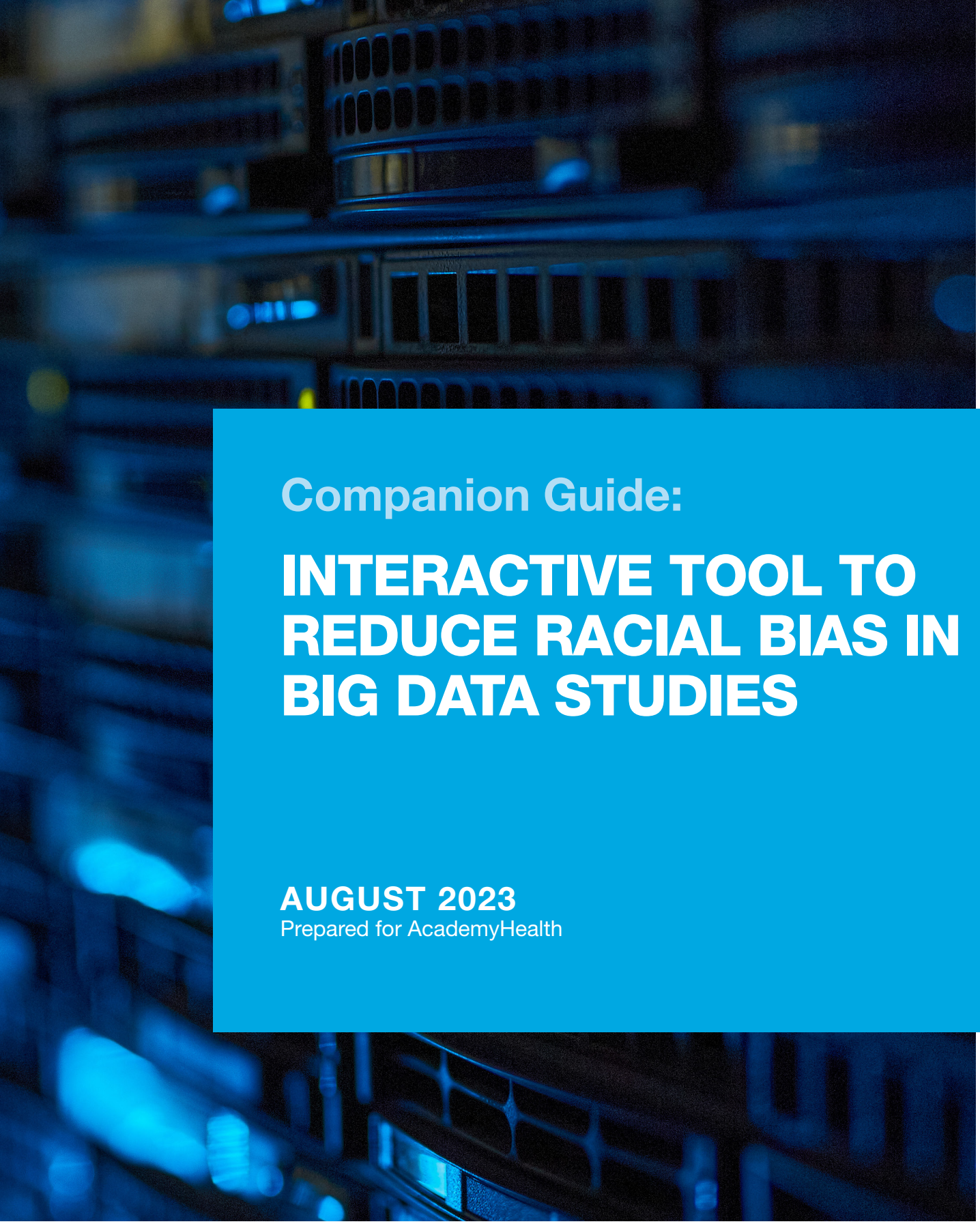


Companion Guide:

INTERACTIVE TOOL TO REDUCE RACIAL BIAS IN BIG DATA STUDIES

AUGUST 2023
Prepared for AcademyHealth

AcademyHealth

ABOUT THE INTERACTIVE TOOL TO REDUCE RACIAL BIAS IN BIG DATA STUDIES:

The **Interactive Tool to Reduce Racial Bias** in Big Data Studies was designed as an easy-to-use resource that guides researchers in addressing potential sources of racial bias in their work. The Tool presents researchers with a sequence of questions designed to:

- Prompt thinking about ways to identify "blind spots" and potential sources of racial bias in their big data sources and methods.
- Provide them with linked and follow-on resources that guide them in anticipating and addressing sources of racial bias in their work.

Available online, at **bit.ly/Tool2ReduceBias**, the Tool can be used throughout the research lifecycle: from study planning and design, through data collection and analysis, to translation and even dissemination.

- It can help to prompt early thinking about the selection of data sources and messages, as well as research questions and the analysis plan.
- In cases where risk of bias cannot be eliminated, the Tool encourages researchers to explicitly acknowledge and communicate limitations. It also provides guidance on how and where to acknowledge these limitations in final research products, for example, to promote transparency.
- Users of the Tool can opt to receive an email including (1) a summary of their responses, flagging opportunities to address potential sources of bias; and (2) this *Companion Guide,* outlining actionable strategies and helpful resources they can employ moving forward.

CONTENTS

Background: Big Data in Health Research 5
Section 1: Your Big Data Research Study 6
1.1 Motivation for Big Data Research Study 6
1.2 Consider Your Audience and Use Cases When Developing Your Big Data Research Study 8
Section 2: Big Data Sources and Documentation 10
2.1 Collecting and Finding Big Data 11
2.2 Characteristics of a High-Quality Big Dataset 13
Section 3: Race and Ethnicity in Big Data 15
3.1 Why Considering Race and Ethnicity is Important 15
3.3 Key Analysis Decisions for Race and Ethnicity Data 18
3.4 Assessing Algorithmic Bias in Big Data Models by Race and Ethnicity 20
Section 4: Use of Proxy Variables 21
Section 5: Data Completeness 23
Section 6: Big Data Representativeness 26
Final Notes 28
Lead Author 29
Acknowledgements 29
Suggested Citation 29
References 30

KEY TERMS

Artificial Intelligence (AI): A non-human program or model that can solve sophisticated tasks.

Big Data: Datasets requiring design and decisions beyond traditional computational or statistical tools (see A Note on Defining Big Data in Health Care Research)

Machine Learning (ML): A program or system that trains a model from input data. The trained model can make useful predictions from new (never-before-seen) data. Traditionally considered a subfield of AI, but often used interchangeably.

Natural Language Processing (NLP): Branch of computer science focused on making computer systems understand and analyze written or spoken human language. May use machine learning methods.

Sources:

Machine Learning Glossary, Google Developers. https://developers.google.com/machine-learning/glossary

Data Science Glossary: Definitions for Common Data Science Terms. https://www.datacamp.com/blog/data-science-glossary

INTRODUCTION

Over the last few years, the number of big data studies in health care research has continued to increase.[1] Researchers have explored numerous health care big data sources including clinical data such as electronic health records, biometric data from wearables and medical devices, genomics data from studying biological systems, and social media and web data.[1,2] The term "big data" is commonly understood as referring to datasets that require design and decision factors beyond the scope of traditional statistical software.[3] The strength of big data analytics is in using various forms of data mining to identify associations and patterns.[4] Big data research may help to address significant health care challenges such as those associated with personalizing treatment, predicting outbreaks of epidemics, drug and medical device safety surveillance, and improving patient outcomes and quality of care.[5,6]

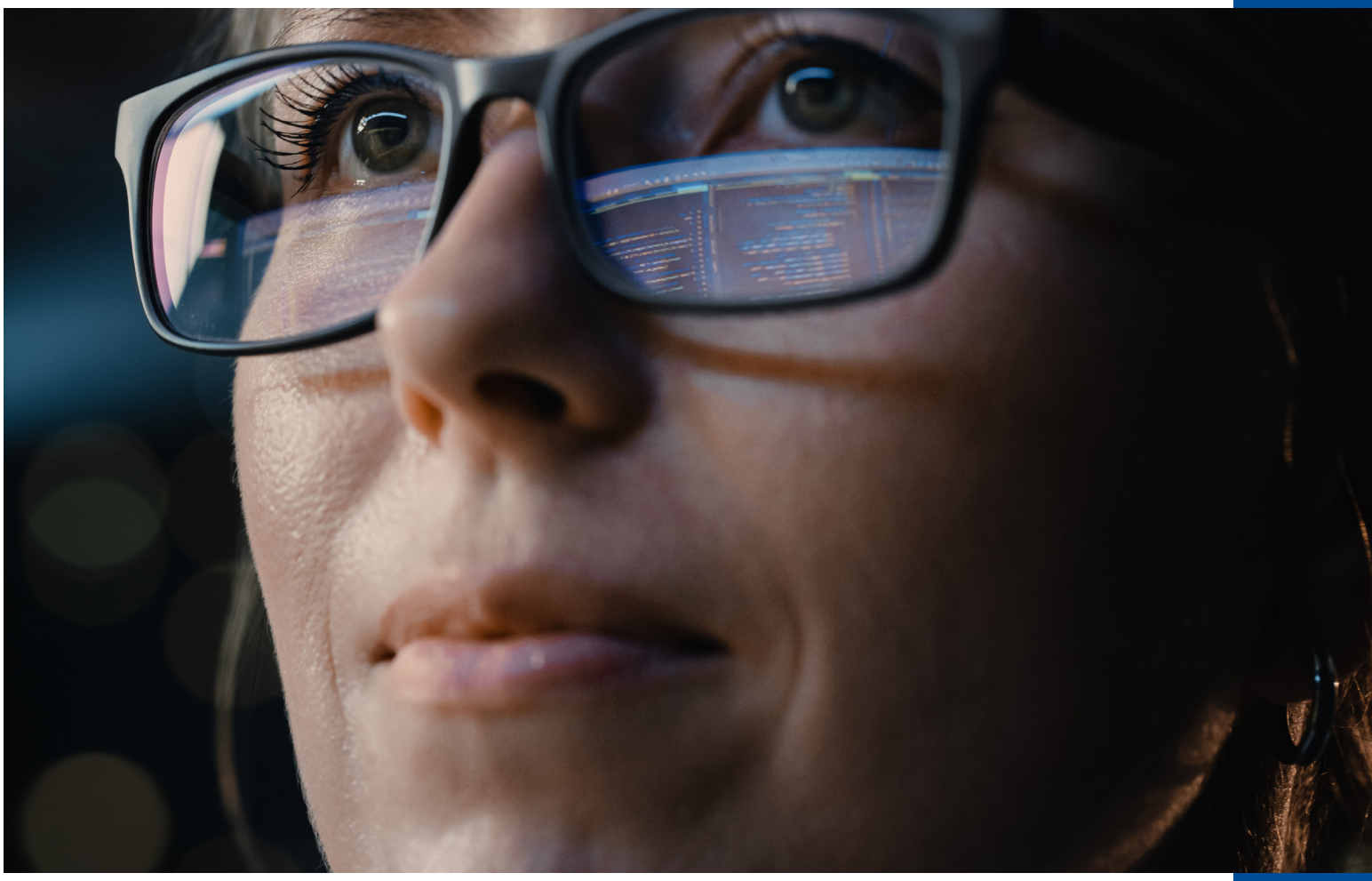
Though big data research offers considerable promise for humanity, without safeguards, this same research can reproduce and amplify existing societal biases and disparities.[7–9] Given the increased public awareness and acknowledgment of structural racism, and its cascading effects, bias based on race and ethnicity in big data research is particularly concerning.[10] This report is grounded in the understanding that race and ethnicity are social constructs without scientific or biological meaning, and that traditional research methods have contributed to racial and ethnic inequities and disparities.[11–14] However, it is important to consider how race and ethnicity interact with health outcomes because the experiences of discrimination based on race or ethnicity (i.e., the experience of racism, marginalization, lack of access) may be connected or contributing to physical or mental health—yielding health inequities across racial or ethnic groups. Thus, the datasets, algorithms, and methods used to support big data research are not objective and neutral; instead, they are capable of reproducing human bias and racist stereotypes.[153] Big data researchers can mitigate the risk of perpetuating bias based on race or ethnicity by intentionally incorporating best practices that reduce or eliminate bias.

This report is intended as a companion resource for the *Interactive Tool to Reduce Racial Bias in Big Data Studies* (hereafter, referred to as the "Tool"). This Tool was developed through an iterative co-design process funded by the Robert Wood Johnson Foundation and led by AcademyHealth, as part of its *Paradigm Project.* The Tool itself prompts people to think critically about the ways they might intentionally adjust their methods

for collecting, accessing, and using data to avoid unintentionally perpetuating race- or ethnicity-based bias, and to look at race in ways that can lead to solutions. In tandem, this guide includes additional detail and outlines action steps aimed at helping people to proactively adjust their approaches—bringing these into alignment with best practices for mitigating bias in big data studies. The guide focuses on assessing bias based on race or ethnicity throughout six key portions of the big data workflow: defining the study, sourcing big data, evaluating race and ethnicity attributes, identifying proxies, addressing data completeness, and assessing representativeness. As AcademyHealth has led this work, and its organizational mission focuses on health services research (HSR), examples and resources referenced throughout are primarily health care-oriented; however, some of the key data science insights and practices may translate into big data research on broader health or other topics. The Tool can be used in multiple ways, including but not limited to:

- serving as a reference resource and prompting researchers' critical thinking around related issues, prior to starting a study;
- supporting training and educational activities, in both professional and academic settings; and
- providing "checkpoint" that published, conference hosts, or other entities can use to ensure their affiliates' work is aligned with best practices in the field.

Researchers can review this report in its entirety or focus on specific sections highlighted in the output of the Tool.



BACKGROUND: BIG DATA IN HEALTH RESEARCH

There is no single consensus on the definition of big data.[16,17] Initial conceptualizations of big data have focused on three key factors: volume (the size of the data), variety (the types of data, such as unstructured text or image data), and the velocity (speed of data flow and analysis).[18] Recent studies since suggested that these three factors are too limiting, vary by field, and not uniformly applied, even to datasets that are commonly recognized as "big data". Further work has sought to define big data through additional attributes such as veracity (data may be messy and hard to verify), relationality (linkage of multiple datasets), and exhaustivity (covers entire populations).[17] In *Toward a Literature-Driven Definition of Big Data in Health Care,* Baro et al. (2015) specifically focus on defining big data by volume in the medical field and suggest that a study should be considered "big data" when the decimal logarithm of the number of statistical individuals (n) multiplied by the number of number of variables (p) is greater than 7 (i.e., $\text{Log}(n*p) > 7$).[19] Yet even this threshold is considerably smaller than "big data" in domains like astronomy and internet research which encounter terabytes and petabytes of data, and neglects other computational ways in which data may seem "big".[4,20]

For the purposes of this report, we understand "big data" as a dataset that challenges existing statistical and computational tools. We recognize that what constitutes big data may look different for studies across varying health care fields and leave each researcher to determine whether "big data" is an appropriate descriptor for their work. We have attempted to write this report such that the key principles and ideas are broadly applicable to varying understandings of "big data".

SECTION 1: YOUR BIG DATA RESEARCH STUDY

The motivation and audience for a research study can frame and shape crucial study decisions regarding data collection, modeling, and dissemination of findings. Clearly identifying the motivation and audience can help researchers to recognize the potential for introducing bias before even beginning their work. The challenge of addressing these human biases is not unique to big data research, but researchers have raised particular concern about these issues in the context of machine learning and big data research.[1,2] This affords for the important and necessary step of taking intentional steps to address this bias.

1.1 Motivation for Big Data Research Study

The field of health services research (HSR) is organized around the shared ethos of improving health systems and outcomes; however, researchers' needs and motivations can differ from those individuals represented in the data or using the research end products. First, researchers' own understanding of how to prioritize topics or frame questions for investigation (for example) may be shaped by the theories, literature, or practices with which they are most familiar. Additionally, researchers may prioritize their own investment in a specific outcome, also known as motivational or self-serving bias, to the detriment of individuals in the data who are most likely to be impacted by the findings.[23] *Motivational bias* can be particularly hard to eliminate as even with awareness of potential conflicts of interest and attempts to make objective and honest judgments, motivational biases can still distort researcher judgment and decision-making.[23] Bojke L et al. (2021) emphasize the importance of ensuring representation from a range of viewpoints to dilute the effect of motivational bias, which we encourage through both identifying the motivations of individuals in the data as well involving multiple stakeholders (see Section 1.2).[23]

Exhibit 1 provides a template for naming the motivations of researchers, relative to those of individuals represented in the data. The blank lines allow for additions of unlisted motivations. Listed motivations are sourced from the literature.[24–27] *Answer intriguing questions* refers to the desire of both researchers and participants to explore the unknown and contribute to general knowledge.[24,25] Some questions may be of greater interest or importance to researchers than participants or vice versa. *Improve societal systems* focuses on the opportunity for research to influence policy and systems.[26] *Altruism/helping others* includes the desire for researchers or participants to give back to other individuals.[24] *Benefit for self* for researchers may include career advancement, glory, or fame.[25] *Benefit for self* for participants may include improved outcomes or care, such as additional health monitoring or access to new treatments.[24,27] *Financial motivation* refers to opportunities for financial benefits, such as remuneration for a researcher or compensation for the participant.[24,25] Financial motivation is particularly important for big data researchers to evaluate as, historically, a substantial amount of AI and data research funding has come from large technology companies.[28] Further, literature suggests that corporate interests can push research agendas away from the questions that are most relevant to public health.[29]

Understanding the motivations driving individual researchers, or the histories and incentive structures underlying different fields, can prove challenging. Assuming the motivations of individuals or groups represented in the data may also pose challenges; this is especially the case for big data research that is often performed with large, anonymized datasets where data may have been recorded without the direct supervision of a human, and where individuals may never have consented to data collection or usage of their data for research purposes.[4] While these individuals may not have explicit motivations related to inclusion in the study, researchers should still aim to consider and understand their goals, needs, preferences, and values (GNPV) either by reviewing relevant literature or consulting people with relevant lived experience or *context* expertise.[30]

Exhibit 1: Identifying Motivations of Researchers versus Individuals in Data

MOTIVATIONS OF RESEARCHERS	MOTIVATIONS OF INDIVIDUALS IN DATA
☐ Answer intriguing question	☐ Answer intriguing question
☐ Improve societal systems	☐ Improve societal systems
☐ Altruism/helping others	☐ Altruism/helping others
☐ Benefit for self	☐ Benefit for self
☐ Financial motivation	☐ Financial motivation
☐ _____	☐ _____
☐ _____	☐ _____

Sources: Coccia (2018); Fecher & Hebing (2021); Sheridan et al. (2020); Soule et al. (2016)

This is important in an era where individuals may be increasingly concerned about how their data, particularly sensitive data related to race and ethnicity, may be misused.[31] If no motivation or positive impact for individuals in the data can be identified, researchers should earnestly consider restructuring, pausing, or abandoning the study. Researchers should also critically assess how their own history, or the history of fields and disciplines in which they were trained, influences their motivations; to the extent that intended work may cause harm and/or fail to return value to those represented in the data, researchers may similarly want to reconsider their planned approaches.

After completing the template in **Exhibit 1**, researchers should consider the differences in motivations and the potential impacts on the project. **Exhibit 2** provides potential questions to consider. For example, if improving societal systems is important to individuals represented in the data but not to researchers, the researchers could consider adding a dissemination plan that includes translating and communicating their results with policymakers.

Researchers should also consider how external factors influence these motivations. For example, time may be an important external factor; on a project with short, strict deadlines, researchers may make decisions to save time that result in bias such as using a nonrepresentative dataset (as described in Section 6) because there is limited time to collect data.

Exhibit 2: Questions to Ask Related to Differences in Motivations

Item	Questions to Ask Before Proceeding
Answer Intriguing Question	• Is the research question one that both researchers and participants care about? If not, why not? • Can the research question be modified?
Improve Societal Systems	• Is there a plan to communicate findings with: - Policymakers? - Participants?
Altruism/Helping Others	• Is there sufficient benefit for the individuals in the data? • Does the study rely too heavily on altruism? • Are participants being taken advantage of?
Benefit for Self	• Do the potential benefits to the researchers outweigh the potential benefits to individual participants?
Financial Motivation	• Does financial motivation impact the balance of power? • Are participants sufficiently compensated for their data?

Clearly identifying and addressing differences in motivations can help researchers better plan studies that have positive impacts for individuals included in the data, thus reducing opportunities for bias and harm. Researchers should consider intentionally adding a dissemination plan or impact statement to their workflow to ensure that they are meeting ethical due diligence.[32]

1.2 Consider Your Audience and Use Cases When Developing Your Big Data Research Study

One reason that bias in big data studies causes concern is that big data tools and algorithms are rapidly being adopted for use in high-consequence settings such as personalized medicine, pandemic planning, and inference related to drugs and alternative treatments supported by discovery analytics.[2] Insufficient consideration of audience needs and use cases can lead to biased outcomes and harm

individuals.[33] **Exhibit 3** provides a template for developing a dissemination plan by identifying the audience, anticipating how that audience might use study findings, and creating a research product that is clear and comprehensive enough to meet audience needs.

The *Audience* includes stakeholders who are affected by the research and those who might find the research valuable.[34] These could include participants in the research or individuals otherwise included in the data as well as other researchers and local, state, or national policymakers. The *Dissemination Channel* column should include details on the preferred communication channels of the audience. This may include traditional research channels (e.g., journals, conferences) and; it can also include innovative approaches aligned with the information needs or communication preferences of the audience (e.g., news channels, direct communication with political staffers and legislators).[34,35] *Involvement in Research Cycle* details the times at which the audience is involved in the research. For example, some argue that research participants or patient representatives should be involved throughout the project, particularly when participants have different perspectives than researchers. Norori et al. (2021) called for the use of participatory science, including by patient groups, in the development of novel AI algorithms.[22] This type of participation may not only increase participant comfort with the use of sensitive attributes such as race and ethnicity; their meaningful involvement can also inform the creation of algorithms that more directly reflect their lived experiences and improve their outcomes. Thus, those algorithms can be designed to yield insights more directly and sustainably beneficial in real-world settings. The *Messaging* column should include how messaging will be framed or the mechanism through which it will be delivered such as a paper, visualization, or announcement.

Creating a formal dissemination plan at the outset of a project helps assign roles, structure activities, and allocate funding for dissemination of findings.[34] This is particularly important for addressing the "translation gap" that occurs when research is not communicated in a way that translates to policy changes.[35] Brownson (2018) highlights that successful dissemination must include stakeholders; be active rather than passive; tell a story that evokes emotion, interest, and usefulness; include findings that are understandable, concise, unbiased, and preferably localized; and account for the specific needs and culture of policy audiences.[35] Effective communication of findings with diverse audiences can prevent harm and improve the positive impact of research.

Exhibit 3: Template Identifying the Audience and Dissemination Plan

Audience Who will be impacted by the findings or find them valuable?	**Impact & Use Cases** How might this audience use or benefit from these research findings?	**Dissemination Channel** What methods of communication does this audience use?	**Involvement in Research Cycle** How will this audience be involved in research (if at all)?	**Messaging** What will be communicated with this audience?
Ex: State legislators	Insights may help to inform state-level policymaking	Relevant information is shared by staffers	Communicate findings with staffers	Prepare a white paper or policy brief with findings

Case Study 1: Maine Organizations Use Stakeholder-Oriented Practices to Improve Care for Patients with Intellectual/Developmental Disabilities

One of the Agency for Healthcare Research and Quality (AHRQ) Impact Case Studies highlights the partnership between the Maine Developmental Disabilities Council (MDDC) and two patient safety organizations to improve care for patients with intellectual and developmental disabilities (IDD). Recognizing that patients with IDD may avoid health care settings due to uncomfortable interactions with primary care practitioners, who often lack experience treating IDD patients.[39,40] Motivated to address this challenge and improve societal systems, researchers sought to improve the health care experience for IDD patients. Data on historical incidents between IDD patients and physicians indicated that both patients and physicians wanted to improve the IDD patient experience. The researchers facilitated conversations among patients, parents of patients, and clinicians using a "Safe Table" protected forum approach where participants could feel comfortable discussing their experiences. Findings were collated into a patient safety brief and prompted the inclusion of IDD needs in training. Researchers highlighted that the focus on conversational communication with doctors, rather than mandates, was important to the success of this work.[41]

ACTION ITEMS: YOUR BIG DATA RESEARCH STUDY

- Complete the **Exhibit 1** template to identify and compare motivations of researchers and individuals included in the data.
- Complete **Exhibit 2** to assess the impact of differences in motivation and address potential consequences.
- Use the **Exhibit 3** template to initiate a formal dissemination plan, including identification of audience(s) and how they may be involved in or impacted by the research.

SECTION 2: BIG DATA SOURCES AND DOCUMENTATION

Researchers have an opportunity to make choices during the data collection, sourcing, and documentation phases that can reduce potential for bias and harm. Health care big data may come from various sources including administrative claims records, clinical registries, electronic health records, biometric data, patient-supplied data, medical imaging, genomics data, biomarker data, and large clinical trials.[2,4] Relative to other disciplines, big data in health care settings may be more likely to be collected according to a specific protocol and thus may be relatively structured.[4] Health care big data may be relatively difficult to access due to legal requirements or the risk of misuse, and data may be costly to collect if the involvement of personnel or expensive instrumentation is required.[4] Several sources of uncertainty may impact health care big data including measurement error, missing data, or quantitative information buried in textual reports.[4] Key decisions throughout the process of collecting and finding data can improve the quality of the dataset and mitigate bias.

2.1 Collecting and Finding Big Data

Researchers often must decide between collecting primary data, finding secondary data, or combining these two approaches. Collecting primary data could include deploying a survey, scraping data from a website, or making measurements. Finding secondary data may include discovering existing datasets, using government data, or obtaining permission to access health records.

Collecting primary data, often considered a more traditional research approach, offers many advantages. These include the ability to better establish alignment between the data collected and a focus on ensuring ethical and responsible practices. Section 3 explicitly describes the challenges of collecting race and ethnicity data, many of which may be more directly addressed when collected as primary data. Yet, collecting primary data can be expensive and often requires specific expertise. Primary data also do not guarantee high-quality data, as primary data collection can still be unrepresentative, flawed, or biased. Thus, even when working with primary data sources, researchers must still seek to make a high-quality dataset (see Section 2.2).

Secondary datasets have exploded in popularity with the increasing accessibility of data collection, decreasing costs of data storage, and general interest in using data for decision-making. Many publicly available datasets can now be easily found with search engines like Dataset Search from Google and data.gov from the U.S. government.[42,43] **Exhibit 4** provides a template for identifying risks related to secondary data use as with, for example, the major concerns arising when the purpose of the original data collection is substantially different from the purpose of the research study. Pointing out that routinely acquired medical data differs from data collected primarily for research purposes, Martin-Sanchez et al. (2017) provide an example that the failure of a clinician to record a patient's disease does not mean the patient does not have that disease.[44] A critique of using electronic health records (EHRs) for big data research is that EHRs are a byproduct of the health system in which a substantial portion of the U.S. population remains uninsured or uses health care rarely; thus, EHR data are not sufficiently representative.[45] This may contribute to sampling bias and inhibit generalizability of findings.

Researchers should also consider the methods and individuals involved in the data collection process.[46] For example, the collection of data by the government may be motivated by legal requirements or specific protocols and thus resulting datasets may be relatively structured. Alternatively, data collection in health care settings may be intended primarily to meet the information needs of practitioners and insurers; thus, these datasets may omit information not required to make a diagnosis or determine an insurance charge code.[44,47] Data collection practices may also impact the likelihood of typographical or other (such as documentation or measurement) errors in the data, as well as the degree of missingness. Veracity, or the ability to confirm the accuracy of a dataset, is a substantial concern in big data studies as researchers often do not have the resources to independently confirm the accuracy of each record.[48] Inaccurate or missing datasets can lead to biased outcomes.[47] For secondary data collection, timeliness is important because (unless data are needed from a particular time frame) using more recent data can ensure conclusions are current and relevant. Finally, knowing about the chain of custody or provenance of the dataset is also crucial for understanding how the data may have changed over time.[49]

Data collection methods should be scrutinized, particularly if the research was conducted in loose regulatory environments.[34] The ethics review of human subjects research generally requires consent of participants, though this is considerably less common for big datasets generated in non-research settings.[50,51] Researchers should consider the implications of using data in a manner that subjects may not have agreed to, if they had been asked for consent.[50] Researchers should be aware of and seek to address their own biases related to the dataset, such as having prior knowledge of the dataset contents and therefore a bias toward seeking a particular answer. Similarly, researchers should consider the extent to which institutional norms, standards, protocols, or other factors may have biased or influenced data collection processes. These biases can shape the manner in which research is carried out. If appropriate, researchers can pre-register a study with a service such as clinicaltrials.gov; studies preregistered with pre-specification of outcome variables are generally considered to have a higher level of credibility.[47]

Exhibit 4: Template for Evaluating Risks in Secondary Data Analysis

Question	Response
What was the purpose of the original study?	
Who was responsible for collecting the data?	
When was the data collected?	
How was the data collected?	
Did subjects provide consent?	
Is there evidence of researcher bias, such as prior knowledge of the dataset contents?	

Sources: Johnston (2014); Baldwin et al. (2022)

Carefully making decisions when finding or collecting data can improve the quality of the data, contribute to better transparency, and prevent opportunities for perpetuating bias.

2.2 Characteristics of a High-Quality Big Dataset

A high-quality dataset is a cornerstone of transparent and reproducible results, and it is especially crucial for preventing bias in big data and machine learning studies.[52] Low-quality data can yield inaccurate results that may even result in retraction of findings.[53] **Exhibit 5** provides a template for evaluating data quality drawn from the Clinical Information Quality (CLIQ) framework for digital health and the Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML) framework.[52,54] Researchers can enter answers into this template and assess the quality of their dataset.

Exhibit 5: Template for Evaluating Data Quality

To what extent are the data…	Response	To what extent are the data…	Response
Accurate?		In a desired format?	
Complete?		Appropriately secure?	
Interpretable?		Timely?	
Plausible?		Maintained?	
Consistent?		Of sufficient size?	
Representative?		Documented?	

Sources: Al-Zaiti et al. (2022); Fadahunsi et al. (2021),

A high-quality big dataset should be accurate, meaning that the contained information is correct. Accuracy is not trivial and should not be presumed, particularly for medical big data such as lab results or wearables data. Medical data collection is often subject to measurement error or uncertainty, yet few studies use methods to investigate or correct for it.[55] It is a myth that a large number of observations alone can compensate for measurement error.[56] In large-scale physical activity monitoring with smartphones and wearable devices, unconscious bias on the part of developers can further exacerbate these measurement

errors as consumer wearable devices appear (for example) unable to detect steps in people with slow, short, or non-stereotypical gait patterns; this includes but is not limited to women, obese individuals, and individuals from different ethnic groups.[57] Measurement errors can differ by race and ethnicity, such as the finding that pulse oximeter errors occurred more often for Black and Hispanic COVID-19 patients than for white patient counterparts, which may have contributed to disparate patient outcomes.[58] A smaller study with accurate measurement data may be preferable to a big data study with many measurement inaccuracies.[59]

A high-quality big dataset should ideally also be complete, with little missing data for each attribute. Completeness is especially important when using machine learning algorithms that often require complete records with no missingness. While no clear standard has been set for a big data missingness threshold, Emmanuel et al. (2021) tested imputation methods for machine learning for up to 20% missing in a given attribute.[60] If data are missing, they should be handled with appropriate imputation or removal methods (see Section 5). The data must be interpretable (i.e., able to be understood) and consistent (i.e., routinely collected and recorded following accepted standards for the field). The importance of representativeness in big data is contextual and depends on the project.[59,61] While unrepresentative datasets can lead to generalizability issues, representativeness may not be required for associative analyses.[62] See Section 6 for more on representative datasets.

Assessing the fit for purpose of a high-quality big dataset may require taking steps (e.g., cleaning the data and generating appropriate features) as needed to transfer the data into the required format. A high-quality big dataset is often, though not always, recently collected and regularly maintained after collection.[54] The dataset must be of sufficient size, particularly for machine learning research where guidance suggests there should be at least five occurrences of the outcome being predicted (known as *positive labels*) per input variable (known as an *input feature*).[54] Finally, a high-quality big dataset should be well documented. Examples of data documentation templates include Datasheets for Datasets,[63] the Data Biography,[64] and the Dataset Nutrition Label.[65]

Data security and privacy are critical issues in big data research.[1,4,6] While a full description of best practices is beyond the scope of this text, several key themes related to security and privacy for health care big data are worth mentioning:

- Health care big data may be subject to data protection or privacy laws that vary by country, such as HIPAA in the United States.[1]
- Data security precautions must be integrated throughout the research lifecycle to prevent data breaches, protect important assets, and satisfy compliance requirements; particular attention may be needed for streaming or cloud-based big datasets.[6]
- Various deidentification, anonymization, and differential privacy approaches for big datasets exist or are in development to support privacy-preserving machine learning; however, many still have limitations or remain vulnerable to attack.[6]

In completing **Exhibit 5**, researchers should assess what "Appropriately Secure" means in the context of their dataset and consult with appropriate technology personnel as required.

Researchers should assess their responses to **Exhibit 5** and compare with the above information to determine the quality of their dataset. If the dataset is determined to be of poor quality, researchers should pause and consider options to improve their data before proceeding.

Case Study 2: Quantifying the Impact of Data Quality on an Outcome Measure

In a 2017 study titled *Quantifying the Effect of Data Quality on the Validity of an eMeasure*, Johnson et al. artificially generated data quality issues affecting EHR data. They then calculated the impact of these issues on the calculation of patients who had a catheter removed within 48 hours of surgery (given the best practice to remove a catheter soon after surgery to prevent infection). This measure is calculated using a variety of variables and requires that patients meet a series of inclusion criteria. For their study, Johnson et al. (2017) modified up to 10% of the records by entering null or inaccurate values for up to 15 variables per individual; this modification was intended to simulate the messiness of real EHR data. In many cases, the modified data no longer met inclusion criteria. For example, every 1% reduction in data quality of birth date and admission type caused a 1% increase in the number of patients who were excluded from the calculation who should have been included. While 1% reduction may seem small, it can have a sizable effect on a big dataset and may be magnified if other calculations are also impacted. These findings highlight the importance of creating high-quality datasets and interrogating the quality of secondary datasets (especially for EHR data), as both can significantly affect research results.[66]

ACTION ITEMS: BIG DATA SOURCES AND DOCUMENTATION

- If considering use of secondary big datasets (e.g., government data, publicly available health care data), complete **Exhibit 4** to assess the potential risks.
- For both primary and secondary big datasets, complete **Exhibit 5** to assess the quality of the datasets.
- If the dataset is of low quality, consider whether a smaller study with higher quality data would be possible.

SECTION 3: RACE AND ETHNICITY IN BIG DATA

Though new methods are being proposed, measuring and addressing racial bias in big data research still typically requires using race and ethnicity data.[67] However, fairness practitioners report that challenges to accessing high-quality race and ethnicity data create significant barriers for implementing fairness techniques.[31] Although some access issues are related to legal regulations like the General Data Protection Regulation (GDPR) and corporate policy, other challenges are related to the methods used to collect race and ethnicity data.[31] In this section, we comment on the importance of using race and ethnicity data and challenges to collecting these data that are most salient for big data researchers.

3.1 Why Considering Race and Ethnicity is Important

As noted in the introduction, big data sources offer considerable promise but are neither neutral nor objective.[15] Gillborn et al. (2018) point out that even in big data research, all data are manufactured and all analysis is driven by human decisions.[15] Big data research is susceptible to reproducing existing biases and perpetuating existing disparities among racial and ethnic groups, as documented in a robust body of evidence.[68] Racial health disparities can emerge not only from issues related to the collection or quality of race and ethnicity data but also from failing to be (1) thoughtful and intentional about how those data are used, or (2) aware and open about limitations in the kinds of causal inferences that can be drawn using those data. While the body of research on the adverse effects of racism is growing, Williams et al. (2019) called for researchers to give more explicit attention to racism and discrimination in health-related research.[69] Disparities are also a consequence of both overt and implicit racism, including the three levels of racism defined by the Centers for Disease Control and Prevention (CDC) style guide:

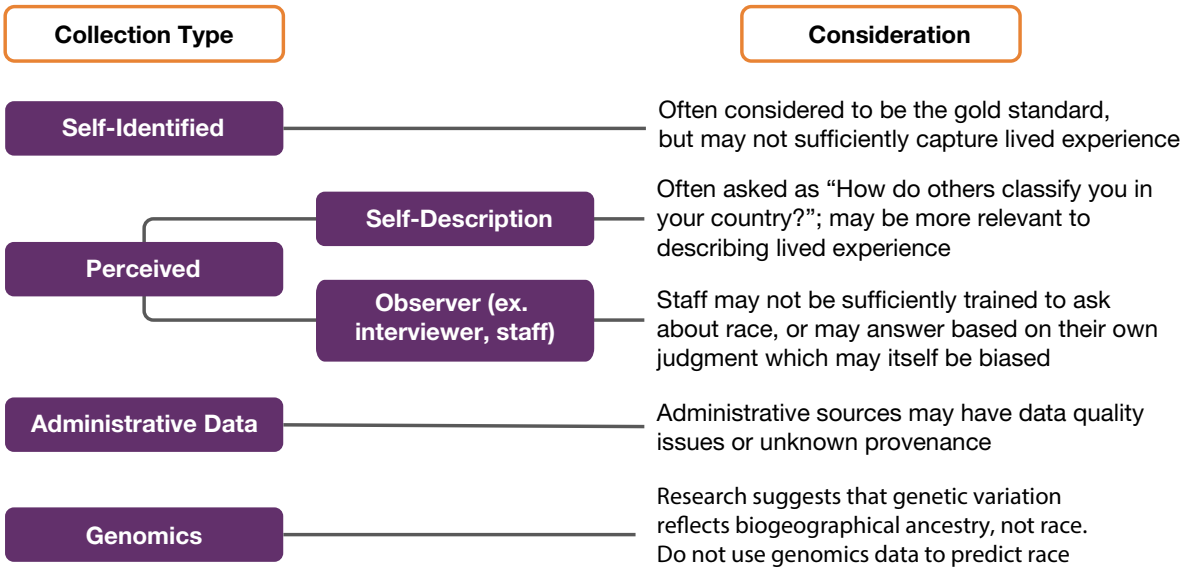
- **Systemic, institutionalized, and structural racism:** "Structures, policies, practices, and norms resulting in differential access to the goods, services, and opportunities of society by 'race' (i.e., how major systems—the economy, politics, education, criminal justice, health—perpetuate unfair advantage).[70]
- **Interpersonal and personally mediated racism:** "Prejudice and discrimination, where prejudice is differential assumptions about the abilities, motives, and intents of others by 'race,' and discrimination is differential actions towards others by 'race.' These can be either intentional or unintentional."[70]
- **Internalized racism:** "Acceptance by members of the stigmatized 'races' of negative messages about their own abilities and intrinsic worth."[70]

Race and ethnicity are social constructs with no biological or scientific meaning, but the experience of all three levels of racism has a direct impact on livelihood and outcomes, particularly as it relates to health.[71,72] Historically, race-based medicine has treated Black and Brown people as distinct from white people, but research shows that distinct physical characteristics and genetic differences better correspond to geography, not race.[10] Medical equations that traditionally have included race as a biological explanatory variable, such as the calculation of risk for chronic kidney disease (CKD) using estimated glomerular filtration rate (eGFR) or for diabetes using body mass index (BMI), have potentially harmed patients.[10] Researchers are called to abandon race-based approaches and turn to race-conscious methods.[10] Race and ethnicity should be used to assess for experiences of discrimination and account for the impact of racism.[10] Big data analyses should include assessment of bias by race and sex to ensure that the big data approach is not itself sustaining or exacerbating race and ethnicity disparities and inequities.[15]

3.2 The Challenges of Collecting Race and Ethnicity Data

Many methods for assessing bias based on race and ethnicity require the collection of race and ethnicity data.[31] Although big data researchers may not be involved directly in data collection, they should still seek to familiarize themselves with the challenges of race and ethnicity data collection so they can acknowledge the potential limitations of their datasets and resulting findings. The American Medical Association (AMA) and the academic Journal of the American Medical Association (JAMA) now explicitly provide guidance that, "The Methods section [of research papers] should include an explanation of who identified participant race and ethnicity and the source of the classifications used."[11] This guidance recognizes both the importance of collecting race and ethnicity data, as well as the challenges of collection. **Exhibit 6** highlights important considerations for different types of race and ethnicity data collection, which are also discussed in further detail below.

Exhibit 6: Considerations by Type of Race and Ethnicity Data Collection



The U.S. Office of Management and Budget (OMB) indicates that preference for self-identified or self-reported race and ethnicity is the preferred method and, as such, this is frequently used in medical research, often as a population descriptor.[73] Research has consistently documented differences in health outcomes by self-identified race and ethnicity, even after accounting for factors such as socioeconomic status, health behaviors, and health.[74] Yet a number of scholars argue that self-identified race and ethnicity does not sufficiently represent the lived experience of race in a racialized society, and that socially assigned race is an important factor to consider in health research.[74–77] This can be asked as a question

of perception (i.e., "How do other people usually classify you in this country?", as asked on the Behavioral Risk Factor Surveillance System) or through observation (i.e., race and ethnicity are assigned by interviewer or health care staff member).[74] Research has found that race observed by another person, such as a health care worker observing the race of a patient, may differ substantially from self-identified race and should be used with caution.[78] Similar caution should be taken when using administrative data, such as claims data, to impute race. If the source of the racial assignment is unknown, it is important to note this limitation on interpretation and generalization.[79] Finally, in recent years, increased attention has been paid to using genomics data to impute or override self-identified race or ethnicity. Research suggests that genomics data and related genetic variation reflect biogeographical ancestry, and not racial differences.[80] Genetic history is not equivalent to race and ethnicity; therefore, it should be treated as a distinct concept.

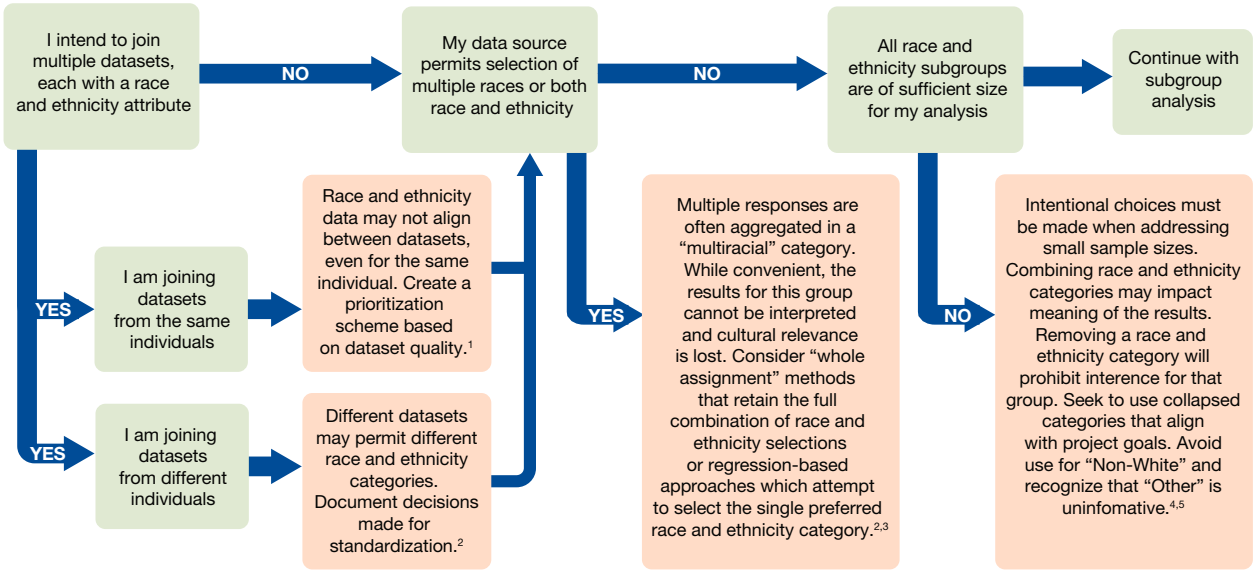
While important, the collection of race and ethnicity data can be especially challenging in health care settings due to challenges such as the lack of standardization for data entry and lack of training among staff. AHRQ suggests that explicitly expressing the motivation for collection of race and ethnicity data, and advocating for valid collection mechanisms with staff and organizational leadership, may improve the quality of race and ethnicity data collection.[81] Patient-facing tools that record race and ethnicity information before, during, or after health care encounters are especially promising.[82]

Big data researchers using federal health data should be aware of the ways in which U.S. government agencies have been called on to improve collection of race and ethnicity data.[83] For example, one recommendation is for the Center for Medicare and Medicaid Services (CMS) to include race and ethnicity categories on Medicare Part C and D enrollment forms.[83] Traditionally, CMS has obtained this data from the Social Security Administration (SSA) which sources data from birth certificates. However, the race and ethnicity categories on birth certificates are limiting, as they predate the 1997 OMB Directive that defined the current standard race and ethnicity categories. SSA is also unable to provide race and ethnicity for beneficiaries born after 1990 and Medicare data are known to be less accurate for beneficiaries identified as American Indian/Alaska Native, Asian/Pacific Islander, or Hispanic.[84] CMS is actively working to address this challenge; in 2022, they collected public comments regarding the intention to pilot inclusion of detailed race and ethnicity categories on Medicare Part C and Part D enrollment forms.[84] CMS intends to use these data to track enrollment and reduce and eliminate health disparities.[84] Other government initiatives related to the collection of race and ethnicity data include identifying opportunities to improve race and ethnicity collection for Veterans Affairs (VA) patients (see Case Study 5),[85] increase validity of race and Hispanic-origin in death certificates in the National Vital Statistics System (NVSS),[86] and address missing race and ethnicity data related to the COVID-19 pandemic.[87]

3.3 Key Analysis Decisions for Race and Ethnicity Data

Researchers often make several key decisions when performing analysis with race and ethnicity attributes. While these decisions are not limited to big data settings, they are critically important as big data researchers generally do not have the opportunity to review all records individually and assess the full impact of what may seem like straightforward decisions. **Exhibit 7** summarizes these choices.

Key Decisions in Preparing for Analysis with Race and Ethnicty Data



Sources: Peltzman et al. (2022); Defining Categorization Needs for Race and Ethnicity Data, AHRQ (2022); Liebler and Halpern-Manners (2008); Ross et al. (2020); Flanigan et al. (2021)

Researchers should first consider whether they need data from multiple datasets, with each dataset including a race and ethnicity variable. If researchers are seeking to join disparate datasets with the same individuals, it is both possible and likely that there may be discrepancies where the race or ethnicity variables for the same individual are different between datasets.[88] Researchers should consider the quality of the datasets being joined and determine a scheme for prioritizing which race and ethnicity response to retain for analysis.[89] If researchers are joining datasets with different individuals, they should familiarize themselves with different race and ethnicity categorization schemes. The Office of Management and Budget (OMB) maintains the most common race and ethnicity definitions, but also encourages additional granularity in race and ethnicity subgroup analysis beyond the minimum standard set.[90] The U.S. Department of Health and Human Services (DHHS) suggests a standard disaggregation for the Hispanic ethnicity and the Asian and Native Hawaiian/Other Pacific Islander categories.[90] AHRQ provides even more detailed recommendations for specific subgroups, with rollup to the OMB standard categories in most cases.[91] Researchers should discuss and document decisions made to standardize a race and ethnicity attribute for multiple joined datasets.

Next in **Exhibit 7**, if the data source(s) permitted selection of multiple races or both race and ethnicity, researchers should determine how they plan to perform analysis. Individuals with multiple responses are often aggregated into a "multiracial" category. While convenient, "multiracial" lacks cultural relevance as it does not describe a subgroup with a shared identity (instead lumping multiple disparate subgroups into one category) and cannot be interpreted.[92,93] Researchers should strive to use a "whole assignment" approach that retains the full combination of race and ethnicity for analysis purposes (e.g., *American Indian or Alaska Native and Black or African American* as a category, instead of including these individuals as "multiracial"). Small studies typically do not have enough individuals to utilize this type of granularity; this approach may be particularly promising in big data research with representative datasets. If a single race is still required, researchers should consider regression-based methods which have also been successfully used to suggest the single race that an individual would select to best describe themselves.[93]

The last key decision in **Exhibit 7** is whether the race and ethnicity groups are of sufficient size for analysis. Sufficient sample size may be determined by statistical or publishing thresholds. When a subgroup is small, a researcher may be called upon to make adjustments.[94] This could include data manipulation, such as rolling up race and ethnicity responses into larger categories.[94] Researchers should carefully consider decisions to combine race and ethnicity categories, perhaps using suggestions in the resources described in the above paragraph on considerations for joining datasets, and ensure alignment with the research question.[94] Guidance from the *Journal of the American Medical Association* (JAMA) says to avoid use of "Non-white" or study designs that compare white versus "Non-white" groups.[11] Additionally, JAMA recommends careful consideration of using an "Other" group label as the results are uninformative and the label may be considered pejorative. "Other" should be used sparingly, not for convenience, and all subgroups included in "Other" should be listed.[11]

After reviewing key analysis decisions, authors should perform subgroup analysis with race and ethnicity data. Example analyses could include reporting outcomes by race and ethnicity and assessing potential associations or relationships of race and ethnicity with other variables. Subgroup analysis is common in clinical trials, though improvements have been suggested for race and ethnicity reporting.[95,96] Subgroup analysis can provide important information to researchers as well as hospitals and care systems by helping to identify local disparities, develop patient-centered resources, and drive decisions on where to invest and deploy resources.[97,98] Disregarding opportunities for disaggregation or inattentively aggregating race and ethnicity during subgroup analysis can conceal important findings. For example, one study found that aggregation of Asian-American subgroups masked meaningful differences in health and health risks among Asian ethnicities.[99] Researchers should pay careful attention to decisions made regarding analysis of race and ethnicity data, and the impact those decisions may have on findings.

3.4 Assessing Algorithmic Bias in Big Data Models by Race and Ethnicity

A research project may include the development of models and algorithms, particularly in big data research, and these models themselves may introduce bias. Algorithmic bias is defined as bias that is introduced during the modeling process.[100] High quality race and ethnicity data are generally important to assessing algorithmic bias, although researchers are exploring innovative approaches that do not require sensitive attributes.[31] A bias audit can be used by researchers to assess bias during the modeling process, including modeling with big datasets.[101] Liu et al. (2022) proposed a medical algorithm audit that includes items previously discussed in this text, such as motivation and stakeholder identification as well as more technical testing and algorithmic review.[102] Technical considerations include exploratory error analysis of mistakes made by the model, subgroup analysis related to possible confounding and stratifying factors, and adversarial testing for simulating how the model behaves to changes in input data.[102] The Algorithmic Bias Playbook (Obermeyer, 2021) provides a clear framework with health care examples for how to define, measure, and mitigate racial bias in live algorithms.[103] Huang et al. (2022) suggests opportunities for inclusion of specific bias techniques into the machine learning development workflow, including resampling existing data during preprocessing, adversarial debiasing during model processing, and modifying decision thresholds in post-processing.[104] The authors go on to discuss several examples of racial bias evaluation in the clinical machine learning literature and strategies used to mitigate this bias.[104] Specific bias techniques and approaches will likely depend on the context and needs for a project; researchers should identify and select appropriate race and ethnicity bias assessment and mitigation techniques appropriate to their workflow.[100,105]

Case Study 3: A Real-World Example of Improving Collection of Patient Race and Ethnicity Data

A 2022 study titled *Improving Patient Race and Ethnicity Data Capture to Address Health Disparities: A Case Study from a Large Urban Health System* provides insights from a real-world approach to improving race and ethnicity data collection at a hospital in New York. This hospital had approximately 60,000 patient visits per year. Prior to the initiative, the health system did not systematically collect race and ethnicity, but rather requested patients enter their own information. This resulted in substantial missingness and inaccurate responses. The study details a five-phase systematic patient registration data collection improvement process (PRDCIP) including (1) assessment and evaluation (2) infrastructure modification (3) training and education (4) implementation and response to results and (5) acknowledging limitations and lessons learned. This innovative approach included training and discussion with front line staff on the importance of improved data capture for addressing health care disparities and how to communicate with patients regarding the strict confidentiality of their information. This study resulted in a 76% improvement in the completeness of race and ethnicity data, and this newly improved dataset is being used in core equity dashboards across the health system.

ACTION ITEMS: RACE AND ETHNICITY IN BIG DATA

- Consider the importance of race and ethnicity in your research study.
- Review **Exhibit 7** and assess the potential impact and opportunities to mitigate choices made when analyzing race and ethnicity data.
- Identify and select appropriate bias assessment and mitigation techniques for your model or algorithm, if applicable.

SECTION 4: USE OF PROXY VARIABLES

Proxies are variables used in place of an unobservable or immeasurable quantity of interest.[107] Proxies are strongly correlated with the variable of interest, even if they are not directly relevant to the question at hand. Proxies raise three major concerns related to racial bias in big data studies: wrongful proxy discrimination, unintentional proxy discrimination, and inappropriate use of race and ethnicity as a proxy for some other variable.

Wrongful proxy discrimination is the known use of a proxy variable to substitute for a sensitive attribute that is not permitted in modelling. For example, race and ethnicity attributes are not permitted in credit models. However, zip code may be permitted, even though zip code is often considered to be a proxy for race and ethnicity.[108] Not all proxy discrimination is unlawful, and in many cases, there may be valid reasons to use a proxy. However, the explicit use of a proxy like zip code to substitute for race when the latter is not authorized for use is considered wrongful and should be avoided.

Unintentional proxy discrimination occurs when a variable serves as a proxy for a protected attribute like race or ethnicity, even when the model developer did not explicitly intend for this variable to serve as a proxy. The equity concept of "Fairness through Unawareness" suggests that a model cannot be unfair if it is unaware of a protected attribute like race or ethnicity, however the presence of unintentional proxies violates this assumption.[109] The unintentional use of a proxy for a protected attribute can lead to discriminatory outcomes, as illustrated in **Case Study 4**. This is particularly concerning in big data and machine learning studies where even if sensitive attributes like race and ethnicity are not included, an algorithm can still leverage the scale of the data with a proxy variable to reconstruct the unused variable.[110] This highlights the importance of interrogating not only the potential for bias in the data, but also in the modeling itself. Researchers have proposed approaches for detecting unintentional proxies in linear regression and machine learning models.[111,112] Three approaches researchers can use to detect unintentional proxies include:

- Advocating for collection of sensitive attribute data, even if not used in the model, to allow for detection of unintentional proxies.
- Evaluating basic correlations between protected sensitive attributes like race and ethnicity and variables in the model.
- Considering more complex analyses, particularly in high-risk use cases, to detect potential bias resulting from unintentional proxies. Due to the potential presence of unintentional proxies, it is not sufficient to say definitively that an algorithm is unbiased simply because it does include a protected variable like race or ethnicity.

Finally, researchers should be aware of the inappropriate use of *race or ethnicity itself as a proxy*. For example, researchers may be interested in understanding the relationship of some health outcome with characteristics such as socioeconomic status, housing insecurity, or health behaviors (e.g., diet). However, race and ethnicity data are often collected more frequently than these other characteristics; as such, researchers may use race and ethnicity as a proxy based on assumptions that people of the same race or ethnicity are likely to have shared or similar experiences. These are substantial assumptions that may not stand up to scrutiny, could be offensive, and may perpetuate a culture of stereotyping or even racism.[113] Yet, these assumptions have historically shown up in preclinical lectures and clinical vignettes used in medical teaching settings as well as race-based algorithms used in medical practice.[10,114] Researchers should gain awareness of race-conscious alternatives and seek to utilize attributes most closely aligned with the research question rather than use race and ethnicity as a proxy.[10,115]

Case Study 4: Unintentional Proxy Discrimination Leads to Racial Bias in Health Care Risk Prediction Algorithm

In 2019, a comprehensive study titled *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations* revealed that a major health care-risk prediction algorithm demonstrated racial bias against Black individuals. The algorithm was used by hospitals and insurance companies to identify which chronically ill patients would receive access to specially trained nursing staff and extra primary-care visits for closer monitoring. This algorithm did not include race or ethnicity as an input variable and instead used other variables to make predictions. One of these variables was previous patient health care spending, which ended up being a proxy for race; even when patients across racial groups spent the same amount, Black patients' costs were generally to cover more intensive procedures like emergency visits for diabetes or hypertension complications. This inappropriate use of a proxy masked that the Black individuals were actually sicker, even though their costs were the same, which resulted in Black individuals receiving lower scores than they should have received. This case study highlights an important finding: even if an algorithm does not include race or ethnicity, it can still be discriminatory if there is an unintentional proxy.[116]

ACTION ITEMS: USE OF PROXY VARIABLES

- Eliminate any instances of wrongful proxy discrimination.
- Use simple correlations or more complex methods to assess potential unintentional proxies in the big dataset.
- Consider if race and ethnicity are being used as an inappropriate proxy and if a different attribute would be preferable for analysis or modeling.

Section 5: Data Completeness

Data completeness describes the extent to which all data are available. Missing data, especially missing race and ethnicity data, requires careful consideration and intentional decision-making. Race or ethnicity data missingness may be especially common in administrative datasets and can occur for various reasons including individual preference, health care provider actions, and administrative policies.[117] Race and ethnicity missingness in big health data gained attention during the COVID-19 pandemic when policymakers sought to quantify disparate impacts, with studies highlighting the importance of attention to imputation methods.[87,118] Race and ethnicity data are crucial for identifying and understanding health disparities and addressing bias in algorithms, so it is important that the data are accurate and comprehensive.[84,119]

Several approaches (summarized in **Exhibit 8**) exist for addressing race and ethnicity data missingness. It may be tempting to remove records with missing race and ethnicity data, especially if using algorithmic approaches that require complete records for all variables. However, race and ethnicity data are likely not missing at random; so, while removing these records may be easy, it may also create a biased dataset that is more likely to be missing individuals from underserved populations.[117,120]

Exhibit 7: Options for Addressing Missing Race and Ethnicity Data

Approach	Remove Records	Imputation	Machine Learning	Linkage
Description	Remove records with missing race and ethnicity data	Impute missing values with options like single regression, hot-deck, MICE, BISG	ML missing data approaches, such as extracting relevant race and ethnicity info with NLP	Link or augment data with information about missing data from other data sources
Advantages	Easy to implement	Keeps all records, prevalent in literature, and improved accuracy	May have better accuracy than other methods; growing research awareness	Can be considerably more accurate if external source is high quality
Concerns	Race and ethnicity likely not missing at random; more likely to drop records from underserved populations and create a biased dataset	Perhaps challenging to implement; approach may have assumptions or be less accurate for some subgroups (ex. BISG)	Challenging to implement; requires sufficient data and computational resources; should be explainable and validated	Requires acquiring and assessing additional data; linkage may require prioritization of validity of data

Sources: Lines (2021); Randall et al. (2021)

Imputation is the act of filling in a missing value with a best estimate.[121] Imputing race and ethnicity is different from imputing other variables, as it introduces a number of ethical considerations. Some practitioners disapprove of imputation or inference of sensitive attributes as it can introduce both privacy risks and dignity concerns.[31] **Exhibit 9** highlights five ethical risks that the Urban Institute suggests researchers should consider before imputing race and ethnicity data.[120] The related questions can help researchers impute race and ethnicity data in alignment with the values and needs of the individuals likely to be most impacted by the work.

Exhibit 8: Ethical Risks of Imputation for Missing Race and Ethnicity Data

Ethical Risk	Questions to Ask Before Proceeding
Excluding people and communities of color from ownership of their data and from decisions on research process and methods	• Does the research include direct engagement with members of the community? • If this is not possible, have researchers attempted to incorporate the perspective of opportunities by asking for guidance from the data collector or collaborating with researchers more proximate to the affected communities?
Violating individual informed consent	Generating race and ethnicity values with advanced analytical methods may override an initial refusal to provide this data. • Is this ethical? • When asking for consent, have researchers clearly communicated how data will be used?
Compromising individual privacy or confidentiality	• Have researchers addressed concerns related to re-identification, such as by using aggregation, synthetic data, or privacy protection measures? • Has a privacy impact assessment been completed?
Producing inaccurate estimates and misleading conclusions	• Have researchers calculated and clearly communicated the degree of uncertainty in findings that result from race and ethnicity data imputation?
Generating data for purposes that harm people or communities of color	• Have researchers sufficiently addressed the concern that imputed race and ethnicity identifiers could be weaponized against individuals from communities of color?

Sources: Randall et al. (2021)

Various imputation methods exist and have been used with race and ethnicity data.[60] Using the mean and mode of a variable for imputation is an older approach that has generally been replaced by more complex analyses; with single regression, for example, values for other variables that are not missing get used to predict the missing value.[122] Hot-deck imputation is used to randomly select a value from a similar record.[123] However, given the importance of race and ethnicity data, researchers generally advocate for using approaches that better capture both nuance and uncertainty.[122] Examples include multiple imputation with chained equations (MICE) and Bayesian and random forest approaches. A final approach to consider, particularly when there is little or no race or ethnicity data, is Bayesian Improved Surname Geocoding (BISG) or Modified Bayesian Improved First Name Surname (mBIFSG).[124] This imputation approach involves using surnames and location information to predict race and ethnicity. It is important to be aware that BISG may exhibit differing accuracy based on context used and by subgroup (i.e., performing worse for younger and American Indian/Alaska Native subgroups), so researchers should validate findings where possible and note limitations alongside results.[124,125]

Machine learning approaches for addressing missingness in race and ethnicity data have gained increasing attention, as in the case of natural language processing (NLP): whereby derived race data drawn from patient health notes can be used to supplement structured EHR data Sholle et al. (2019 found that this approach to addressing missing race and ethnicity data in patient records led to a 20% increase in documented Hispanic patients and a 26% increase in documented Black patients in a cross-sectional study of EHR data from 16,665 patients.[126] Access to large EHR datasets can also support more complex machine learning research. For example, Kim et al. (2018) proposed Race and ethnicity Imputation from Disease history with Deep LEarning (RIDDLE) which yielded significantly better performance in predicting race and ethnicity than other assessed ML approaches.[127] Machine learning approaches appear to be gaining traction for imputing race and ethnicity, though they generally require large datasets and may be quite complicated to implement.

A final option to consider is using other data sources for linkage or augmentation. Linkage can improve the accuracy of health care data, especially for underserved populations.[128] This option is particularly common with administrative data sources. For example, Espey et al. (2013) linked the U.S. National Death Index (NDI) records with Indian Health Service (IHS) registration records to identify AI/AN deaths misclassified as non-AI/AN deaths.[129] Linkage can be *direct*, via a shared identifier, or *probabilistic*, which entails connecting information from separate sources based on the probability of two records representing the same entity.[120] However, linkage may also require specifying rules (i.e., *If multiple sources include conflicting race data, which should be prioritized?*) and introduce additional considerations (i.e., *Is it appropriate to impute a missing race or ethnicity value with the value of a family member?*). The answers to these questions are generally context dependent, though research suggests that family race is suggestive of (though not equivalent to) individual race.[130] Researchers should recognize that linkage still requires high-quality data, and incomplete linkage can contribute to bias.[131]

The completeness of race and ethnicity data can improve the accuracy and reliability of a study, and researchers should strive to make intentional decisions when addressing missingness of race and ethnicity data.

Case Study 5: Incorporating Data from Multiple Sources to Improve Completeness of Race and Ethnicity

In 2019, the U.S. Government Accountability Office (GAO) made suggestions for the Veterans Administration (VA) improve accuracy for specific minority groups as part of its commitment to addressing health disparities.[132] Related to these recommendations, Hernandez et al. (2019) explored approaches to link existing VA survey data from the Survey of Healthcare Experiences of Patients (SHEP) with administrative data from the VA Corporate Data Warehouse (CDW), VA Defense Identity Repository (VADIR), and Medicare. The study aimed to compare the accuracy of administrative data versus self-reported data, and to develop a hierarchy for combining datasets. This work required harmonizing different racial and ethnic categorizations, used across these datasets. Similar to previous studies, this work found agreement between datasets to be good for white and Black individuals, but generally poor for other groups. Using lower performing datasets would significantly underreport these groups and hinder the ability to accurately detect disparities. To accurately classify race and ethnicity for veterans, researchers recommended that VHA administrators should use SHEP data when available first—then CDW data, then VADIR data, then Medicare data. The researchers used a simple precedence hierarchy, as it did not require a veteran to have data in each dataset and recommended this specific ordering as it prioritized accuracy for minority groups. This case study can illustrate the benefits and challenges of linking multiple datasets to improve the completeness of race and ethnicity.[85]

ACTION ITEMS: BIG DATA COMPLETENESS

- Assess the amount of missing data in the big dataset.
- Determine the best approach (**see Exhibit 8**) for addressing the missing data relative to project context and complexity.
- Consider the ethical implications of imputing missing race and ethnicity data in **Exhibit 9**.

SECTION 6: BIG DATA REPRESENTATIVENESS

Data representativeness is the degree to which a dataset represents the population of interest.[133] Race and ethnicity datasets are unrepresentative when the data do not include sufficient coverage of a particular race and ethnicity group.[134] Even when big datasets are extremely large, they may not represent the underlying population; researchers should not assume that big data supports unbiased estimation of population parameters even when large enough to render standard uncertainty estimates negligible.[62,134,135] For example, white individuals are more likely than other groups to have used a wearable health care device in the previous 12 months and more willing to share their wearable data with a health care provider.[136] Since the generalizability of models depends on representative datasets, it would likely be inappropriate to make broad generalizations from analysis of this dataset to populations not well-represented in wearables data.[137]

Furthermore, increasing data size also introduces the Big Data Paradox where confidence intervals shrink but small biases become magnified.[59,138] This can lead to studies with large sample sizes that have misleadingly narrow confidence intervals around biased results.[138] Bradley et al. (2021) explored an example of the Big Data Paradox in survey data of U.S. vaccine uptake by comparing the Delphi-Facebook survey of 250,000 individuals per week with an Axios-Ipsos online panel of 1,000 respondents. The former overestimated uptake by 17 percentage points with miniscule margins of error, while the latter provided reliable estimates and uncertainty quantification. The researchers suggested that the overrepresentation of white adults and people with college degrees in the Delphi-Facebook survey contributed to this error.[134]

Representativeness is less of a concern when a big dataset is used for investigation of associations and dependencies between variables.[62] Still, experts suggest that the importance of data quality continues to matter more than data quantity and that researchers should assess the importance of representativeness relative to the project context.[59,61]

Some degree of representativeness may be within the control of a researcher. For example, a researcher can use stratification and sampling techniques that increase the likelihood of achieving a representative sample. Approaches like enrichment sampling have also shown promise for improving representativeness.[139] Researchers can use the Data Representativeness Criterion (DRC) to assess the how representative a training dataset is of a new unseen dataset.[140] Other representativeness may be beyond the control of a researcher. For example, as of 2019, nonelderly American Indian and Alaskan Native, Hispanic, Native Hawaiian or Pacific Islander, and Black people are less likely to have health insurance than their white counterparts.[141] Because of this, a researcher using claims data from health insurance companies should acknowledge the limitation that the dataset is likely not representative for population level inference.

Even if a dataset is representative, it may still be unbalanced such that the representative sample is distributed unevenly over an outcome of interest.[142] For example, a cancer researcher may be working with data where only a fraction of respondents has cancer.[142] Many machine learning algorithms for classification problems on big datasets cannot handle class imbalance; thus, they generate inaccurate estimates for the less common outcome, even if that is the main outcome of interest. Researchers have three main options for addressing class imbalance: resampling, algorithmic modification, and cost-sensitive learning.[143] The latter two are employed less often due to challenges in application. Instead, more popular options include oversampling, undersampling, and a combination sampling approach.[144] Oversampling produces new synthetic samples and adds them to the less common category. Undersampling removes some repeated samples from the original dataset. Both approaches have criticisms: oversampling can lead to overfitting and an increase in computational cost while undersampling may remove potentially valuable data.[143,144] Researchers continue to explore new options for handling imbalanced data, including ensemble methods.[143] These techniques may also be applicable to addressing representation more broadly.

CASE STUDY 6: RAISING AWARENESS OF THE LACK OF REPRESENTATION IN DATASETS USED FOR TRAINING MEDICAL MACHINE LEARNING ALGORITHMS

In a 2020 study titled *Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms*, Kaushal et al. investigated the geographic distribution of patient cohorts used to train machine learning algorithms for medical purposes. The researchers were concerned that the lack of geographic diversity may introduce bias to the algorithms. A total of 2,606 studies were identified in a search, and 74 met inclusion criteria. Researchers found that the machine learning algorithms trained on U.S. patient data were disproportionately trained on data from California, Massachusetts, and New York, with little to no representation from the remaining 47 states. The authors say that these states may have economic, educational, social, behavioral, ethnic, or cultural features unrepresentative of the entire nation; therefore, algorithms trained primarily on data from these states may generalize poorly to new geographies. The researchers call for greater awareness of the lack of representation in these data and greater action to ensure that machine learning training datasets mirror the population that they are being designed for.[7]

ACTION ITEMS: BIG DATA REPRESENTATIVENESS

- Consider if the data are representative by race and ethnicity and assess the need for representativeness based on the research question or application.
- If needed, explore options to improve representativeness like stratification and sampling techniques.
- Evaluate if the dataset set is imbalanced and if so, explore oversampling or undersampling approaches.

FINAL NOTES

This text has discussed opportunities to assess bias throughout six key portions of the big data workflow: defining the study, sourcing big data, evaluating race and ethnicity attributes, identifying proxies, addressing data completeness, and assessing representativeness. This text is not exhaustive; no one approach, metric, or checklist can verify that a big data study is free of bias. Instead, researchers should continue furthering their education on how to address bias and take intentional steps such as those suggested in this report.

Suggested Further Reading

- Aragon C, Guha S, Kogan M, Muller M, Neff G. **Human-Centered Data Science**. The MIT Press; 2022.
- **Inventory of Resources for Standardized Demographic and Language Data Collection**. Centers for Medicare and Medicaid Services; 2022.
- Imbrahim S, Charlson, M, Neill D. **Big Data Analtyics and the Struggle for Equity in Healthcare: The Promise and Perils**. Health Equity 2020; 4(1):99-101. doi:10.1089/heq.2019.0112
- Tong M, 2021. **Use of Race in Clinical Diagnosis and Decision Making: Overview and Implications**. KFF. Published December 9, 2021.
- Vokinger KN, Feuerriegel S, Kesselheim AS. **Mitigating bias in machine learning for medicine**. Commun Med. 2021;1(1):1-3. doi:10.1038/s43856-021-00028-w

LEAD AUTHOR

Emily Hadley, MS – Research Data Scientist, RTI International

Emily Hadley is a Research Data Scientist with RTI International, an independent nonprofit research institute dedicated to improving the human condition. In her work, Emily collaborates with subject matter experts to solve complex problems in health, education, and criminal justice using data science and statistical methods. Emily's main research interests are exploring technical and policy approaches to addressing bias, equity, and ethics in data science. Emily has contributed multiple publications, presentations, and workshops to this area of research.

ACKNOWLEDGEMENTS

This report was produced with funding from The Robert Wood Johnson Foundation, which AcademyHealth gratefully acknowledges. Special thanks also go to the subject matter experts who contributed time and shared insights about their own experiences:

- **Jonathan Schwabish,** PhD, MA – Senior Fellow, The Urban Institute
- **Anita Chandra,** DrPH – Vice President & Director, RAND Corporation

Additionally, the following AcademyHealth staff members were involved in the development of this guide:

- **Rachel Dungan,** MSSP – Director, AcademyHealth
- **Marley Catlett,** MPH – Senior Research Associate, AcademyHealth
- **Michael Gluck,** PhD, MPP – Vice President, AcademyHealth

SUGGESTED CITATION

Hadley, E. "Companion Guide: Interactive Tool to Reduce Racial Bias in Big Data Studies," AcademyHealth, June 2023.

REFERENCES

1. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data*. 2019;6(1):54. doi:10.1186/s40537-019-0217-0
2. Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. *J Big Data*. 2022;9(1):3. doi:10.1186/s40537-021-00553-4
3. Roger Magoulas, Ben Lorica. *Big Data: Technologies and Techniques for Large-Scale Data*. O'Reilly Media, Inc; 2009.
4. Lee CH, Yoon HJ. Medical big data: promise and challenges. *Kidney Res Clin Pract*. 2017;36(1):3-11. doi:10.23876/j.krcp.2017.36.1.3
5. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. 2016;13(6):350-359. doi:10.1038/nrcardio.2016.42
6. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. *J Big Data*. 2018;5(1):1. doi:10.1186/s40537-017-0110-7
7. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA*. 2020;324(12):1212-1213. doi:10.1001/jama.2020.12067
8. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci*. 2020;117(23):12592-12594. doi:10.1073/pnas.1919012117
9. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol*. 2018;154(11):1247-1248. doi:10.1001/jamadermatol.2018.2348
10. Cerdeña JP, Plaisime MV, Tsai J. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *The Lancet*. 2020;396(10257):1125-1128. doi:10.1016/S0140-6736(20)32076-6
11. Flanagin A, Frey T, Christiansen SL, AMA Manual of Style Committee. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA*. 2021;326(7):621. doi:10.1001/jama.2021.13304
12. Ibrahim SA, Charlson ME, Neill DB. Big Data Analytics and the Struggle for Equity in Health Care: The Promise and Perils. *Health Equity*. 2020;4(1):99-101. doi:10.1089/heq.2019.0112
13. Racism and Mental Health: The African American experience: Ethnicity & Health: Vol 5, No 3-4. Accessed April 17, 2023. https://www.tandfonline.com/doi/abs/10.1080/713667453
14. Whitfield KE. *Closing the Gap: Improving the Health of Minority Elders in the New Millennium*. Gerontological Society of America; 2004.
15. David Gillborn, Paul Warmington, Sean Demack. QuantCrit: education, policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethn Educ*. 2018;21(2):158-179.
16. Snijders C, Matzat U, Reips UD. "Big Data": Big Gaps of Knowledge in the field of Internet science.
17. Kitchin R, McArdle G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data Soc*. 2016;3(1):2053951716631130. doi:10.1177/2053951716631130
18. Douglas Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety, Application Delivery Strategies*. META Group Inc; 2001.
19. Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. *BioMed Res Int*. 2015;2015:e639021. doi:10.1155/2015/639021
20. Pence HE. What is Big Data and Why is it Important? *J Educ Technol Syst*. 2014;43(2):159-171. doi:10.2190/ET.43.2.d
21. Sengupta E, Garg D, Choudhury T, Aggarwal A. Techniques to Elimenate Human Bias in Machine Learning. In: *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*. ; 2018:226-230. doi:10.1109/SYSMART.2018.8746946
22. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. *Patterns*. 2021;2(10):100347. doi:10.1016/j.patter.2021.100347
23. Bojke L, Soares M, Claxton K, et al. *Reviewing the Evidence: Heuristics and Biases*. NIHR Journals Library; 2021. Accessed January 8, 2023. https://www.ncbi.nlm.nih.gov/books/NBK571047/
24. Soule MC, Beale EE, Suarez L, et al. Understanding Motivations to Participate in an Observational Research Study: Why do Patients Enroll? *Soc Work Health Care*. 2016;55(3):231-246. doi:10.1080/00981389.2015.1114064
25. Coccia M. Motivations of Scientific Research in Society. Published online October 30, 2018. Accessed October 23, 2022. https://papers.ssrn.com/abstract=3275318
26. Fecher B, Hebing M. How do researchers approach societal impact? *PLOS ONE*. 2021;16(7):e0254006. doi:10.1371/journal.pone.0254006
27. Sheridan R, Martin-Kerry J, Hudson J, Parker A, Bower P, Knapp P. Why do patients take part in research? An overview of systematic reviews of psychosocial barriers and facilitators. *Trials*. 2020;21(1):259. doi:10.1186/s13063-020-4197-3
28. Knight W. Many Top AI Researchers Get Financial Backing From Big Tech. *Wired*. Accessed January 8, 2023. https://www.wired.com/story/top-ai-researchers-financial-backing-big-tech/
29. Fabbri A, Lai A, Grundy Q, Bero LA. The Influence of Industry Sponsorship on the Research Agenda: A Scoping Review. *Am J Public Health*. 2018;108(11):e9-e16. doi:10.2105/AJPH.2018.304677
30. Johnson H, Davies JM, Leniz J, Chukwusa E, Markham S, Sleeman KE. Opportunities for public involvement in big data research in palliative and end-of-life care. *Palliat Med*. 2021;35(9):1724-1726. doi:10.1177/02692163211002101
31. Andrus M, Spitzer E, Brown J, Xiang A. "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness. Published online January 22, 2021. doi:10.48550/arXiv.2011.02282

32. Room for improvement. *Nat Mach Intell*. 2021;3(1):1-1. doi:10.1038/s42256-021-00294-2
33. Jenna Wiens, Suchi Saria, Mark Sendak, et al. Do no harm: a roadmap for responsible machine learning for health care | Nature Medicine. *Nat Med*. 2019;25:1337-1340.
34. Ross-Hellauer T, Tennant JP, Banelytė V, et al. Ten simple rules for innovative dissemination of research. *PLoS Comput Biol*. 2020;16(4):e1007704. doi:10.1371/journal.pcbi.1007704
35. Brownson RC, Eyler AA, Harris JK, Moore JB, Tabak RG. Getting the Word Out: New Approaches for Disseminating Public Health Science. *J Public Health Manag Pract*. 2018;24(2):102-111. doi:10.1097/PHH.0000000000000673
36. Campo PD del, Gracia J, Blasco JA, Andradas E. A strategy for patient involvement in clinical practice guidelines: methodological approaches. *BMJ Qual Saf*. 2011;20(9):779-784. doi:10.1136/bmjqs.2010.049031
37. Wit M de, Abma T, Loon MK van, Collins S, Kirwan J. Involving patient research partners has a significant impact on outcomes research: a responsive evaluation of the international OMERACT conferences. *BMJ Open*. 2013;3(5):e002241. doi:10.1136/bmjopen-2012-002241
38. Greshake Tzovaras B, Angrist M, Arvai K, et al. Open Humans: A platform for participant-centered research and personal data exploration. *GigaScience*. 2019;8(6):giz076. doi:10.1093/gigascience/giz076
39. Agaronnik N, Campbell EG, Ressalam J, Iezzoni LI. Communicating with Patients with Disability: Perspectives of Practicing Physicians. *J Gen Intern Med*. 2019;34(7):1139-1145. doi:10.1007/s11606-019-04911-0
40. Doherty AJ, Atherton H, Boland P, et al. Barriers and facilitators to primary health care for people with intellectual disabilities and/or autism: an integrative review. *BJGP Open*. 2020;4(3). doi:10.3399/bjgpopen20X101030
41. Maine Groups Improve Care for Patients with Intellectual/Developmental Disabilities. Accessed October 23, 2022. https://www.ahrq.gov/news/newsroom/case-studies/202201.html
42. Dataset Search. Accessed October 23, 2022. https://datasetsearch.research.google.com/
43. Data.gov. Data.gov. Accessed October 23, 2022. https://www.data.gov/
44. Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, Sacchi L. Secondary Use and Analysis of Big Data Collected for Patient Care. *Yearb Med Inform*. 2017;26(1):28-37. doi:10.15265/IY-2017-008
45. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol*. 2021;21(1):234. doi:10.1186/s12874-021-01416-5
46. Johnston MP. Secondary Data Analysis: A Method of which the Time Has Come. *Qual Quant Methods Libr*. 2017;3(3):619-626.
47. Kaplan RM, Chambers DA, Glasgow RE. Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *Clin Transl Sci*. 2014;7(4):342-346. doi:10.1111/cts.12178
48. Favaretto M, De Clercq E, Elger BS. Big Data and discrimination: perils, promises and solutions. A systematic review. *J Big Data*. 2019;6(1):12. doi:10.1186/s40537-019-0177-4
49. Roemerman S. *Council Post: Four Reasons Data Provenance Is Vital For Analytics And AI*. Accessed August 28, 2022. https://www.forbes.com/sites/forbestechcouncil/2019/05/22/four-reasons-data-provenance-is-vital-for-analytics-and-ai/
50. Ferretti A, Ienca M, Hurst S, Vayena E. Big Data, Biomedical Research, and Ethics Review: New Challenges for IRBs. *Ethics Hum Res*. 2020;42(5):17-28. doi:10.1002/eahr.500065
51. Metcalf J, Crawford K. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data Soc*. 2016;3(1):2053951716650211. doi:10.1177/2053951716650211
52. Al-Zaiti SS, Alghwiri AA, Hu X, et al. A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML). *Eur Heart J - Digit Health*. 2022;3(2):125-140. doi:10.1093/ehjdh/ztac016
53. Two elite medical journals retract coronavirus papers over data integrity questions. Accessed August 28, 2022. https://www.science.org/content/article/two-elite-medical-journals-retract-coronavirus-papers-over-data-integrity-questions
54. Fadahunsi KP, O'Connor S, Akinlua JT, et al. Information Quality Frameworks for Digital Health Technologies: Systematic Review. *J Med Internet Res*. 2021;23(5):e23479. doi:10.2196/23479
55. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. *J Clin Epidemiol*. 2018;98:89-97. doi:10.1016/j.jclinepi.2018.02.023
56. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *Int J Epidemiol*. 2020;49(1):338-347. doi:10.1093/ije/dyz251
57. Brodie MA, Pliner EM, Ho A, et al. Big data vs accurate data in health research: Large-scale physical activity monitoring, smartphones, wearable devices and risk of unconscious bias. *Med Hypotheses*. 2018;119:32-36. doi:10.1016/j.mehy.2018.07.015
58. Fawzy A, Wu TD, Wang K, et al. Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19. *JAMA Intern Med*. 2022;182(7):730-738. doi:10.1001/jamainternmed.2022.1906
59. Meng XL. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann Appl Stat*. 2018;12(2). doi:10.1214/18-AOAS1161SF
60. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. *J Big Data*. 2021;8(1):140. doi:10.1186/s40537-021-00516-9
61. Huang JY. Representativeness Is Not Representative: Addressing Major Inferential Threats in the UK Biobank and Other Big Data Repositories. *Epidemiology*. 2021;32(2):189. doi:10.1097/EDE.0000000000001317

62. Cox DR, Kartsonaki C, Keogh RH. Big data: Some statistical issues. *Stat Probab Lett*. 2018;136:111-115. doi:10.1016/j.spl.2018.02.015

63. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for Datasets. Published online December 1, 2021. doi:10.48550/arXiv.1803.09010

64. Krause H. An Introduction to the Data Biography » We All Count. We All Count. Published January 21, 2019. Accessed August 28, 2022. https://weallcount.com/2019/01/21/an-introduction-to-the-data-biography/

65. The Data Nutrition Project. Accessed August 28, 2022. https://datanutrition.org/labels/

66. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Quantifying the Effect of Data Quality on the Validity of an eMeasure. *Appl Clin Inform*. 2017;8(4):1012-1021. doi:10.4338/ACI-2017-03-RA-0042

67. Michael Veale, Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data Soc*. 4(2). doi:10.1177/2053951717743530

68. *Evidence of Disparities among Ethnicity Groups*. Agency for Healthcare Research and Quality; 2018. Accessed September 11, 2022. https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata2.html

69. Williams DR, Lawrence JA, Davis BA. Racism and Health: Evidence and Needed Research. *Annu Rev Public Health*. 2019;40(1):105-125. doi:10.1146/annurev-publhealth-040218-043750

70. *Health Equity Style Guide for the COVID-19 Response: Principles and Preferred Terms for Non-Stigmatizing, Bias-Free Language.* Centers for Disease Control and Prevention; 2020.

71. How Structural Racism Works — Racist Policies as a Root Cause of U.S. Racial Health Inequities | NEJM. Accessed January 16, 2023. https://www.nejm.org/doi/full/10.1056/NEJMms2025396

72. Burnett-Bowie SAM, Bachmann GA. Racism: the shameful practices that the medical profession is finally addressing. *Womens Midlife Health*. 2021;7(1):9. doi:10.1186/s40695-021-00068-1

73. Bonham VL, Green ED, Pérez-Stable EJ. Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. *JAMA*. 2018;320(15):1533. doi:10.1001/jama.2018.13609

74. White K, Lawrence JA, Tchangalova N, Huang SJ, Cummings JL. Socially-assigned race and health: a scoping review with global implications for population health equity. *Int J Equity Health*. 2020;19(1):25. doi:10.1186/s12939-020-1137-5

75. Campbell ME, Troyer L. The Implications of Racial Misclassification by Observers. *Am Sociol Rev*. 2007;72(5):750-765. doi:10.1177/000312240707200505

76. Garcia JA, Sanchez GR, Sanchez-Youngman S, Vargas ED, Ybarra VD. RACE AS LIVED EXPERIENCE: The Impact of Multi-Dimensional Measures of Race/Ethnicity on the Self-Reported Health Status of Latinos. *Bois Rev Soc Sci Res Race*. 2015;12(2):349-373. doi:10.1017/S1742058X15000120

77. Campbell ME, Bratter JL, Roth WD. Measuring the Diverging Components of Race: An Introduction. *Am Behav Sci*. 2016;60(4):381-389. doi:10.1177/0002764215613381

78. Witzig R, Dery M. Subjectively-assigned race versus self-reported race and ethnicity in US healthcare. *Soc Med*. 2014;8(1):5.

79. Nead KT, Hinkston CL, Wehner MR. Cautions When Using Race and Ethnicity in Administrative Claims Data Sets. *JAMA Health Forum*. 2022;3(7):e221812. doi:10.1001/jamahealthforum.2022.1812

80. Mersha TB, Abebe T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. *Hum Genomics*. 2015;9(1):1. doi:10.1186/s40246-014-0023-x

81. *Improving Data Collection across the Health Care System*. Agency for Healthcare Research and Quality; 2018. Accessed September 11, 2022. https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata5.html

82. Polubriaginof FCG, Ryan P, Salmasian H, et al. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc JAMIA*. 2019;26(8-9):730-736. doi:10.1093/jamia/ocz113

83. Cara James, Smita Pamar, Sarah Hudson Scholle, Philip Saynisch, Jeni Soucie, Barbara Lyons. *Federal Action Is Needed to Improve Race and Ethnicity Data in Health Programs*. Grantmakers In Health; 2021.

84. *Inaccuracies in Medicare's Race and Ethnicity Data Hinder the Ability To Assess Health Disparities*. US Department of Health and Human Services; 2022. https://oig.hhs.gov/oei/reports/OEI-02-21-00100.pdf

85. Hernandez SE, Sylling PW, Mor MK, et al. Developing an Algorithm for Combining Race and Ethnicity Data Sources in the Veterans Health Administration. *Mil Med*. 2020;185(3-4):e495-e500. doi:10.1093/milmed/usz322

86. Rothwell CJ, Madans JH, Atkinson D, Ni H. *The Validity of Race and Hispanic-Origin Reporting on Death Certificates in the United States: An Update*. National Center for Health Statistics; 2016.

87. Kader F, Smith CL. Participatory Approaches to Addressing Missing COVID-19 Race and Ethnicity Data. *Int J Environ Res Public Health*. 2021;18(12):6559. doi:10.3390/ijerph18126559

88. Magaña López M, Bevans M, Wehrlen 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*. 2017;4(5):812-818. doi:10.1007/s40615-016-0283-3

89. Peltzman T, Rice K, Jones KT, Washington DL, Shiner B. Optimizing Data on Race and Ethnicity for Veterans Affairs Patients. *Mil Med*. 2022;187(7-8):e955-e962. doi:10.1093/milmed/usac066

90. *HHS Implementation Guidance on Data Collection Standards for Race, Ethnicity, Sex, Primary Language, and Disability Status*. US Department of Health and Human Services Accessed September 11, 2022. http://aspe.hhs.gov/reports/hhs-implementation-guidance-data-collection-standards-race-ethnicity-sex-primary-language-disability-0

91. *Template of Granular Ethnicity Category Lists and Coding Schemes with Rollup to the OMB Race and Hispanic Ethnicity Categories*. Agency for Healthcare Research and Quality; 2018. Accessed September 11, 2022. https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldataaptabe1.html

92. Chapter 3: Defining Categorization Needs for Race and Ethnicity Data | Agency for Healthcare Research and Quality. Accessed September 10, 2022. https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata3a.html

93. LIEBLER CA, HALPERN-MANNERS A. A Practical Approach to Using Multiple-Race Response Data: A Bridging Method for Public-Use Microdata. *Demography*. 2008;45(1):143-155.

94. Ross PT, Hart-Johnson T, Santen SA, Zaidi NLB. Considerations for using race and ethnicity as quantitative variables in medical education research. *Perspect Med Educ*. 2020;9(5):318-323. doi:10.1007/s40037-020-00602-3

95. Liu P, Ross JS, Ioannidis JP, Dhruva SS, Vasiliou V, Wallach JD. Prevalence and significance of race and ethnicity subgroup analyses in Cochrane intervention reviews. *Clin Trials*. 2020;17(2):231-234. doi:10.1177/1740774519887148

96. Gil-Sierra MD, Fénix-Caballero S, Abdel kader-Martin L, et al. Checklist for clinical applicability of subgroup analysis. *J Clin Pharm Ther*. 2020;45(3):530-538. doi:10.1111/jcpt.13102

97. HPOE.org - Reducing Health Care Disparities: Collection and Use of Race, Ethnicity and Language Data. Accessed September 11, 2022. http://www.hpoe.org/resources/ahahret-guides/1431

98. *A Framework for Stratifying Race, Ethnicity and Language Data*. Health Research & Educational Trust; 2014. www.hpoe.org/stratifyingdata

99. Gordon NP, Lin TY, Rau J, Lo JC. Aggregation of Asian-American subgroups masks meaningful differences in health and health risks among Asian ethnicities: an electronic health record based cohort study. *BMC Public Health*. 2019;19:1551. doi:10.1186/s12889-019-7683-3

100. Belenguer L. AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *Ai Ethics*. 2022;2(4):771-787. doi:10.1007/s43681-022-00138-8

101. *Examining the Black Box: Tools for Assessing Algorithmic Systems*. Ada Lovelace Institute; 2020. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

102. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health*. 2022;4(5):e384-e397. doi:10.1016/S2589-7500(22)00003-6

103. Ziad Obermeyer. *Algorithmic Bias Playbook*. Center for Applied AI at Chicago Booth; 2021. https://www.ftc.gov/system/files/documents/public_events/1582978/algorithmic-bias-playbook.pdf

104. Huang J, Galal G, Etemadi M, Vaidyanathan M. Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review. *JMIR Med Inform*. 2022;10(5):e36388. doi:10.2196/36388

105. Hasan A, Brown S, Davidovic J, Lange B, Regan M. Algorithmic Bias and Risk Assessments: Lessons from Practice. *Digit Soc*. 2022;1(2):14. doi:10.1007/s44206-022-00017-z

106. Vega Perez RD, Hayden L, Mesa J, et al. Improving Patient Race and Ethnicity Data Capture to Address Health Disparities: A Case Study From a Large Urban Health System. *Cureus*. 14(1):e20973. doi:10.7759/cureus.20973

107. Graham Upton, Ian Cook. *A Dictionary of Statistics*. 2nd ed. Oxford University Press; 2008.

108. Tschantz MC. What is Proxy Discrimination? In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM; 2022:1993-2003. doi:10.1145/3531146.3533242

109. Osoba O, Boudreaux B, Saunders J, Irwin J, Mueller P, Cherney S. *Algorithmic Equity: A Framework for Social Applications*. RAND Corporation; 2019. doi:10.7249/RR2708

110. Hooker S. Moving beyond "algorithmic bias is a data problem." *Patterns*. 2021;2(4):100241. doi:10.1016/j.patter.2021.100241

111. Yeom S, Datta A, Fredrikson M. Hunting for Discriminatory Proxies in Linear Regression Models. In: *32nd Conference on Neural Information Processing Systems*. ; 2018:11.

112. Datta A, Fredrikson M, Ko G, Mardziel P, Sen S. Proxy Non-Discrimination in Data-Driven Systems. Published online 2017. doi:10.48550/ARXIV.1707.08120

113. Briggs AH. Healing the past, reimagining the present, investing in the future: What should be the role of race as a proxy covariate in health economics informed health care policy? *Health Econ*. 2022;31(10):2115-2119. doi:10.1002/hec.4577

114. Tong M, 2021. Use of Race in Clinical Diagnosis and Decision Making: Overview and Implications. KFF. Published December 9, 2021. Accessed September 15, 2022. https://www.kff.org/racial-equity-and-health-policy/issue-brief/use-of-race-in-clinical-diagnosis-and-decision-making-overview-and-implications/

115. Krause H. Introduction to Proxy Variables » We All Count. We All Count. Published June 12, 2020. Accessed September 11, 2022. https://weallcount.com/2020/06/12/introduction-to-proxy-variables/

116. Vartan S. Racial Bias Found in a Major Health Care Risk Algorithm. Scientific American. Accessed September 11, 2022. https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

117. Spangler KR, Levy JI, Fabian MP, et al. Missing Race and Ethnicity Data among COVID-19 Cases in Massachusetts. *J Racial Ethn Health Disparities*. Published online September 2, 2022. doi:10.1007/s40615-022-01387-3

118. Labgold K, Hamid S, Shah S, et al. Measuring the missing: greater racial and ethnic disparities in COVID-19 burden after accounting for missing race/ethnicity data. *medRxiv*. Published online October 2, 2020:2020.09.30.20203315. doi:10.1101/2020.09.30.20203315

119. Ansari B, Hart-Malloy R, Rosenberg ES, Trigg M, Martin EG. Modeling the Potential Impact of Missing Race and Ethnicity Data in Infectious Disease Surveillance Systems on Disparity Measures: Scenario Analysis of Different Imputation Strategies. *JMIR Public Health Surveill*. 2022;8(11):e38037. doi:10.2196/38037

120. Randall M, Stern A, Su Y. Five Ethical Risks to Consider before Filling Missing Race and Ethnicity Data. Urban Institute. Accessed September 13, 2022. https://www.urban.org/research/publication/five-ethical-risks-consider-filling-missing-race-and-ethnicity-data

121. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59(10):1087-1091. doi:10.1016/j.jclinepi.2006.01.014

122. Imputing Race/Ethnicity: Part 1. RTI. Published September 2, 2021. Accessed October 27, 2022. https://www.rti.org/insights/imputing-raceethnicity-part-1

123. Andridge RR, Little RJA. A Review of Hot Deck Imputation for Survey Non-response. *Int Stat Rev Rev Int Stat*. 2010;78(1):40-64. doi:10.1111/j.1751-5823.2010.00103.x

124. Branham DK, Finegold K, Chen L, et al. Trends in Missing Race and Ethnicity Information After Imputation in HealthCare.gov Marketplace Enrollment Data, 2015-2021. *JAMA Netw Open*. 2022;5(6):e2216715. doi:10.1001/jamanetworkopen.2022.16715

125. Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study - Adjaye-Gbewonyo - 2014 - Health Services Research - Wiley Online Library. Accessed October 27, 2022. https://onlinelibrary.wiley.com/doi/10.1111/1475-6773.12089

126. Sholle ET, Pinheiro LC, Adekkanattu P, et al. Underserved populations with missing race ethnicity data differ significantly from those with structured race/ethnicity documentation. *J Am Med Inform Assoc JAMIA*. 2019;26(8-9):722-729. doi:10.1093/jamia/ocz040

127. Kim JS, Gao X, Rzhetsky A. RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning. *PLOS Comput Biol*. 2018;14(4):e1006106. doi:10.1371/journal.pcbi.1006106

128. Bigback KM, Hoopes M, Dankovchik J, 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(S1):1390-1402. doi:10.1111/1475-6773.12331

129. Espey DK, Jim MA, Richards TB, Begay C, Haverkamp D, Roberts D. Methods for Improving the Quality and Completeness of Mortality Data for American Indians and Alaska Natives. *Am J Public Health*. 2014;104(S3):S286-S294. doi:10.2105/AJPH.2013.301716

130. Silva GC, Trivedi AN, Gutman R. Developing and evaluating methods to impute race/ethnicity in an incomplete dataset. *Health Serv Outcomes Res Methodol*. 2019;19(2):175-195. doi:10.1007/s10742-019-00200-9

131. Bohensky MA, Jolley D, Sundararajan V, et al. Data Linkage: A powerful research tool with potential problems. *BMC Health Serv Res*. 2010;10(1):346. doi:10.1186/1472-6963-10-346

132. Office USGA. VA Health Care: Opportunities Exist for VA to Better Identify and Address Racial and Ethnic Disparities. Accessed January 8, 2023. https://www.gao.gov/products/gao-20-83

133. OECD Glossary of Statistical Terms - Representative sample Definition. Accessed January 15, 2023. https://stats.oecd.org/glossary/detail.asp?ID=3831

134. Bradley VC, Kuriwaki S, Isakov M, Sejdinovic D, Meng XL, Flaxman S. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*. 2021;600(7890):695-700. doi:10.1038/s41586-021-04198-4

135. Kevin C. Desouza, Kendra L. Smith. *Big Data for Social Innovation*. Stanford Social Innovation Review; 2014.

136. Chandrasekaran R, Katthula V, Moustakas E. Patterns of Use and Key Predictors for the Use of Wearable Health Care Devices by US Adults: Insights from a National Survey. *J Med Internet Res*. 2020;22(10):e22443. doi:10.2196/22443

137. Cahan EM, Hernandez-Boussard T, Thadaney-Israni S, Rubin DL. Putting the data before the algorithm in big data addressing personalized healthcare. *Npj Digit Med*. 2019;2(1):1-6. doi:10.1038/s41746-019-0157-2

138. Msaouel P. The Big Data Paradox in Clinical Practice. *Cancer Invest*. 2022;40(7):567-576. doi:10.1080/07357907.2022.2084621

139. Mercaldo ND, Brothers KB, Carrell DS, et al. Enrichment sampling for a multi-site patient survey using electronic health records and census data. *J Am Med Inform Assoc JAMIA*. 2018;26(3):219-227. doi:10.1093/jamia/ocy164

140. Schat E, van de Schoot R, Kouw WM, Veen D, Mendrik AM. The data representativeness criterion: Predicting the performance of supervised classification based on data set similarity. *PLoS ONE*. 2020;15(8):e0237009. doi:10.1371/journal.pone.0237009

141. Artiga S, Hill L, Orgera K, Damico A. Health Coverage by Race and Ethnicity, 2010-2019. KFF. Published July 16, 2021. Accessed September 11, 2022. https://www.kff.org/racial-equity-and-health-policy/issue-brief/health-coverage-by-race-and-ethnicity/

142. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J Biomed Inform*. 2019;90:103089. doi:10.1016/j.jbi.2018.12.003

143. Sánchez-Hernández F, Ballesteros-Herráez JC, Kraiem MS, Sánchez-Barba M, Moreno-García MN. Predictive Modeling of ICU Healthcare-Associated Infections from Imbalanced Data. Using Ensembles and a Clustering-Based Undersampling Approach. *Appl Sci*. 2019;9(24):5287. doi:10.3390/app9245287

144. A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique. *Int J Recent Trends Eng Res*. 2017;3(4):444-449. doi:10.23883/IJRTER.2017.3168.0UWXM