of race and ethnicity.(48) Self-identified race and ethnicity are often dependent on physical attributes that, although heritable, correlate poorly with genetic similarity or ancestry. Nevertheless, race and ethnicity are well-established predictors of health outcomes and access to care. However, a multitude of factors that are both correlated with and are independent of race and ethnicity may affect group differences in health and health care. Race and ethnicity are only one of many elements considered to be social determinants of health (SDoH) - non-medical factors that influence health outcomes and known to have a significant relationship to these disparities.(49,50) Teasing out which factors influence health outcomes is challenging,(51) and issues with data quality and inappropriate or poorly applied standards around race and ethnicity can greatly lessen our understanding of health disparities.

Though there are some limitations to the racial representation in this dataset, it nevertheless remains a unique resource for COVID-19 research on racial disparities. COVID-19 has served to emphasize the deadliness of these disparities and has made social conditions far worse for many Black, Hispanic, and American Indian persons living in the US. But these inequities are not immutable.

The COVID-19 pandemic provides an opportunity for clinicians, health systems, scientists, and policy makers to address social disparities, and thereby improve the health and well-being of all persons in the US for both known and future illnesses.

Databases such as N3C spur discovery by collecting and centralizing clinical data, making national, federated datasets available to researchers. Although intended to increase the accessibility of data, governance can paradoxically create further restrictions. In addition, these efforts require that data be transformed numerous times, and differences in how race and ethnicity is conceptualized by healthcare institutions affects the quality of the harmonized data. Transparency about these numerous decisions is critical if researchers are to make accurate inferences from analyses.

Acknowledgements and Funding. The analyses described here were conducted with data and tools accessed through the National Center for Advancing Translational Science (NCATS) N3C Data Enclave (<https://covid.cd2h.org>) and supported by NCATS U24TR002306. This research was possible because of the patients whose information is included within the data and the organizations and scientists who have contributed to the ongoing development of this community resource. Lily A. Cook has received support from the National Library of Medicine under Award Number T15LM007088. Nicole G. Weiskopf has received funding under National Library of Medicine R21LM013645 and U24TR002306 from the National Center for Advancing Translational Science (NCATS) of the U.S. National Institutes of Health. Juan Espinoza has received funding from grants UL1TR001855 and UL1TR000130 from the National Center for Advancing Translational Science (NCATS) of the U.S. National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to acknowledge the assistance of Siegfried Gold, the N3C Data Liaison to the Social Determinants of Health domain team, and Adam Wilcox, who provided encouragement and guidance on the initial structure and goals of this work.

3.7 References

1. Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. COVID-19 Dashboard. [Internet]. [cited 2021 Aug 7]. Available from: <https://coronavirus.jhu.edu/map.html>
2. Mude W, Oguoma VM, Nyanhanda T, Mwanri L, Njue C. Racial disparities in COVID-19 pandemic cases, hospitalisations, and deaths: A systematic review and meta-analysis. *J Glob Health*. 2021 Jun 26;11:05015.
3. Gold JAW, Rossen LM, Ahmad FB, Sutton P, Li Z, Salvatore PP, et al. Race, Ethnicity, and Age Trends in Persons Who Died from COVID-19 - United States, May-August 2020. *MMWR Morb Mortal Wkly Rep*. 2020 Oct 23;69(42):1517–21.
4. Wortham JM, Lee JT, Althomsons S, Latash J, Davidson A, Guerra K, et al. Characteristics of Persons Who Died with COVID-19 - United States, February 12-May 18, 2020. *MMWR Morb Mortal Wkly Rep*. 2020 Jul 17;69(28):923–9.
5. Hatcher SM, Agnew-Brune C, Anderson M, Zambrano LD, Rose CE, Jim MA, et al. COVID-19 Among American Indian and Alaska Native Persons - 23 States, January 31-July 3, 2020. *MMWR Morb Mortal Wkly Rep*. 2020 Aug 28;69(34):1166–9.
6. Dagliati A, Malovini A, Tibollo V, Bellazzi R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. *Brief Bioinform*. 2021 Mar 22;22(2):812–22.
7. Ferguson C. It took a pandemic, but the US finally has (some) centralized medical data | MIT Technology Review. *MIT Technology Review* [Internet]. 2021 Jun 21 [cited 2021 Aug 20]; Available from: <https://www.technologyreview.com/2021/06/21/1026590/us-covid-database-n3c-nih-privacy/>
8. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc*. 2021 Mar 1;28(3):427–43.
9. National Institutes of Health (NIH), National Center for Advancing Translational Sciences (NCATS). National COVID Cohort Collaborative Data Enclave Repository. [Internet]. Bethesda, Maryland: U.S. Department of Health and Human Services, National Institutes of Health; 2021. Available from: <https://covid.cd2h.org/>
10. Jirkovsky V, Obitko M. Semantic Heterogeneity Reduction for Big Data in Industrial Automation. *CEUR Workshop Proceedings*. 2014;1214.
11. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, et al. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc*. 2019 Aug 1;26(8–9):730–6.
12. Grafova IB, Jarrin OF. Beyond Black and White: Mapping Misclassification of Medicare Beneficiaries Race and Ethnicity. *Med Care Res Rev*. 2020 Jul 7;1077558720935733.
13. Jarrin OF, Nyandeghe AN, Grafova IB, Dong X, Lin H. Validity of Race and Ethnicity Codes in Medicare Administrative Data Compared With Gold-standard Self-reported Race Collected During Routine Home Health Care Visits. *Med Care*. 2020 Jan;58(1):e1–8.
14. Karen L Pellegrin, Jill B Miyamura, Carolyn Ma, Ronald Taniguchi. Improving Accuracy and Relevance of Race/Ethnicity Data: Results of a Statewide Collaboration in Hawaii. *J Healthc Qual*. 2016;38(5):314–21.
15. Ford C, Harawa NT. A NEW CONCEPTUALIZATION OF ETHNICITY FOR SOCIAL EPIDEMIOLOGIC AND HEALTH EQUITY RESEARCH. *Soc Sci Med* 1982. 2010 Jul;71(2):251–8.
16. Griffith DM, Moy E, Reischl TM, Dayton E. National data for monitoring and evaluating racial and ethnic health inequities: where do we go from here? *Health Educ Behav Off Publ Soc Public Health Educ*. 2006 Aug;33(4):470–87.
17. Mays VM, Ponce NA, Washington DL, Cochran SD. Classification of race and ethnicity: implications for public health. *Annu Rev Public Health*. 2003 Jan 1;24:83–110.
18. Williams DR, Lavizzo-Mourey R, Warren RC. The concept of race and health status in America. *Public Health Rep Wash DC*. 1994 Jan 1;109(1):26–41.
19. Office of Management and Budget. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. *Fed Regist*. 1997 Oct 30; Available from:

20. Hong SS. N3C Data Ingestion and Harmonization Workstream Race Mapping Crosswalk [Internet]. National COVID Cohort Collaborative; 2020 [cited 2021 Aug 23]. Available from: https://github.com/National-COVID-Cohort-Collaborative/Data-Ingestion-and-Harmonization/blob/1f48cb945b186292aec07ce7e2751796967a3a83/ETLProcess/scripts/n3c_xwalk_mapping.sql
21. Institute of Medicine Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: National Academies Press; 2003.
22. Center for Disease Control and Prevention. Race and Ethnicity Code Set Version 1.0. 2000 [cited 2020 Jan 20]; Available from: <https://www.cdc.gov/phinf/resources/vocabulary/documents/CDC-Race-Ethnicity-Background-and-Purpose.pdf>
23. HL7 International. Structure Definition: US Core Race Extension [Internet]. [cited 2021 Aug 23]. Available from: <https://www.hl7.org/fhir/us/core/StructureDefinition-us-core-race.html>
24. Ulmer C, McFadden B, Nerenz DR, Institute of Medicine (U.S.), editors. Race, ethnicity, and language data: standardization for health care quality improvement. Washington, D.C: National Academies Press; 2009. 181 p.
25. Klinger EV, Carlini SV, Gonzalez I, Hubert SS, Linder JA, Rigotti NA, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. J Gen Intern Med. 2015 Jun;30(6):719–23.
26. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. Journal of the American Medical Informatics Association. 2022 Jan 1;29(1):187–96. Available from: <https://doi.org/10.1093/jamia/ocab199>
27. Madlock-Brown C, Reynolds RB. Identifying obesity-related multimorbidity combinations in the United States. Clin Obes. 2019 Dec;9(6):e12336.
28. CDC. Demographic Trends of COVID-19 cases and deaths in the US reported to CDC [Internet]. COVID Data Tracker. 2021 [cited 2021 Aug 26]. Available from: <https://covid.cdc.gov/covid-data-tracker>
29. National Center for Health Statistics. Health Disparities - COVID-19 Provisional Counts [Internet]. Centers for Disease Control and Prevention. 2021 [cited 2021 Aug 26]. Available from: https://www.cdc.gov/nchs/nvss/vsrr/covid19/health_disparities.htm
30. National Center for Advancing Translational Sciences. NIH Tribal Consultation on the National COVID Cohort Collaborative (N3C) Framing Letter [Internet]. Undated. Available from: <https://ncats.nih.gov/files/Framing-Letter-N3C-508.pdf>
31. PEDSnet. PEDSnet Common Data Model [Internet]. [cited 2021 Aug 23]. Available from: <https://pedsnet.org/data/common-data-model/>
32. The National Patient-Centered Clinical Research Network. PCORnet Common Data Model (CDM) Specification, Version 6. 2020. Available from: https://pcornet.org/wp-content/uploads/2022/01/PCORnet-Common-Data-Model-v60-2020_10_221.pdf
33. Compton E, Bentley M, Ennis S, Rastogi S. 2010 Census Race and Hispanic Origin Alternative Questionnaire Experiment. 2013.
34. Osman I, Urban Indian Health Institute. Best Practices for American Indian and Alaska Native Data Collection. Seattle, WA: Seattle Indian Health Board; 2020. Available from: <https://aiapi.asu.edu/sites/default/files/best-practices-for-american-indian-and-alaska-native-data-collection.pdf>
35. Tsosie R. Tribal Data Governance and Informational Privacy: Constructing “Indigenous Data Sovereignty.” Mont Law Rev. 2019 Aug 1;80(2):229–68.
36. Hull SC, Wilson (Diné) DR. Beyond Belmont: Ensuring Respect for AI/AN Communities Through Tribal IRBs, Laws, and Policies. Am J Bioeth. 2017 Jul 3;17(7):60–2.
37. Garrison NA, Barton KS, Porter KM, Mai T, Burke W, Carroll SR. Access and Management: Indigenous Perspectives on Genomic Data Sharing. Ethn Dis. 29(Suppl 3):659–68.
38. National Institutes of Health. NIH Tribal Consultation Report: NIH Draft Policy for Data Management and Sharing. 2020 Sept. Available from: https://osp.od.nih.gov/wp-content/uploads/Tribal_Report_Final_508.pdf

39. Sibelius K. U. S DEPARTMENT OF HEALTH AND HUMAN SERVICES TRIBAL CONSULTATION POLICY. 2010.
40. Rutter J, National Center for Advancing Translational Sciences. Dear Tribal Leader Letter. December 20th, 2021. Available from: <https://ncats.nih.gov/files/Dear-Tribal-Leader-Letter-N3C-508.pdf>
41. Eisler RM, Hersen M, editors. Handbook of Gender, Culture, and Health. 1st ed. Psychology Press; 2012.
42. U.S. Department of Health and Human Services. Implementation guidance on data collection standards for race, ethnicity, sex, primary language, and disability status. Office of the Assistant Secretary for Planning and Evaluation; 2011.
43. Office of Management and Budget. Review and Possible Limited Revision of OMB's Statistical Policy Directive on Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. Fed Regist [Internet]. 2016 Sep 30 [cited 2021 Aug 26];81(190). Available from: <https://www.govinfo.gov/content/pkg/FR-2016-09-30/pdf/2016-23672.pdf>
44. US Census Bureau. About Race [Internet]. United States Census Bureau. 2020 [cited 2021 Aug 26]. Available from: <https://www.census.gov/topics/population/race/about.html>
45. NIH Revitalization Act. [Internet]. Sect. Section 289a-2, PL 103-43 Jun 10, 1993 p. 630–1. Available from: <https://www.govinfo.gov/content/pkg/USCODE-2011-title42/pdf/USCODE-2011-title42-chap6A-subchapIII-partH-sec289a-2.pdf>
46. Flores LE, Frontera WR, Andrasik MP, del Rio C, Mondríguez-González A, Price SA, et al. Assessment of the Inclusion of Racial/Ethnic Minority, Female, and Older Individuals in Vaccine Clinical Trials. JAMA Netw Open. 2021 Feb 19;4(2):e2037640.
47. Landale NS, Oropesa RS. White, Black, or Puerto Rican? Racial Self-Identification among Mainland and Island Puerto Ricans. Soc Forces. 2002 Sep 1;81(1):231–54.
48. Hahn RA. The State of Federal Health Statistics on Racial and Ethnic Groups. JAMA. 1992 Jan 8;267(2):268–71.
49. World Health Organization (WHO). Social determinants of health [Internet]. [cited 2021 Aug 23]. Available from: <https://www.who.int/health-topics/social-determinants-of-health>
50. Social Determinants of Health - Healthy People 2030 | health.gov [Internet]. [cited 2021 Aug 23]. Available from: <https://health.gov/healthypeople/objectives-and-data/social-determinants-health>
51. Cottrell EK, Hendricks M, Dambrun K, Cowburn S, Pantell M, Gold R, et al. Comparison of Community-Level and Patient-Level Social Risk Data in a Network of Community Health Centers. JAMA Netw Open. 2020;3(10):e2016852–e2016852.

Chapter 4. The care and feeding of integrated data: the impact of data preparation methods on the results of observational studies

Abstract

Objective. This research aims to: (a) demonstrate that choices made while preparing real-world data have a substantial impact on results, and (b) identify which steps in the data preparation pipeline are the most impactful.

Methods. We utilized a two-stage approach. First, we designed a cohort study for which we created an integrated dataset containing both individual and community-level variables. That dataset was created in order to estimate the association between a community-level exposure (poor air quality days caused by wildfire smoke) and an individual-level health outcome documented in the clinical record (pulmonary exacerbations in people with cystic fibrosis). Second, we examined how varying the choices made during the data preparation process impact the results of such a study. To show this, we developed an algorithm that prepared the data a variety of different ways and then generated results from every combination of decisions that could have been made at each step in the data preparation process. The algorithm relied on Fisher's exact test to generate the resulting point estimate odds ratios. We then characterized the impact of each decision on the result by using Hedges' g to calculate the standardized mean difference.

Results. The algorithm ran through a total of 135 different possible data preparation pathways, generating 93 distinct odds ratios from the same baseline dataset and statistical test. The resulting odds ratio estimates varied between 0.83 (CI = 0.53 – 1.29, $p = 0.40$) to 2.93 (CI = 1.83– 4.70, $p < 0.001$), with a mean OR of 1.31 (SD +/-0.37). 50.37% ($n = 68$) of the odds ratios had a p -value ≤ 0.05 . The effect size of the data preparation decisions varied between $g = 0.01$ and $g = 0.65$. Different data preparation techniques removed up to 87.23% ($n = 519$) of the patients from the baseline dataset, and up to 93.51% ($n = 565,067$) of the patient days. Characteristics of the patient cohort, such as age and the percentage of patients who lived in urban areas, also differed depending on data preparation methods.

Conclusion. Because data preparation methods have a significant impact on the results of the analysis, researchers should provide documentation of and justification for these methods. These results suggest that making information about data preparation decisions available may enhance reproducibility.

4.1 Introduction

Although EHR data has been used increasingly often as a source for observational research, EHRs do not capture many important health risks; critical factors such as socio-economic status and exposure to environmental pollutants are either sparsely collected in the patient health record or are simply not collected at all. As a result, researchers often choose to integrate individual medical records with data from external, non-clinical sources, usually data from the community or population level. Linking, or integrating, data this way allows for the incorporation of novel variables into clinical research,⁽¹⁾ deepening our understanding of the many factors that affect our health and aiding in the identification of disparities that create health inequity. These non-medical factors are often referred to as *social determinants of health*.⁽²⁾

While linking data can add a great deal of richness to the dataset by introducing dozens - if not hundreds - of novel variables, it also has the potential to multiply data quality problems. These integrated datasets can compound data quality problems that plague real-world data by introducing an additional potential source of error: the linkage process itself.⁽³⁻⁵⁾ Error and bias can be introduced as a result of faulty data from the patient record, faulty community-level data, matching the patient to the wrong community, or possibly all three. Because patient addresses are typically used to link their health information to information about their community, social determinants of health research is unusually dependent on patient address data. However, address data can be a shaky linkage to rely upon when conducting data integration. Because these linkages are imperfect, there is inevitably some data either lost or incorrectly linked during the process of integrating two datasets. As researchers, it is our responsibility to ask which stages of the research process are most vulnerable to the introduction

of error.

Fortunately, thoughtful data preparation can help prevent error, improving the validity and usability of data and allow more data to be retained for analysis. As data users inevitably encounter data quality issues— whether this is missing or inaccurate data - they make choices about how to handle these issues. These choices are most frequently made during the data preparation process, particularly during data cleaning. Data preparation, also referred to as data munging, data wrangling, or preprocessing, is intended to address data quality issues prior to analysis, thus reducing error and bias. The data preparation process entails steps such as cleaning, transforming, and integrating the data.

Currently, most researchers utilize an ad hoc, a priori, and intuition-based approach to data preparation. This lack of a scientific, formalized approach may have repercussions to research outcomes, however.(6) For example, a 2015 experiment that appeared in the journal *Nature* gave the same dataset to 29 different research teams for analysis and interpretation – and yielded 29 different results across a variety of different statistical methods and data preparation practices.(7)

The objective of our research is to demonstrate the effect of only one of these factors – data preparation. We isolated the effect of data preparation choices on the results of the analysis by keeping the dataset, research team, and statistical methods constant. By varying only the choices made while cleaning and preparing the data, this research aims to: (a) demonstrate that choices made while preparing secondary use data have a substantial impact on results, and (b) identify which steps in data pipeline are the most impactful. To the best of our knowledge, this research is the first to demonstrate that integrated datasets are vulnerable to data preparation methods, and to attempt to quantify exactly which steps of the data preparation process have the most influence on the result.

4.2 Background

Social determinants of health and data integration. There are numerous definitions of social determinants of health. The Center for Disease Control and Prevention (CDC) relies upon what they

call a “place-based framework” for their definition, stating “social determinants of health are conditions in the places where people live, learn, work, and play that affect a wide range of health risks and outcomes.”(2) They add that most health inequity can be traced back to these social determinants factors.

In 2014, the Institute of Medicine published a detailed report intended to provide guidance about social determinants of health data – which concepts are the most clinically relevant, which should be collected in the EHR, and how best to represent them.(8) Although some of the core social determinants domains identified by the Institute of Medicine report are collected at the point of care, others are challenging to collect at the patient level. To obtain these less commonly collected social determinants, the Institute of Medicine recommended drawing upon what is known about a patient’s “neighborhood and community composition,” writing:

“A number of domains of special relevance to the social determinants of health can be characterized by use of the patient’s residential address. The exponential growth of geocoded datasets will likely allow linkage to a large set of potentially useful variables in coming years.”

In other words, some information relevant to health - such as income - is rarely collected in the medical record but we can make an educated guess about that information by looking at the neighborhood in which the patient lives. This is typically accomplished by linking the patient’s address with community-level information about their neighborhood, a process called *geocoding*. The Institute of Medicine report notes, “an important point is that these geographically linked data can capture health-relevant information that cannot be obtained directly from the patient.”(8)

Data integration and data quality. The ability to link community-level data to health data therefore hinges on patient addresses, and thus its accuracy is largely dependent on the success of the geocoding process, i.e., whether a patient’s address can be matched to the correct neighborhood information. However, research projects linking patient health records to community-level information are often unable to match a significant portion of their patient records, thus excluding a large number

of patients from the research cohort and introducing selection bias.(9) For example, researchers working on the Pediatric Big Data Project geocoded their database in order to link data from one million children with “800 physical, built, and social environmental datasets.”(10) Of the approximately one million patients in the database, only 727,555 (73%) had valid addresses; among those with valid addresses, 127,502 (18%) were inexplicably recorded as having lived at the address for zero days. What effect does the exclusion of 27% of the original patient cohort have on that research? What effect does the inclusion of the 18% of health records with questionable address data have?

Results from research conducted using geocoded patient data from the Duke University Health System give us some clues about how these issues may affect research results. In 2012, Duke initiated a massive geocoding project, aiming to link the patient records they had stored in their data warehouse with community-level data from the Census and other sources.(11) To accomplish this, they built a Geographic Health Information System by geocoding the patient addresses in their enterprise data warehouse and linking individual patient records to publicly available community level data, adding 10,000 new demographic and socioeconomic variables to their data warehouse.(12) However, one of the first research projects to utilize geocoded patient records came to a conclusion that seemed to contradict several decades of research about the relationship between wealth and health: that community-level data about socio-economic status adds little to the EHR in terms of predicting health outcomes.(13)

In order to understand this seemingly contradictory result it may help to contextualize it. We know little about how this particular research project handled any missing data. However, in 2014 the group tasked with implementing the geocoding of Duke’s research warehouse stated that they were able to geocode only 72% of the patient data,(11) meaning that 28% of it could not be matched to community-level data and presumably would not be available for analysis. We also know little about how this project classified exposure for this study – in other words, how long someone needed to live at an address in order to be considered of a certain socio-economic status - or about the numerous other data preparation decisions that that led to that result. Is it possible that the choices made about

this data influenced the result?

4.3 Methods

To answer this question, we designed a study utilizing a two-stage approach. During the first stage, we created an integrated dataset containing both community and individual-level variables. This dataset was intended to estimate the association between a community-level exposure and the individual-level health outcome documented in the clinical record. For the second stage we examined the impact that varying the methods used to prepare that data might have on that association.

Stage one: cohort study. First, a dataset was built by linking patient records with air quality data. Dr. Aaron Trimble obtained the clinical data by querying the Cystic Fibrosis Foundation Patient Registry (14) and pulling the data of patients treated at Oregon Health & Science University's Cystic Fibrosis Care Center and Doernbecher Children's Hospital between 2010 to 2019 (inclusive). All patients consented to be included in the CF Foundation's Patient Registry. The outcome of interest was severe pulmonary exacerbation, which was defined as a clinical event with respiratory symptoms requiring medical intervention – either home intravenous antibiotics or hospitalization. Community-level data about fine particulate matter (PM_{2.5}) was obtained from the EPA's Air Quality System DataMart.(15)

Based on their residential zip code centroid, individuals were assigned the measurements from the air quality monitor located nearest their home.(16) Those residing in areas with a daily Air Quality Index (AQI) beyond the limit set by the EPA as 'unhealthy for sensitive groups,' (AQI > 100) were considered exposed. Using patient-days as the increment of analysis, we examined whether the likelihood of experiencing an exacerbation was different for patients who had been exposed to wildfire smoke-driven air pollution within the designated lag time. We tabulated the number of wildfire-exposed patient-days and unexposed patient-days and determined whether each patient-day contained the start of an episode of care for an exacerbation. Both exacerbations and exposure status were treated as binary variables, and a contingency table was created to determine summary statistics.(17) The

association between wildfire smoke exposure and pulmonary exacerbations was examined using Fisher's exact test to generate point estimate odds ratios.

Stage two: the effect of data preparation decisions. The second stage examined how the choices made while preparing data for analysis could impact the results of the cohort study conducted in stage one. To do this, we developed an algorithm that prepared the same dataset a variety of different ways. The algorithm was created by examining the data preparation process used in the first stage, and then splitting the decisions made into four integral steps:

- (1) decisions about validating addresses,
- (2) decisions made about handling missing addresses,
- (3) decisions about classifying exposures, and
- (4) decisions made about lag times.

We then identified several choices that could plausibly be made at each step and treated each possibility as a node in a directed network. We used a deductive approach to select the possible choices available at each step, using our prior work - a systematic review of data quality problems in social determinants of health research – as a foundation.⁽¹⁸⁾ Based on a review of 76 of peer-reviewed articles, that work derived six recommendations for conducting observational research with social determinants data:

1. Avoid complete case analysis
2. Impute data
3. Rely on multiple sources
4. Use validated software tools
5. Select addresses thoughtfully
6. Be transparent

We developed some of the options based on these recommendations. We also searched peer-reviewed literature for specific methods that would be commonly utilized in such a study – methods for addressing missing data in health research, methods utilizing geocoding for researching health

outcomes, and methods used in environmental health studies on either wildfire or and cystic fibrosis. We selected plausible options with the assistance of a clinician familiar with cystic fibrosis, Dr. Aaron Trimble. The options available at each step are as follows (see Table 1):

Table 3: Data preparation steps and the options available

DATA PREPARATION PROCESS						
Step:	1. Validate addresses	2. Handle missing addresses	INTEGRATE DATA	3. Classify exposures	4. Decide lag times	ANALYSIS
Possible choices:	1a. Validate zip codes using USPS; set invalid zip codes to missing 1b. Ignore invalid zip codes (drops invalid zips during data integration step) 1c. Validate zips using number of digits; set invalid zip codes to missing	2a. Drop rows missing address data (complete case analysis) 2b. Interpolate first, then drop missing (last observation carried forward) 2c. Triangulate first (i.e., try to find address elsewhere), then drop missing 2d. Impute all missing addresses (k-nearest neighbors method) 2e. Use auxiliary data for imputation - triangulate first, then impute w/knn	<i>This is the point in the data preparation process where the clinical dataset is integrated (or "matched") with the community-level dataset.</i>	3a. Drop data from patients living > 3 miles from the closest air quality monitor 3b. Drop data from patients living > 30 miles from the closest air quality monitor 3c. Use all data, regardless of how far patient lives from closest air quality monitor	4a. Consider outcomes associated if they occur within ONE week after exposure 4b. Consider outcomes associated if they occur within TWO weeks after exposure 4c. Consider outcomes associated if they occur within FOUR weeks after exposure	<i>After all data preparation decisions have been made and conducted, the prepared dataset is ready for analysis.</i>
Purpose:	Linking (or integrating) data in order to determine exposure			Determining validity of exposure		Associating exposure and outcome

Step 1: Methods for validating addresses (three possibilities). Due to data entry mistakes, incorrect values are occasionally entered into the zip code field of the patient. We compared three different methods for identifying and handling invalid address data (in this case, zip codes).

Option A: Validate patient zip codes using a list of residential zip codes from the US Postal Service (USPS). All zip codes not on the USPS list were set to missing.

Option B: Do not validate addresses. It is important to note that choosing not to validate addresses is equivalent to dropping rows of data containing invalid zip codes, because only valid zip codes will match up to an air quality monitor during the integration process.

Option C: Validate zip codes by checking whether they have five digits. Zip codes with something other than five digits were then set to missing.

Step 2: Methods for handling missing addresses (five possibilities).

Option A: Drop rows missing address data (AKA listwise or casewise deletion). Although casewise deletion is the most commonly used method for handling missing data, it can be problematic because it assumes that information is missing completely at random.⁽¹⁹⁾ Studies have shown that filtering electronic health records for patients with complete data carries with it several biases.^(20, 21)

Option B: Interpolate first, then drop missing addresses. For patients whose addresses were missing only for some years, their missing addresses were interpolated using the “last observation carried forward” method – i.e., it was assumed that they continued to reside at the same address. This single imputation method was selected because it is commonly used for missing data during clinical trials. Like casewise deletion (Option A, above), last observation carried forward assumes that information is missing completely at random; however, it is less likely to. Because this method makes guesses about a patient’s location using data from that same patient, we were only able to interpolate zip codes for patients who had a confirmed address at some point in their record, so data from patients who did not have a known address to draw upon for interpolation were therefore dropped.

Option C: Triangulate first, then drop missing addresses. This method looked in other data pulls to try to find the patient’s address elsewhere. Rather than abandoning or deleting earlier data pulls, this method combined all versions of the dataset, linked patient data by ID number and the date, then filled in the data horizontally (across the rows). We refer to this method as “triangulation,” because it fills in missing data about a patient from other places across that patient’s own record at that same time point. As with the interpolation method described above, rows of data were dropped from patients whose address data was incomplete at all pulls.

Option D: Impute all missing addresses using bootstrap resampling. The zip codes here were imputed by creating a list of zip codes found in the dataset and filling in missing values by randomly selecting from this list with replacement.

Option E: Triangulate, then impute. This first uses the triangulation method described in 2c but, rather than simply dropping the missing rows, adds the bootstrap resampling technique described

above to impute the rows of data that remain incomplete triangulation. This method is the most “inclusive” of the imputation methods we tested, in the sense that it attempts to include as much auxiliary data as possible. Results from simulated models have suggested that inclusive imputation methods are preferable to restrictive methods,(22) but to the best of our knowledge this is the first time this strategy has been tested using real-world data.

After these first two steps, the algorithm then integrated the individual-level clinical data with the community-level data by merging the two datasets. The next step makes decisions about classifying exposures.

Step 3: Classify exposures (three possibilities). The cohort study conducted in stage one estimated exposure based on air quality monitor readings closest to each patient, also known as the proxy method. However, approximating exposure this way can result in misclassification.(23) The most common way to avoid misclassifying participants when using the proxy method is to assume that the air quality monitor readings are only relevant for within a certain radius – in other words, some patients live so far away from the exposure that they could no longer reasonably be consider exposed.(24)

Option A: Drop rows containing data from patients whose residential zip code centroid was more than 3 miles from the zip code centroid of the nearest air quality monitor.(25)

Option B: Drop data from patients who live more than 30 miles from the nearest air quality monitor. (26)

Option C: Use data from all patients regardless of how far away they live from the nearest air quality monitor.

Step 4: Decide lag times (three possibilities). Finally, we need to consider how long it may take after an exposure for a patient to develop the outcome of interest. We selected the following potential intervals after consulting a clinician familiar with cystic fibrosis, Dr. Aaron Trimble, and examining the literature about wildfire smoke exposures.

Option A: Use a rolling one-week window. Assumes that outcomes experienced within 7 days following an exposure are associated with that exposure, and that an outcome experienced outside

that window is unassociated.

Option B. Use a rolling two-week window. As above but uses 14 days rather than 7.

Option C: Use a rolling four-week window. As above but uses 28 days.

Next, we used the algorithm to generate the results from every possible pathway through the data preparation process, where each pathway was a different combination of options that could be selected at each step (see Fig. 1). We then compared the results by grouping the results from each step together, taking the mean of the odds ratios that resulted from the transformation conducted at that step. Finally, we characterized the effect size of each decision by calculating the standardized mean difference using Hedges' g :(27-29)

$$g = \frac{\mu_{i1} - \mu_{i0}}{\sigma_i}.$$

Here, σ_i is the pooled sample standard deviation,

$$S_{iP} = \sqrt{\frac{(n_{i0} - 1) s_{i0}^2 + (n_{i1} - 1) s_{i1}^2}{n_{i0} + n_{i1} - 2}},$$

while μ_{i0} , s_{i0}^2 , and n_{i0} are the sample mean, sample variance, and number of samples for the control group, and μ_{i1} , s_{i1}^2 , and n_{i1} are the sample mean, sample variance, and number of samples for the treatment group. The grouped odds ratios for each step were considered the “treatment” group, while the pooled means odds ratio for all results was used as the “control.”

All data cleaning, validation, statistical analysis, and figures were documented in a Python notebook and conducted in Python 3.7. The code for the algorithm is available in the supplementary material.

4.4 Results

Odds ratios. The algorithm generated a total of 135 different possible combinations of data

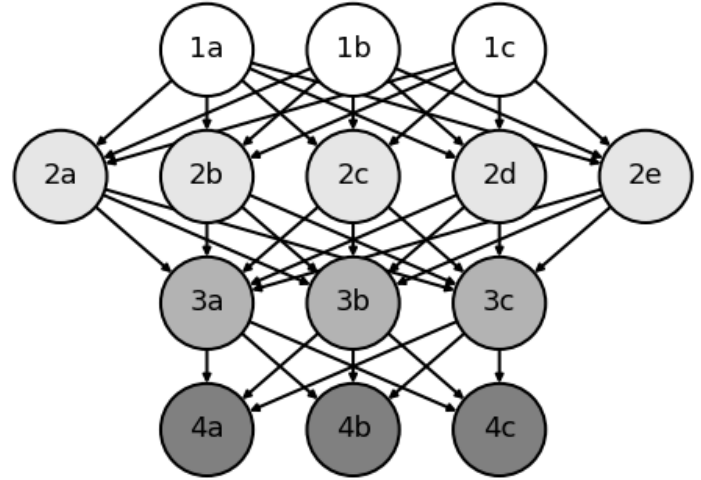


Figure 1: Directed acyclic graph of data preparation pathways

preparation decisions that could have been made, generating 59 distinct results from the same baseline dataset and statistical test (see Fig. 2). The point odds ratio estimates of the association varied between 0.83 (CI = 0.53 – 1.29, $p = 0.40$) to 2.93 (CI = 1.83– 4.70, $p < 0.001$), with a mean OR of 1.31 (SD +/-0.37). 50.37% ($n = 68$) of the odds ratios had statistically significant p-values, defined as $p \leq 0.05$.

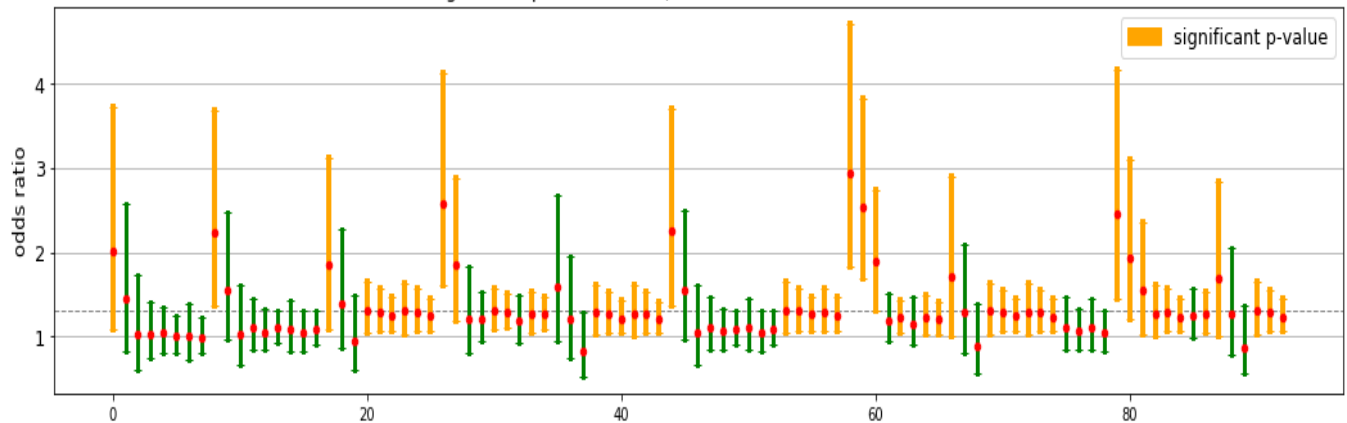


Figure 2: Unique odds ratios, shown with confidence intervals

Cohort comparison. Prior to undergoing data preparation, the cohort in the baseline dataset contained 595 individual patients and information about 604,266 patient days spread between 2010 and 2019. The mean age of patients in the baseline cohort was 16.09 years (SD +/-14.56), and the mean distance to the nearest AQ monitor was 9.84 miles (SD +/-8.91). Patients in rural areas were further away from the nearest air quality monitor than average, with a mean distance of 17.21 miles (SD +/-14.12), while patients in urban areas tended to be closer (6.39 miles, SD +/-3.74). In the baseline dataset 3.55% of the patient days represented data from patients in rural areas, while 67.54% was from urban areas.

After undergoing data preparation, characteristics of the cohorts differed depending on which techniques were used. Decisions about classifying exposures (node 3) exerted a strong influence on patient demographics. Figure 3 shows how more restrictive rules requiring patients to be closer to the nearest air quality monitor in order to be included in the study resulted in a higher percentage of data

being contributed by patients in urban areas; on average, 98.73% of the patient days from Node 3a (drop patients living more than three miles from the nearest air quality monitor) came from patients residing in urban areas. Contrast this with Node 3c, where only 69.35% of the patient days were

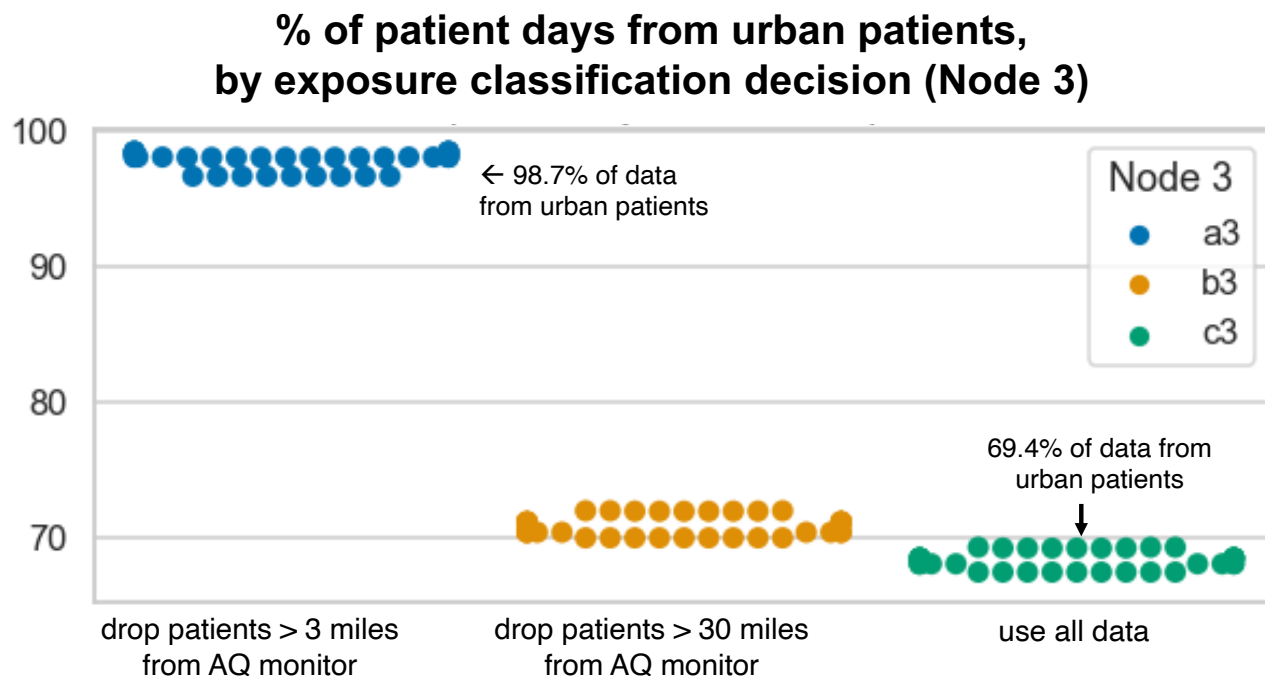


Figure 3: Effect of different exposure classification options (Node 3) on the percentage of patient days contributed by patients in urban areas.

contributed by urban patients – a percentage which more closely aligns with the numbers from the baseline dataset.

The combination of nodes 2b (last known zip code carried forward) and 3c (use data from all patients regardless of how far away they live from the nearest air quality monitor) produced the lowest mean age of patients (8.81 years), while 2d (impute using bootstrap resampling) and 3a (drop data from patients living further than 3 miles from the nearest zip code) produced the highest (18.77 years).

Missing data. Initially, 49.91% of the rows in the baseline dataset were missing information about the patient’s current address; these data therefore could not be linked to community-level data and consequently could not be used in the analysis without transformation. The average age of

patients who were missing address data was 26.92, while the average age of the patients who had an address present in their record was 9.36. Because some data preparation methods removed more data than others, the analysis datasets generated by the algorithm retained between 12.76% ($n = 76$) and 100% of the patients from the baseline dataset, and between 6.49% ($n = 39,199$) and 100% of the patient days.

Because the registry did not confirm patient zip codes prior to 2012, a disproportionate amount of the of the missing data (34.72%) was from the years 2010 to 2011. This meant that the methods which dropped missing data (Nodes 2a, 2b, and 2c) were more likely to omit data from those years. Coincidentally, Oregon and southwest Washington also happened to have very few days with harmful levels of ambient wildfire smoke on those particular years.

Effect size. Decisions exposure classification (Step 3, $g = 0.65, 0.39$, and 0.43) was the most impactful on the results (see Table 2).

Table 4: Effect size of data preparation steps

Step:	1. Validate addresses			2. Handle missing addresses					3. Classify exposures			4. Decide lag times		
Choice:	1a. Validate zip codes using USPS; set invalid zip codes to missing	1b. Ignore invalid zip codes (drops invalid zips during data integration)	1c. Validate zips using number of digits; set invalid zip codes to missing	2a. Drop rows missing data (complete case analysis)	2b. Interpolate first, then drop missing (last observation carried forward)	2c. Triangulate first (try to find address elsewhere), then drop missing	2d. Impute all missing addresses using bootstrap resampling	2e. Triangulate first, then impute	3a. Drop data if patients live > 3 mi. from AQ monitor	3b. Drop data if patients live > 30 mi. from AQ monitor	3c. Use all data regardless of distance	4a. Outcomes occurring w/in ONE week of exposure are associated	4b. Outcomes occurring w/in TWO weeks of exposure are associated	4c. Outcomes occurring w/in FOUR weeks of exposure are associated
mean OR:	1.29	1.34	1.31	1.18	1.26	1.33	1.53	1.27	1.58	1.19	1.17	1.49	1.31	1.14
Effect size:	0.06	0.06	0.01	0.38	0.14	0.04	0.55	0.13	0.65	0.39	0.43	0.44	0.01	0.52
Interpretation:	Minimal/ no effect	Minimal/ no effect	Minimal/ no effect	Small effect	Small effect	Minimal/ no effect	Medium effect	Small effect	Medium effect	Small effect	Medium effect	Medium effect	Minimal/ no effect	Medium effect

4.5 Discussion

Previous studies have demonstrated the importance of data preparation decisions to the field of pharmacoepidemiology;(30, 31) to the best of our knowledge, this is the first study to show how these decisions can impact the results of social determinants of health research or on integrated datasets.

Our analysis has implications for studies utilizing integrated datasets containing both community-level and individual-level data. Integrating health data from the EHR with community-level data is a crucial step for identifying all the factors that affect our health, be they environmental, economic, or societal. Because missing or inaccurate addresses will not match up with any community-level data, methods for handling missing and inaccurate data have a significant impact on the data linkage process. Because we are matching patients to the nearest air quality monitor using their zip codes, an incorrect zip code will fail to match a patient with information about their community's air quality values. While integrating datasets can introduce many new variables into health research, as noted in a recent article in the Journal of Innovation in Health Informatics: "improvements need to occur in the quantity and quality of data input so that proper conclusions between public health and environmental factors may be made."⁽³²⁾

Health equity implications for the use of integrated datasets. We have demonstrated that data preparation techniques have an impact on the results of health research that relies on patient address to link clinical outcomes to neighborhood data. Address data are necessary for linking patients to information about their environment because they provide us with the simplest way to estimate where a patient spends most of their time (and therefore where the patient breathes). However, it has been well-documented that residential instability is high for low-income families – to put it simply, poor people move a lot.^(10, 33-35) Frequent address changes result not only in disrupted continuity of care, but also make it more likely that the health record will have missing, incomplete, or unreliable address data.⁽³⁶⁾ Thus, patients with higher residential instability - a vulnerable group that includes migrant workers, the houseless, and people in poverty - are more likely to be underrepresented in studies that have used complete case analysis and may therefore be inaccurately represented in the research results. In fact, this was confirmed by a review of health data linkage research, which found that people of low socioeconomic status and low educational attainment were particularly likely to be excluded from studies that utilized linkage due to lower rates of consent and less complete data.⁽⁴⁾

Our hope is that understanding the impact of data preparation choices will help data users

make evidence-based data preparation decisions. Although the data cleaning and integration process is usually time-consuming – often acknowledged to be the longest (and most tedious) steps in most research projects (37) - there is currently little guidance on how it should be conducted. Education and training in research methodology focuses on study design and statistical analytic techniques, not on data preparation.(38) Moreover, although literature focuses exhaustively on just about every other step of the research cycle,(39) producing manuscripts on study protocols, study design and methodology, participant recruitment, database structures, and of course, data analysis, it is curiously silent on the topics of data cleaning and on data integration.

Perhaps as a result of the inadequate literature on the topic, there are no standard methods for resolving data quality problems. Given the fact that the needs of individual users tend to vary greatly, perhaps there shouldn't be. Further research into data cleaning, data integration, and data quality improvement does not imply that standards are necessary or desirable. Instead, we hope to further the dialogue about the importance of these steps to the research process by documenting how they are conducted in a particular field.

Limitations: The cohort study we describe in stage one would typically rely on multivariate logistic regression as a statistical method. However, we chose to use Fisher's exact test to generate odds ratios because running the algorithm through numerous possibilities already required considerable computing time; requiring it to generate a complex logit model for every possibility was untenable. Although this means that the relationship we found between wildfire smoke and pulmonary exacerbation may be prone to confounding, our primary focus with this research is not to identify the precise nature of this association. Instead, we aim to show how results from integrated studies are vulnerable to the impacts of different data preparation methods.

Because cystic fibrosis is a rare disease, the number of subjects in our study was small. As a result, more conservative methods for cleaning the dataset, such as those that restricted the sample to only those patients who lived in the same zip code as an air quality monitor, were most likely underpowered and suffered from wide confidence intervals.(40) It is possible that these variations in

sample size, in turn, influenced the effect sizes we used to characterize the impact of the decisions. We hope to replicate this technique with a larger dataset at some point in the future in order to examine whether the effect size was impacted by the order in which these steps were conducted, and to explore whether combinations of multiple steps were more impactful than the individual steps we analyzed here.

4.6 Conclusion

Clinical trials are considered to be more rigorous than the use of EHR data.⁽⁴¹⁾ However, nearly everything we know about social determinants of health comes from studies conducted upon observational data. Our intention is to further knowledge about methods for improving the quality and reliability of research conducted on observational data. Our results also strengthen the argument that transparency is critical for good science and suggest that the “reproducibility crisis” might simply be a failure to provide adequate information about data preparation methods.

Supplementary materials. [\(A\) Python code for the algorithm](#); [\(B\) Forest plot of odds ratios and confidence intervals of all nodes](#).

4.7 References

1. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA*. 2014;311(24):2479-80.
2. Center for Disease Control and Prevention. About Social Determinants of Health (SDOH) [updated 10th March 2021. Available from: <https://www.cdc.gov/socialdeterminants/about.html>.
3. Aalsma MC, Schwartz K, Haight KA, Jarjoura GR, Dir AL. Applying an Electronic Health Records Data Quality Framework Across Service Sectors: A Case Study of Juvenile Justice System Data. *EGEMS (Wash DC)*. 2019;7(1):26.
4. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, et al. Data linkage: a powerful research tool with potential problems. *BMC Health Serv Res*. 2010;10:346.
5. Zandbergen PA. Geocoding Quality and Implications for Spatial Analysis. *Geography Compass*. 2009;3(2):647-80.
6. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*. 2018;1(3):337-56.
7. Silberzahn R, Uhlmann EL. Many hands make tight work. *Nature*. 2015;526:189 - 91.
8. IOM (Institute of Medicine). Capturing social and behavioral domains and measures in electronic health records: Phase 2. Washington, DC: The National Academies Press; 2014.

9. Delgado-Rodríguez M, Llorca J. Bias. *J Epidemiol Community Health*. 2004;58(8):635-41.
10. Schinasi LH, Kenyon CC, Moore K, Melly S, Zhao Y, Hubbard R, et al. Heavy precipitation and asthma exacerbation risk among children: A case-crossover study using electronic health records linked with geospatial data. *Environ Res*. 2020;188:109714.
11. Rusincovitch SA, Pruitt S, Gray R, Li K, Anderson ML, Brinson SW, et al. Design and Implementation of an Automated Geocoding Infrastructure for the Duke Medicine Enterprise Data Warehouse. *AMIA Joint Summits on Translational Science*; San Francisco, 2014.
12. Miranda ML, Ferranti J, Strauss B, Neelon B, Califf RM. Geographic health information systems: a platform to support the 'triple aim'. *Health Aff (Millwood)*. 2013;32(9):1608-15.
13. Bhavsar NA, Gao A, Phelan M, Pagidipati NJ, Goldstein BA. Value of Neighborhood Socioeconomic Status in Predicting Risk of Outcomes in Studies That Use Electronic Health Record Data. *JAMA Netw Open*. 2018;1(5):e182716.
14. Knapp EA, Fink AK, Goss CH, Sewall A, Ostrenga J, Dowd C, et al. The Cystic Fibrosis Foundation Patient Registry. Design and Methods of a National Observational Disease Registry. *Ann Am Thorac Soc*. 2016;13(7):1173-9.
15. Air Quality System Data Mart [internet database]. US Environmental Protection Agency. Available from <http://www.epa.gov/ttn/airs/aqsdatamart>. Accessed 6/25/2020.
16. Omohundro SM. Five Balltree Construction Algorithms. Berkeley, CA: International Computer Science Institute; 1989. Contract No.: TR-89-063.
17. Kim HY. Statistical notes for clinical researchers: Risk difference, risk ratio, and odds ratio. *Restor Dent Endod*. 2017;42(1):72-6.
18. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. *J Am Med Inform Assoc*. 2021;29(1):187-96.
19. National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. Washington, DC: The National Academies Press; 2010.
20. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, et al. Biases introduced by filtering electronic health records for patients with "complete data". *J Am Med Inform Assoc*. 2017;24(6):1134-41.
21. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inform Decis Mak*. 2014;14(1):51.
22. Collins LM, Schafer JL, Kam C-M. A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods*. *Psychol Methods*. 2001; 6(4):330-51.
23. Brugha R, Grigg J. Urban air pollution and respiratory infections. *Paediatr Respir Rev*. 2014;15(2):194-9.
24. Kravets N, Parker JD. Linkage of the Third National Health and Nutrition Examination Survey to air quality data. *Vital Health Stat* 2. Nov 2008;2(149):1-16.
25. Roca-Barcelo A, Douglas P, Fecht D, Sterrantino AF, Williams B, Blangiardo M, et al. Risk of respiratory hospital admission associated with modelled concentrations of *Aspergillus fumigatus* from composting facilities in England. *Environ Res*. 2020;183:108949.
26. Goss CH, Newsom SA, Schildcrout JS, Sheppard L, Kaufman JD. Effect of ambient air pollution on pulmonary exacerbations and lung function in cystic fibrosis. *Am J Respir Crit Care Med*. 2004;169(7):816-21.
27. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Revised edition. New York.: Academic Press; 1977.
28. Kelley K, Preacher KJ. On effect size. *Psychol Methods*. 2012;17(2):137-52.
29. Hedges LV. Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*. 1981;6(2):107-28.
30. Pye SR, Sheppard T, Joseph RM, Lunt M, Girard N, Haas JS, et al. Assumptions made when preparing drug exposure data for analysis have an impact on results : An unreported step in pharmacoepidemiology studies. *Pharmacoepidemiol Drug Saf*. 2018;27(7):781-8.
31. Eskin M, Eurich DT, Simpson SH. Impact of Drug Exposure Definitions on Observed Associations in Pharmacoepidemiology Research. *J Popul Ther Clin Pharmacol*. 2018;25(1):e39-e52.

32. Shaw NT, McGuire SK. Understanding the use of geographical information systems (GIS) in health informatics research: A review. *J Innov Health Inform.* 2017;24(2):940.
33. Phinney R. Exploring Residential Mobility among Low-Income Families. *Soc Serv Rev.* 2013;780 - 815.
34. DeLuca S, Rosenblatt P, Wood H. Why Poor People Move (and Where They Go)- Residential Mobility, Selection and Stratification.
35. Knighton AJ. Is a Patient's Current Address of Record a Reasonable Measure of Neighborhood Deprivation Exposure? A Case for the Use of Point in Time Measures of Residence in Clinical Care. *Health Equity.* 2018;2(1):62-9.
36. Fankhauser K, Marino M, Giebultowicz S, Cowburn S, Lucas J, Kaufmann J, et al. Integration of community-level determinants of health with longitudinal patient records accounting for patient mobility. *North American Primary Care*; 11/17/20192019.
37. Lohr S. For Data Scientists, 'Janitor Work' Is Hurdle to Insights. *The New York Times.* 2014 2014/08/18/.
38. Zhu Y, Hernandez LM, Mueller P, Dong Y, Forman MR. Data Acquisition and Preprocessing in Studies on Humans: What Is Not Taught in Statistics Classes? *Am Stat.* 2013;67(4):235-41.
39. Johnson SB. Clinical Research Informatics: Supporting the Research Study Lifecycle. *Yearb Med Inform.* 2017;26(1):193-200.
40. White E, Kushiz LH, Pepe MS. The effect of exposure variance and exposure measurement error on study sample size: Implications for the design of epidemiologic studies. *J Clin Epidemiol.* 1994;47(8):873-80.
41. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health.* 2016;37:61-81.

Appendix 4a. Python code for the algorithm described in Chapter 4

Nota bene: the code below was originally created and run in Python 3.7. I cannot guarantee that it will operate the same in different versions of Python.

```
import numpy as np
import pandas as pd
import random
from scipy import stats
import statsmodels.api as sm

#Import dataframe created earlier of all dates in the study (summertime dates, as that is when
wildfire season #occurs in Oregon and SW Washington.
annual_dates = pd.read_csv('all_summer_days.csv',
                           header=0, sep=',', parse_dates=['date']).convert_dtypes()

#Import dataframe of all zip codes in OR and WA and the air quality monitor closest to that code,
# which also includes indicator of whether that zip is urban or rural
zips_mons = pd.read_csv(
    'zipcodes_and_AQmonitors.csv',
    sep=',', header=0, index_col=['Unnamed: 0'],
    usecols=['Unnamed: 0', 'zip', 'AQSID', 'distance']).convert_dtypes()
real_zips = list(zips_mons['zip']) #creates a list of valid zip codes for OR and WA (according to the
Post Office)

#Import dataframe of air quality monitor readings for the study period
NW_PM_data = pd.read_csv(
    'AQI_PM_df.csv',
    header=0, sep=',', index_col=['Unnamed: 0'],
    parse_dates=['begin date']).convert_dtypes()

#Import dataframe of patients experiencing exacerbations and the dates of those exacerbations
exacerbation_data = pd.read_csv('/exacerbations.csv',
                                header=0, sep=',', parse_dates=['exacerbation_date']).convert_dtypes()
```

```
#Define a data preprocessing function with randomly selected
#data preparation and imputation methods
def df_preprocess(epoch):

    #Step 1 validates zip codes by identifying implausible (those outside WA or OR)
    # and non-conforming zip codes (those with something other than 5 digits)

    #Step 1 option A: set the value of non-conforming zips to na
    def a1(d_f):
        d_f['Current Zip'].loc[~d_f['Current Zip'].isin(real_zips)] = np.nan
        d_f['Current Zip_x'].loc[~d_f['Current Zip_x'].isin(real_zips)] = np.nan
        d_f['Current Zip_y'].loc[~d_f['Current Zip_y'].isin(real_zips)] = np.nan
        return d_f

    #Step 1 option b: ignore non-conforming and implausible zip codes
    #Note: this just copies the same df because it does not need a filter
```

```

def b1(d_f):
    d_f = d_f.copy()
    return d_f

na
#Step 1 option C: set the value of implausible zips (those with something other than 5 digits) to
na
def c1(d_f):
    d_f.loc[(d_f['Current Zip'] < 10000), 'Current Zip'] = np.nan
    d_f.loc[(d_f['Current Zip'] > 99999), 'Current Zip'] = np.nan

    d_f.loc[(d_f['Current Zip_x'] < 10000), 'Current Zip_x'] = np.nan
    d_f.loc[(d_f['Current Zip_x'] > 99999), 'Current Zip_x'] = np.nan

    d_f.loc[(d_f['Current Zip_y'] < 10000), 'Current Zip_y'] = np.nan
    d_f.loc[(d_f['Current Zip_y'] > 99999), 'Current Zip_y'] = np.nan
    return d_f
step1 = [a1, b1, c1]

#Define the preparation methods:
#Step 2 option A: Casewise deletion
def a2(d_f):
    d_f.dropna(subset=['Current Zip'], inplace = True)
    return d_f

#Step 2 Option B: Interpolation
# (i.e., infers missing zips with zips from the same patient at different years)
# (Uses Only confirmed patient addresses, so rows from patients missing ALL address data are
dropped)
def b2(d_f):
    d_f.update(d_f.groupby('patient_ID')['Current Zip'].fillna(method = 'bfill'))
    d_f.update(d_f.groupby('patient_ID')['Current Zip'].fillna(method = 'ffill'))
    return d_f

#Step 2 option C: Triangulation
# (Fills in missing zip codes across datasets with zips from the same patient)
def c2(d_f):
    d_f[['Current Zip_x', 'Current Zip_y', 'Current Zip']] = d_f[['Current Zip_x', 'Current
Zip_y', 'Current Zip']].replace("", np.nan).bfill(axis = 'columns').ffill(axis = 'columns')
    return d_f

#Step 2 option d: uses bootstrap resampling procedure to fill in missing zip codes
def d2(d_f):
    boot_strap= d_f[["Current Zip_x", "Current Zip_y", "Current Zip"]].copy().astype('Int64')

    #make a list of zip codes present in the dataset
    whole_zip1 = boot_strap["Current Zip"][boot_strap["Current Zip"].notnull()]
    whole_zip2 = boot_strap["Current Zip_x"][boot_strap["Current Zip_x"].notnull()]
    whole_zip3 = boot_strap["Current Zip_y"][boot_strap["Current Zip_y"].notnull()]
    complete_zips = [whole_zip1, whole_zip2, whole_zip3]
    fill_zips = pd.concat(complete_zips).astype('Int64')
    filled_zips = fill_zips.reset_index(drop=True)
    boot_strap["Current Zip"].fillna(filled_zips, inplace = True)
    d_f["Current Zip"] = boot_strap["Current Zip"]

```

```

        return d_f

#Step 2 option E: Triangulation, then uses bootstrap resampling to fill missing zip codes
def e2(d_f):
    d_f[['Current Zip_x', 'Current Zip_y', 'Current Zip']] = d_f[['Current Zip_x', 'Current
Zip_y', 'Current Zip']].replace("", np.nan).bfill(axis = 'columns').ffill(axis = 'columns')
    boot_strap= d_f[['Current Zip_x', "Current Zip_y", "Current Zip"]].copy().astype('Int64')

    #create a series of zip codes present in the dataset
    whole_zip1 = boot_strap["Current Zip"][boot_strap["Current Zip"].notnull()]
    whole_zip2 = boot_strap["Current Zip_x"][boot_strap["Current Zip_x"].notnull()]
    whole_zip3 = boot_strap["Current Zip_y"][boot_strap["Current Zip_y"].notnull()]
    complete_zips = [whole_zip1, whole_zip2, whole_zip3]
    fill_zips = pd.concat(complete_zips).astype('Int64')
    filled_zips = fill_zips.reset_index(drop=True)
    boot_strap["Current Zip"].fillna(filled_zips, inplace = True)
    d_f["Current Zip"] = boot_strap["Current Zip"]
    return d_f

#List of step 2 functions
step2 = [a2, b2, c2, d2, e2]

#Step 3 option a: use only patients who live within close proximity to the AQ monitor
def a3(d_f):
    d_f = d_f[(d_f['distance'] <= 3)]
    return d_f

#Step 3 option b:
#Filters dataset to include ONLY patients who live within 30 miles of the nearest air quality
monitor
def b3(d_f):
    d_f = d_f[(d_f['distance'] <= 30)]
    return d_f

#Step 3 option c:
#Use data from the air quality monitor nearest to the patient regardless of distance.
#Note: this just copies the same df because it does not need a filter
def c3(d_f):
    d_f = d_f.copy()
    return d_f

#List of step 4 functions
step3 = [a3, b3, c3]

#lag times
#one week lag time
def a4(d_f):
    d_f = d_f.dropna(subset=['AQSID']).sort_values(by = ['date'])
    d_f = pd.merge_asof(
        d_f, right = NW_PM_data, by='AQSID',
        left_on = 'date', right_on = 'begin date',
        direction='backward', tolerance= pd.Timedelta(weeks = 1))
    return d_f

```

```

#two week lag time
def b4(d_f):
    d_f = d_f.dropna(subset=['AQSID']).sort_values(by = ['date'])
    d_f = pd.merge_asof(
        d_f, right = NW_PM_data, by='AQSID',
        left_on = 'date', right_on = 'begin date',
        direction='backward', tolerance= pd.Timedelta(weeks = 2))
    return d_f

#four week lag time
def c4(d_f):
    d_f = d_f.dropna(subset=['AQSID']).sort_values(by = ['date'])
    d_f = pd.merge_asof(
        d_f, right = NW_PM_data, by='AQSID',
        left_on = 'date', right_on = 'begin date',
        direction='backward', tolerance= pd.Timedelta(weeks = 4))
    return d_f

#List of step 4 functions
step4 = [a4, b4, c4]

all_decisions = [] #stores which node was selected
test_result = [] #stores ORs from the fisher's exact test
LCB = [] #stores lower confidence boundary from the fisher's exact test
UCB = [] #stores upper confidence boundary from the fisher's exact test
p_val = [] #list holding p-values of all results from fisher's exact
length = [] #list of length of the dataset
mean_age = [] #list of mean age of patients in the dataset
rural_rate = [] #list of % of patient days representing data from patients in rural areas
urban_rate = [] #list of % of patient days representing data from patients in urban areas
num_patients = [] #list of the number of patients in the final dataset

for i in range(epoch):
    #Renew the original dataframe
    data = pd.read_csv('all_patient_data_df.csv',
        header=0, sep=',').convert_dtypes()

    #Randomly select one method in step 1 list
    step_one = random.choice(step1)
    #Get function name
    step_one_name = step_one.__name__
    #Update dataframe
    step_1 = step_one(data)
    del data, step_one

    #Randomly select one imputation method in step 2 list
    step_two = random.choice(step2)
    #Get function name
    step_two_name = step_two.__name__
    #Updated dataframe
    step_2 = step_two(step_1)

```

```

#Join the df selected in Step 1 ("step_1") with df of all the days in the study to create a df of
patient days
step2_ = pd.merge(step_2, annual_dates, how='left', left_on='Annual Review Year', right_on
= 'year')

#Join the new df with df of all patient exacerbations (has to be done after adding the dates in
the annual_dates df
step_2_df = pd.merge(step2_, exacerbation_data, how='left', left_on = ['patient_ ID', 'date'],
right_on = ['patient_ ID', 'exacerbation_date'])
step2_df = pd.merge(step_2_df, zips_mons, how='left', left_on='Current Zip', right_on='zip')
del step_2, step_two, step2_, step_2_df

#Randomly select one method in step 3 list
step_three = random.choice(step3)
#Get function name
step_three_name = step_three.__name__
#Update dataframe
step_3 = step_three(step2_df)
del step2_df, step_three

#Randomly select one method in step 4 list
step_four = random.choice(step4)
#Get function name
step_four_name = step_four.__name__
#Update dataframe
step_4 = step_four(step_3)

#now add exacerbations by patient and year to the table of exposed patients
step4_d = step_4.copy()
step4_d['exposed'] = step4_d['exposed'].fillna(0).astype('int')
step4_d['exacerbation'] = step4_d['exacerbation'].fillna(0).astype('int')
step4_d_f = step4_d.copy()
step4_df = step4_d_f.sort_values(by = ['patient_ ID', 'date'])
del step4_d_f, step_4, step4_d

#Use the last dataframe perform the test:
#build a 2x2 table, starting with dataframes for the outcome variable (i.e., disease state)
no_condition = step4_df[step4_df.exacerbation == 0] #equal to b+d
condition = step4_df[step4_df.exacerbation == 1] #equal to a+c

#create dataframe for all exposures
no_wildfire = step4_df[step4_df.exposed == 0] #equal to c+d
wildfire = step4_df[step4_df.exposed == 1] #equal to a+b

#store number of rows in the analysis dataset
final_df_length = len(step4_df)
length.append(final_df_length)

#store number of patients in the final dataset
tot_patients = step4_df['patient_ID'].nunique()
num_patients.append(tot_patients)

#store mean age of patients in the final dataset

```

```

age_patients = round((step4_df.groupby('patient_ID')['age'].mean()).mean(), 2)
mean_age.append(age_patients)

#store % of patient days representing data from patients in rural areas
rural_zips = round((len(step4_df[(step4_df['rural'] == 1)])/len(step4_df)*100), 2)
rural_rate.append(rural_zips)

a_b = len(wildfire) #all exposed patient days
a_c = len(condition) #all patient-days with exacerbations
b_d = len(no_condition) #all patient-days w/o exacerbations
b = no_condition.exposed.sum() #exposed patient days w/o exacerbations
a = condition.exposed.sum() #exposed patient days WITH exacerbations (should be the
smallest)
c = a_c - a #exposed patient days w/o exacerbations
d = b_d - b #patient days with no exposure and no exacerbation (should be the largest group)
del wildfire, condition, no_condition, no_wildfire, final_df_length

#create a 2x2 table
table = np.asarray([[a, c], [b, d]])
t22 = sm.stats.Table2x2(table)

# perform fisher's exact test on the table to calculate odds ratio
oddsratio = t22.oddsratio
oddsratio = round(oddsratio, 3)
lcb, ucb = t22.oddsratio_confint()
lcb = round(lcb, 3)
ucb = round(ucb, 3)
pvalue = t22.oddsratio_pvalue()
pvalue = round(pvalue, 3)

del table, t22
del a, c, d, b,

#Store test results of each run in a list
test_result.append(oddsratio)
p_val.append(pvalue)
LCB.append(lcb)
UCB.append(ucb)
del oddsratio, pvalue, lcb, ucb

#Store decisions you made for each step in a list
decision_list = [step_one_name, step_two_name, step_three_name, step_four_name]
all_decisions.append(decision_list)

#clear dataframes to prevent data from being carried forward
del step_one_name, step_two_name, step_three_name, step_four_name

#Create a dataframe to store your decisions and test results for each run
output_df = pd.DataFrame(all_decisions, columns = ['Node 1', 'Node 2', 'Node 3', 'Node 4'])
output_df['odds_ratio'] = test_result
output_df['p'] = p_val
output_df['patients'] = num_patients
output_df['rows'] = length

```

```

output_df['LCB'] = LCB
output_df['UCB'] = UCB
output_df['age'] = mean_age
output_df['rural'] = rural_rate
output_df['urban'] = urban_rate
output_df.sort_values(by=['Node 1', 'Node 2', 'Node 3', 'Node 4'], inplace = True)

return output_df

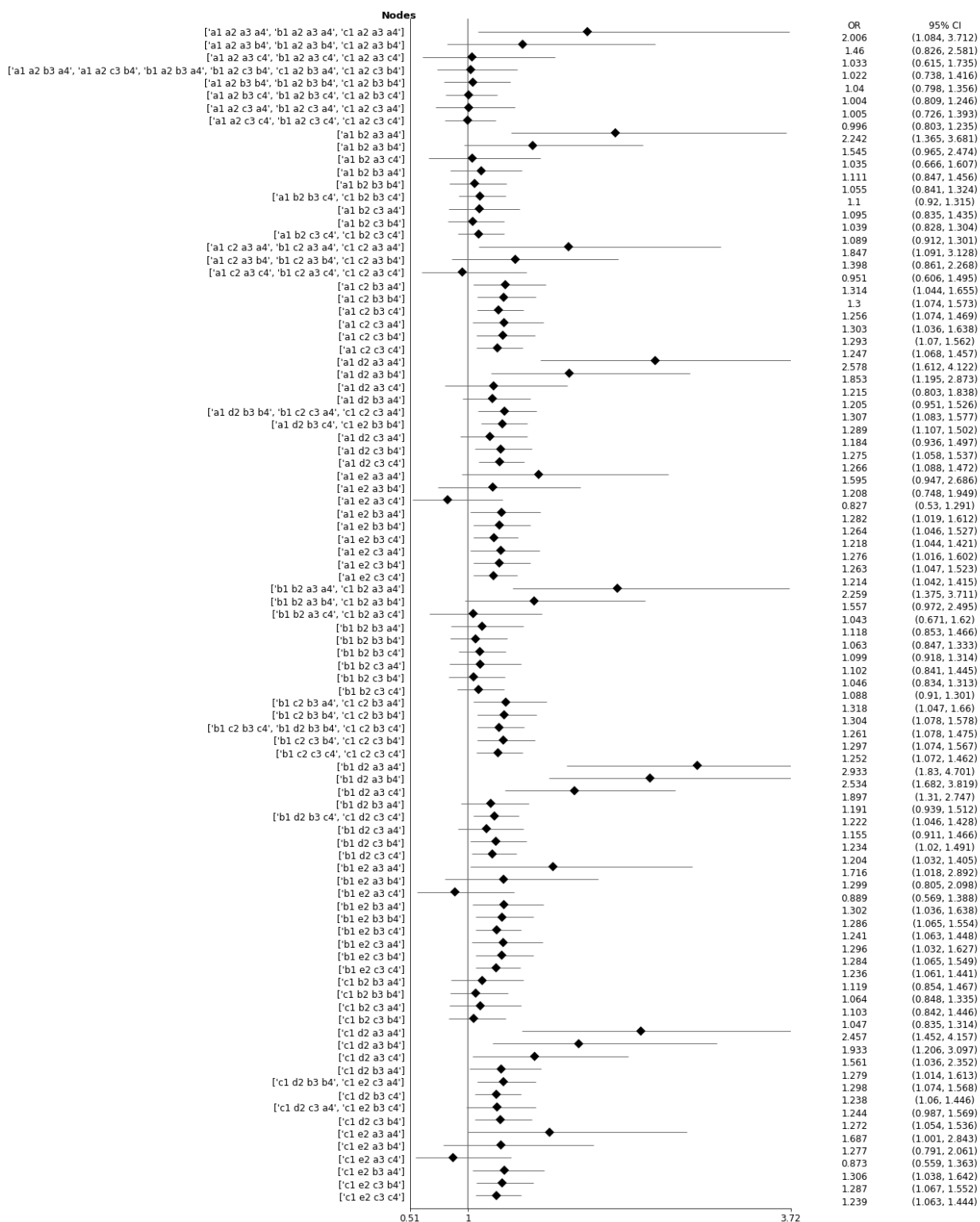
```

```

#runs the above function, then saves if it has run through all possible combinations
# and otherwise runs it again.
result = df_preprocess(1500)
result.drop_duplicates(subset = ['Node 1', 'Node 2', 'Node 3', 'Node 4'], inplace=True)
while len(result) < (3*5*3*3): #if this is run was complete, then save the file; otherwise run it again.
    print("You need to run the function again because it didn't include all the runs.")
    result = df_preprocess(1500)
    result.drop_duplicates(subset = ['Node 1', 'Node 2', 'Node 3', 'Node 4'], inplace=True)
else: #makes a column that lists all the nodes in that row
    result['Nodes'] = result['Node 1'].map(str) + ' ' + result['Node 2'].map(str) + ' ' + result['Node
3'].map(str) + ' ' + result['Node 4'].map(str) + ' ' + result['Node 5'].map(str)
    #create a column, 'signif,' to indicate whether that result was statistically significant
    result['signif'] = np.where(result['p'] <= 0.05, 1, 0)
    results = result.copy()
    results.to_csv('result_df.csv', index = False)
    print("IT WORKED.")

```

Appendix 4b. Forest plot of odds ratios and confidence intervals for all nodes



Chapter 5. Conclusion

The availability of electronic clinical data and the recognition of the importance of real-world evidence have made secondary use research— in which clinical data are used for purposes beyond clinical care— common. The question is no longer “**can** we use electronic health record data for research?” but rather “**how** do we make electronic health record data usable for research?” Simultaneously, there has been increasing focus on health equity and disparities, and a need for high-quality social determinants of health (SDoH) data to assess and characterize these issues. Social determinants of health data is necessary for healthcare systems and public health officials to monitor health disparities and provide researchers the opportunity to quantify the impact of racism on health.(1) Unfortunately, the quality of this data in the electronic health records (EHR) tends to be poor, limiting the potential for valid reuse.

This body of work has contributed solutions to those problems by furthering the development of methods for improving the quality and of the social determinants of health data used in health equity research. I approached the issue by focusing on data quality improvement practices for real-world data, particularly on the data preparation process and the data pipeline. My first study, a systematic literature review, had three major findings: first, the quality of data about race and ethnicity is an important topic with a large body of literature; second, problems with the quality of SDoH data not distributed randomly, with data from Hispanic patients being particularly prone to bias; and third, there are methods that may improve the quality of “real-world” data. These initial findings inspired my subsequent studies. My next study, an analysis of the data about race and ethnicity in the National Covid Cohort Collaborative Data Enclave, demonstrated the relevance of the first aim by showing how those findings generalize across healthcare systems. Finally, my third study presented a data-driven approach to the application of data quality improvement methods.

5.1 Contributions

The research presented here has furthered knowledge about the quality and reliability of the social determinants of health data used in health equity research. It has highlighted several important considerations for users of real-world datasets, shown opportunities to improve the quality of these data at various points in the data pipeline, and identified evidence-based methods available to researchers seeking to improve the quality of this data.

My first study demonstrated that some social determinants data are more prone to problems with data quality, particularly patient race/ethnicity and integrated datasets that merge individual-level and community-level data. Subsequently, I showed that data quality problems can create issues with validity and reliability, and that the impact of these data quality issues was not equal across all races and ethnicities. This literature review also connected the concepts of data quality and bias, showing how failure to identify and address these data quality problems has the potential to limit the validity and reliability of efforts to understand and reduce health inequity. It identified a trend of problems with the quality of data about Hispanic/Latino and American Indian patients.

My second paper showed that there can be significant challenges with the quality of race and ethnicity data in large, centralized databases. My analysis called attention to issues with the quality of some of the race and ethnicity data in the National Covid Cohort Collaborative Data Enclave. Critically, I demonstrated that data from some of the very groups hardest hit by the pandemic – American Indians, Black or African Americans, and Hispanic/Latinos - are disproportionately affected by these data quality problems. This paper set in motion a process which promises to make available the data on American Indian patients currently obscured within the N3C. The manuscript builds an argument that “rescuing” this data will allow for better identification of research cohorts and avoid the erasure of minoritized patients, and that data pipelines which allow more data to be retained for analysis can help improve the usability and validity of these data.(2)

My final paper demonstrates quantitatively that the conclusions drawn from an analysis can be highly dependent on how the data has been prepared. Although previous studies have

demonstrated the importance of data preparation decisions to the field of pharmacoepidemiology,(3, 4) to the best of our knowledge this is the first study to show how these decisions can impact the results of social determinants of health research or on integrated datasets. This finding has important implications for studies utilizing integrated datasets containing both community-level and individual-level data. Integrating health data from the EHR with community-level data is a crucial step for identifying all the factors that affect our health, be they environmental, economic, or societal. Because missing or inaccurate addresses will not match up with any community-level data, methods for handling missing and inaccurate data can have a substantial impact on the data linkage process. While integrating datasets can introduce many new variables into health research, as noted in a recent article in the Journal of Innovation in Health Informatics: “improvements need to occur in the quantity and quality of data input so that proper conclusions between public health and environmental factors may be made.”(5)

5.2 Limitations

The work here is focused on social determinants variables available either as demographic information or by data linkage, and not on variables collected through structured assessments (e.g., PRAPARE) or those documented through diagnostic codes (i.e., ICD V-codes). For the first study, although some social determinants variables may have more thoroughly documented or easily quantifiable data quality issues, it is important to note that this does not necessarily mean those variables are of poor quality. A larger body of research may indicate simply that these variables are more accessible and therefore easier to study. In the second study, the provenance of the data from the National COVID Cohort Collaborative is limited, which makes it difficult to trace exactly how errors may have occurred. Although the primary focus of my third study this research was ultimately not epidemiological, it is nonetheless likely that my final study was underpowered. The findings of that study may therefore have been impacted by the small sample size.

5.3 Conclusion

As the field of data science has shifted from collecting data toward analyzing this data, informatics researchers have been working on the problem of secondary use – how to make patient data intended as clinical documentation usable as a source for observational research. Currently, observational research is known to be a less rigorous than clinical trials. However, nearly everything we know about social determinants of health comes from studies conducted upon observational data. This work furthers the knowledge about methods for improving the quality and reliability of research conducted on the real-world datasets used for observational research. I have shown that consideration of data quality and thoughtfully applied quality improvement methods may help prevent bias, thus improving the validity of research conducted with social determinants of health data. I have demonstrated that transparency about these numerous data preparation decisions is critical if researchers are to make accurate inferences from analyses. I hope to open a dialogue about how bias is introduced into data and normalize the practice of reporting data preparation techniques in the methods section of papers.

5.4 References

1. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences Research. *The Annals of Family Medicine*. 2022;20(2):157-63.
2. Rahm E, Do HH. Data Cleaning: Problems and Current Approaches. *Bulletin of the Technical Committee on Data Engineering*. 2000;23(4):3 - 13.
3. Pye SR, Sheppard T, Joseph RM, Lunt M, Girard N, Haas JS, et al. Assumptions made when preparing drug exposure data for analysis have an impact on results : An unreported step in pharmacoepidemiology studies. *Pharmacoepidemiol Drug Saf*. 2018;27(7):781-8.
4. Eskin M, Eurich DT, Simpson SH. Impact of Drug Exposure Definitions on Observed Associations in Pharmacoepidemiology Research. *J Popul Ther Clin Pharmacol*. 2018;25(1):e39-e52.
5. Shaw NT, McGuire SK. Understanding the use of geographical information systems (GIS) in health informatics research: A review. *J Innov Health Inform*. 2017;24(2):940.