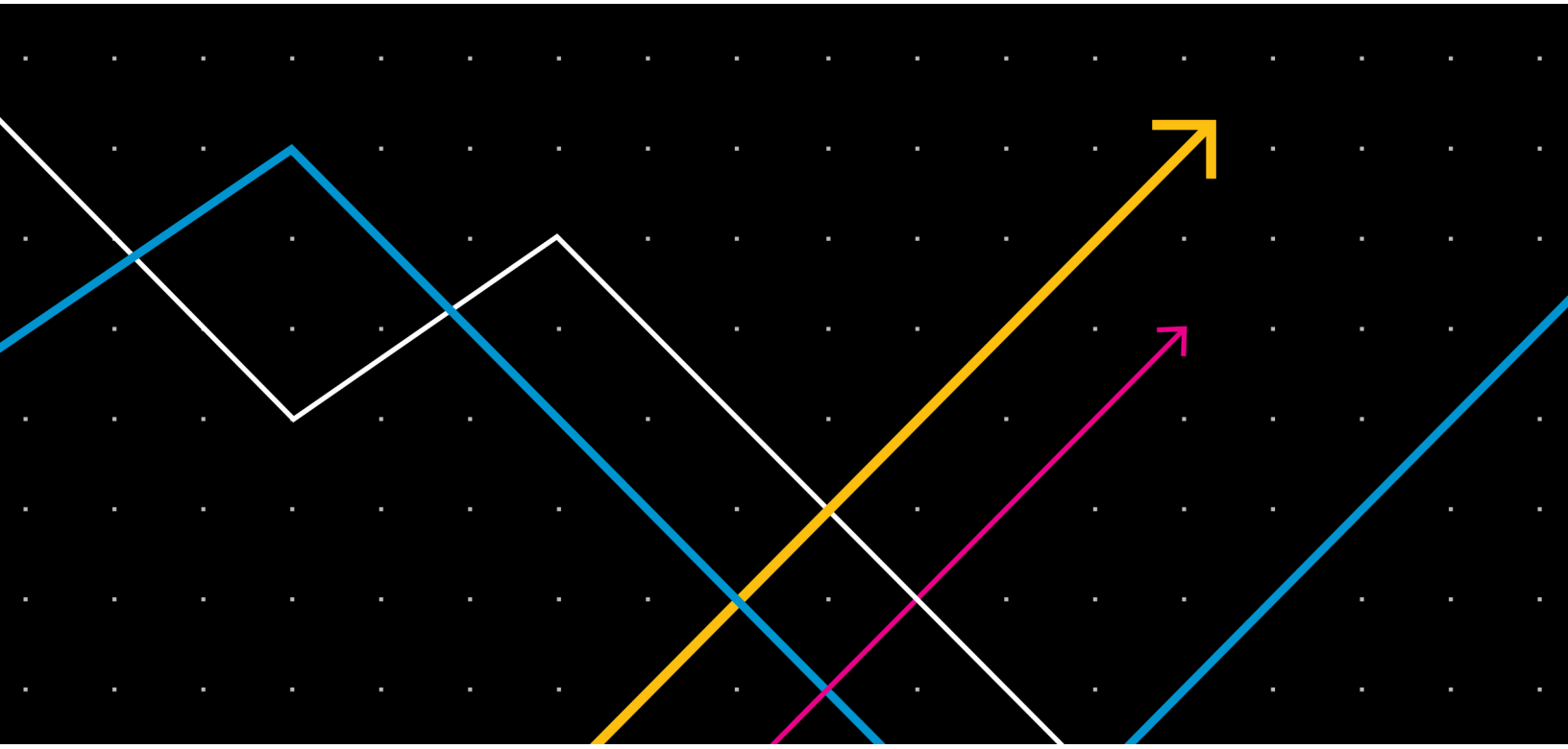

RESEARCH REPORT

Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity

Landscape Scan Findings

K. Steven Brown *Yipeng Su* *Jahnavi Jagganath* *Jacqueline Rayfield*
Megan Randall

August 2021

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

Contents

Acknowledgments

This report was funded by the Robert Wood Johnson Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

We would like to thank our key informants, who graciously lent their time, insights, and expertise as a part of this project, including Rexford Anson-Dwamena, Virginia Department of Health; Richard Chang, UCLA Center for Health Policy Research; Marc Elliott, the RAND Corporation; Amy Hawn Nelson, Actionable Intelligence for Social Policy; Shruti Jayaraman, Chicago Beyond; Chris Kingsley, Annie E. Casey Foundation; Chris Wheat, JPMorgan Chase Institute; and Janelle Wong, University of Maryland and AAPI Data.

We are also grateful to our Urban Institute colleagues who provided valuable internal review, comments, and feedback throughout the research process: Shena Ashley, Elsa Falkenberger, LesLeigh Ford, Graham MacDonald, Ajjit Narayanan, Kathy Pettit, and Alena Stern.

Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity: Landscape Scan Findings

A growing body of literature has demonstrated how seemingly "race-neutral" policies and systems can reinforce white privilege.[1] Alongside historical research about how racist policymaking has affected people of color in US communities, disaggregation of data by race and ethnicity is a critical tool for shining a light on racialized systems of privilege and oppression. There are strong ethical and practical reasons for disaggregating data: disaggregated data can provide clarity in areas where disparities have been suspected but not identified. Moreover, disaggregating data enables people to see themselves reflected in data, which in turns enables them to make decisions, inform policy, and advocate for a more just and equitable distribution of resources.

Despite the benefits of disaggregated data, many high-value datasets lack information on race and ethnicity. This could owe to various factors, such as long-standing practices of not collecting this information, regulatory constraints, concerns about nonresponse, and even the notion that race and ethnicity data are not central to questions asked on forms and surveys. Whatever the reasons, the absence of disaggregated data has harmed communities of color and obscured disparities in health, wealth and financial well-being, justice involvement, safety net benefits, and other policy domains. Moreover, the COVID-19 pandemic has brought long-overdue attention to the need for real-time disaggregated data.[2] For example, according to the Centers for Disease Control and Prevention, race and ethnicity data were missing from nearly 50 percent of vaccine records in the first month of vaccine rollout, which obscured racial and ethnic disparities in vaccinations.[3] And regarding disparities in wealth and financial well-being, the lack of race and ethnicity information in credit bureau data has inhibited efforts to examine how credit scores affect racial homeownership gaps and to challenge the use of credit screens in hiring.[4]

In response, data scientists and researchers have developed, and continue to expand, creative methods for appending race and ethnicity onto datasets lacking those data, allowing policymakers to disaggregate those data and track racial disparities to inform policymaking. These methods are often held up as alternatives to the "gold standard" of collecting original, self-reported data on people's race

and ethnicity when those original data are not feasible to collect (such as with historical data) or are not allowed to be collected (such as with tax data). But these methods do not typically require the input of the people whose data are being combined or augmented, creating ethical risks and a potential lack of empathy for people whose data are used. As this body of methods expands,[5] we believe it should be accompanied by a robust discussion of the ethical risks associated with applying different methodological strategies.

In this report, we focus on risks associated with a method known as imputation, primarily because it is increasingly employed to append race and ethnicity onto new datasets.[6] For example, to address the lack of racially disaggregated COVID-19 data in Virginia, the Virginia Department of Health has used the Bayesian Improved Surname Geocoding method developed by the RAND Corporation to impute missing race and ethnicity data from vaccine records (Anson-Dwamena, Pattah, and Crow 2020). Although imputation is being used more often, the literature on associated ethical concerns remains sparse.

In the sections that follow, we review common methods for appending race and ethnicity data, including imputation. We then discuss our research approach for this landscape scan, and lastly document and discuss the areas of ethical risk reported by our key informants and the literature.

This effort is part of a project being led by the Racial Equity Analytics Lab**,** which seeks to equip today's change agents with data and analyses to advance social and economic policies that help remedy persistent structural racism. We believe more thoughtful, race-conscious policies and practices have the power to forge new avenues of opportunity and prosperity for Black, Latine, Native American, and Asian American and Pacific Islander communities and other communities of color, and that timely, reliable data are essential for designing race-conscious solutions—as well as for holding decisionmakers and institutions accountable for choices that perpetuate oppressive systems.[7]

Addressing Race and Ethnicity Data Gaps

Many valuable public, administrative, and private data sources that could help policymakers understand and predict disparate outcomes by race and ethnicity lack information on race and ethnicity. For example, to protect consumers from racial discrimination, credit bureaus are legally prevented from collecting this information for credit score data,[8] which limits efforts to examine how racial disparities in credit scores affect hiring, homeownership, taxation, and other social and economic domains —and vice versa.

Even fewer datasets disaggregate data at the hyperlocal geographies often necessary for state and local decisionmaking, and many that do are limited in terms of their completeness, accuracy, or granularity.[9] Even credible public data sources are limited in their ability to disaggregate at important levels of detail. The US Office of Management and Budget, for example, only requires the Census Bureau to collect and classify responses for a minimum of five racial categories.[10] Established in 1997, this classification can mask experiences and outcomes within broader racial groups, such as Asian American and Pacific Islanders, a group that comprises many diverse subgroups with different experiences and needs.

Absent this information, researchers are forced to choose between using imprecise methods to estimate race (such as using the predominant race of a person's zip code)[11] and forgoing disaggregation altogether. Statisticians, researchers, and data scientists have developed and are furthering various tools for filling these critical gaps, including imputation, machine learning, and data-linkage methods:

- **Imputation.** Probabilistic methods like imputation can generate data that maintain the statistical properties of "real" data. One imputation method is Bayesian Improved Surname Geocoding, which is used to impute race and ethnicity on administrative data (this method was developed by the RAND Corporation for the US Department of Health and Human Services and is used by the Equal Opportunity Employment Commission) (Fremont et al. 2016; Harris 2020). Another method is multiple imputation, which involves creating multiple copies, or implicates, of an imputed race and ethnicity variable; it is a standard procedure used in many public data products, such as the SIPP Synthetic Beta, the National Survey of Children's Health, and the Survey of Consumer Finances. Using multiple imputation, researchers can analyze variation resulting from the uncertainty in input data sources and, drawing from sets of probabilities, can determine whether results are robust to the randomness inherent in the imputation process.
- **Machine-learning methods.** These can be especially useful when using text data or modeling complex, nonlinear relationships. For example, the Urban Institute applied machine-learning methods to predict property-level zoning-density limits (Nechamkin and MacDonald 2019) and impute sentiment toward police from tweets (Oglesby-Neal, Tiry, and Kim 2019). Several groups are already working on ways to mitigate biases in machine learning, such as the Algorithmic Justice League and researchers from the University of Chicago who developed the Aequitas tool to audit algorithms for bias and fairness (Saleiro et al. 2019).[12]

- **Data linkage.** Probabilistic data linkage, popularly known as fuzzy matching, involves using multiple and/or nonunique keys to connect information from separate sources based on the probability of two records representing the same person or entity. For example, Urban probabilistically linked names and addresses of community development financial institutions to estimate community development financial flows and linked mothers' and infants' records to evaluate the Strong Start for Mothers and Newborns initiative (Hill et al. 2015; Theodos and Hangen 2017). Another type of data linkage is data fusion, which involves integrating multiple data sources to achieve more accuracy than a single data source would (e.g., combining multiple administrative data sources, such as the American Community Survey, city surveys, and United States Postal Service data, to understand change in a neighborhood).

Although this report focuses on imputation, our learnings can be generally applied to the other modifying approaches, including machine learning and data linkage and integration.[13] We focus on imputation in part because of its increasing application, but also because the ethical concerns around it have received less coverage in the literature than those other approaches.

Research Approach

This landscape scan is one of three complementary products in the Racial Equity Analytics Lab's Ethics and Empathy series, the two others being a case study of imputing race and ethnicity onto credit bureau data and a standards guide for researchers, data analysts, and other stakeholders on how to apply imputation methods ethically for racial equity analysis (Brown, Ford, and Ashley 2021; Stern and Narayanan 2021). Our objective for this scan was to surface potential ethical risks associated with using imputation to append race and ethnicity data for racial equity analysis to inform decisions and recommendations as reported in the case study and standards guide. Toward that objective, we spoke to and read papers written by technical experts on imputation and related methods and stakeholders from organizations focused on racial equity and community engagement. As a first step, we performed a scan to identify major voices in the following intersecting spaces:[14]

- research and data science
- data equity advocacy
- data- and research-informed policymaking
- analytics and equity in the private and philanthropic sectors

As part of our **literature review**, we identified published works from organizations and individuals that focus on or highlight data disaggregation, imputation, and racial equity, and for our **key informant interviews**, we sought interviews with a handful of key authorities on these subjects. Throughout our landscape scan, we applied lessons and findings to our case study and ultimately used findings from the scan to inform recommendations in the standards guide. We also incorporated questions and challenges that arose during the case study into our landscape scan to identify and apply existing best practices and to apply relevant measures for mitigating ethical risk described in the literature.

Literature Review

We initially sought to identify published writings from authoritative sources that addressed at least one of four ethical-risk areas surfaced from a previous workshop on the ethics of imputing race and ethnicity for racial equity analysis (box 1) (Randall, Stern, and Su 2021). These risk areas were the following:

- excluding people and communities of color from ownership of their data and from decisions on research process and methods
- violating informed consent, privacy, or confidentiality
- producing inaccurate estimates with misleading conclusions
- generating data for purposes that harm people of color

We identified 10 to 15 sources for each of the four risk areas from various fields and perspectives. Our sources included methodological and technical guidance on data imputation, critical examinations of big data and race and ethnicity, and data governance practices that can mitigate ethical risk (Benjamin 2016; Hawn Nelson, Jenkins, Zanti, Katz, and Berkowitz 2020; Jefferson 2018; Lee et al. 2016; Leslie 2019; Petty et al. 2018).[15] Early selections from the literature and the four risk areas shaped our interview instrument (described in detail in the next section). We added works throughout our landscape scan as we consulted more literature and spoke with interviewees who pointed us to additional sources. Although the four risk areas provided a foundation for our inquiry in the landscape scan, our synthesis went beyond them to speak to broader technical and theoretical themes.

BOX 1

Ethics and Imputation: Examining Four Potential Risk Areas

The four risk areas below are derived from an early review of the literature and from insights from a November 2020 workshop—the Design Thinking Workshop on Ethical Imputation—hosted by the Urban Institute and attended by 24 participants from the fields of advocacy, research, data science, and public policy.[a] These risk areas provided a preliminary foundation upon which we built our more extensive literature and key informant inquiries for this landscape scan.

Risk area 1: excluding people and communities of color from ownership of their data and from decisions on research processes and methods. Power dynamics between people whose data are being collected and the organizations funding, collecting, and using those data can prevent people from exercising authority over their own data. Researchers and data analysts who impute race and ethnicity onto datasets with previously collected secondary data may be disconnected from the people the original data were collected from and even from the purposes they were collected for. Failing to provide channels for critical individual and community-level input increases the likelihood that researchers will overlook how people and communities connect to and identify with the research process.

Risk area 2: violations of informed consent, privacy, and confidentiality. When racial or ethnic identifiers are appended to other identifiers at small units of geography or in otherwise small populations, people are at increased risk of being reidentified, even if the datasets have been anonymized. Moreover, someone consenting to provide sensitive individual financial or health data absent racial identifiers may revoke that consent if plans to append race or ethnicity to their individual data are fully disclosed. Communities of color have historically been systematically deprived of opportunities for informed consent in research. Even today, Black patients are overrepresented in Food and Drug Administration–approved clinical trials that do not require informed consent. Establishing avenues and standards for providing people voice and choice on how and when their data are used is a critical component of equity-centered research.

Risk area 3: producing inaccurate estimates with misleading conclusions. Imputation and related methods often come with a degree of statistical uncertainty and only produce results as accurate as the underlying data, which often reflect structural disparities and racial biases. For example, in 2020, the California attorney general revoked law enforcement departments' access to a database of suspected gang members because of pervasive errors. Gang affiliations had been assigned using largely unsubstantiated (and, in some cases, demonstrably falsified) reports from individual law enforcement officers, reflecting significant racial bias (Benjamin 2019). Without careful examination of the robustness of the estimates and clear definitions of acceptable ranges of uncertainty for different use cases, linking biased datasets together, using them to power data-driven decision systems, and training predictive algorithms can magnify erroneous predictions and lead to misinformed policy choices that harm communities of color.

Risk area 4: generating data for purposes that harm people or communities of color. Data generated from an imputation can be used to target people and communities of color, which can lead to direct harms such as predatory lending or disproportionate policing (Derenoncourt 2019; Hwang, Hankison, and Brown 2015). This concern gets at the heart of a larger debate about the responsible use and presentation of racially and ethnically disaggregated data. Any tool that allows for disaggregation, including imputation and related methods, can contribute to racist narratives if the data reinforce harmful stereotypes about people of color that lead to discrimination against groups and neighborhoods. This can happen through the selection and visual presentation of data and through framing racially disparate outcomes as resulting from individual choices and behaviors rather than structural forces.

Notes:

[a] See Randall, Stern, and Su (2021) for a detailed discussion of the workshop as well as the risk areas in question. Note that although that brief discusses five ethical risk areas, for the purposes of this landscape scan we collapsed our inquiry on informed consent and privacy into one risk area.

Key Informant Interviews

We identified voices in the fields of data science and data equity we believed would bring a broad range of perspectives, experiences, and knowledge to questions on imputation and racial equity analysis. In our interview questionnaire, we included questions about risks in the areas described in box 1 and expanded our inquiry beyond those areas, inviting interviewees to share risks we may have overlooked in the November 2020 workshop that preceded and informed this scan.

In the four groups of stakeholders we identified in our initial actor scan, we sought interviews with data equity advocates, researchers engaged in racial equity analysis or imputation work, policymakers or public-sector data workers, and private-sector data actors, including those in philanthropy. We obtained our list of names from our review of the literature and an inquiry among Urban Institute researchers who work in the data science, racial equity, data equity, and community engagement fields. Between January and April of 2021, we successfully completed interviews with the following eight key informants:

- Rexford Anson-Dwamena, Virginia Department of Health
- Richard Chang, UCLA Center for Health Policy Research
- Marc Elliott, the RAND Corporation
- Shruti Jayaraman, Chicago Beyond
- Chris Kingsley, Annie E. Casey Foundation

- Amy Hawn Nelson, Actionable Intelligence for Social Policy, University of Pennsylvania
- Chris Wheat, JPMorgan Chase Institute
- Janelle Wong, University of Maryland and AAPI Data

We obtained an interview with at least one person in four of our five priority areas, with the exception of actual policymakers, although we did speak with one expert (Anson-Dwamena) in a state policymaking institution. Notably, we conducted interviews during a time when major institutions in the federal government, private sector, and philanthropy were vocally taking steps to advance racial equity in their organizations. Several experts we reached out to, particularly in the data equity and community advocacy spaces, either declined or did not respond to our invitations. Our understanding is that demands on advocates' time for advising and conducting interviews have been intense over the past year, and that many may not have had capacity or interest in participating during this crucial period. Given the importance of their perspectives, we incorporate publicly available writings and speeches into the findings from our scan. While we draw upon the individual and collective insights of the informants, the findings reflect the authors' research and analysis of the full scan of literature and interviews.[16]

From the interviews and scan, we analyzed how the informants weighed the risks of imputing race and ethnicity against the benefits, considered the role of empathy in the research process, and suggested methodological improvements or alternative approaches to collecting data more ethically and empathetically to understand racial and ethnic disparities.

By ethical risks in this context, we mean the ways applying a method could **harm people, put them at greater risk of harm, or benefit them less than other methods**. Of course, when a method of generating disaggregated data can be applied to understand the disparate impact of critical policy decisions, *not* applying it could also do harm and therefore carries its own ethical risk.[17]

Moreover, by **empathy** in this context, we mean **adequately considering the personhood and the expressed concerns and needs of people and communities** reflected in input data and in outputs generated from imputation or other methods. Analysts may assume they know communities' needs and concerns, in some cases because they may feel that their own preconceptions are grounded in existing bodies of research. But *seeing* and *being seen* are critical components of empathy and require that analysts ask and listen to people articulating what they want for themselves. An empathetic approach recognizes that data come from and reflect people and communities at potentially vulnerable moments in their lives. We asked our interviewees what role empathy should play in the

development and deployment of analytic methods like imputation—and what an empathetic approach would look like when using these methods for racial equity analysis.

Key Findings

Although imputation is a powerful tool for amending and appending data, users of this method must consider the associated risks for people and communities whose lives are reflected in the data. The race and ethnicity data that imputation generates can help us understand and address racial disparities, but employing it without considering its limitations and pitfalls can increase the risk of harming people of color and worsening disparities. We examined the literature for, and queried experts in the field on, the potential risks and benefits of employing imputation to generate disaggregated data for racial equity analysis, including the risks of violating **ethics** and **empathy** in the research process.

Our key findings reflect a synthesis of the recurring themes that were present to some extent in both the interviews and the literature review. We share examples of where each main finding appears in the literature we reviewed or in the interviews we conducted, sharing quotes and citations throughout.

Imputation Is a Useful but Imperfect Tool

Many in the field, including researchers and advocates, are excited about analytic advances that can shed light on racial disparities and encourage policymakers to craft more race-conscious policies, and many are excited to invest in analytic tools like imputation that can bolster and expand racial equity analyses. Most people we spoke with said the need for disaggregated data is likely to outweigh the risks of obtaining those data, whether through imputation or other analytic means. For example, Wong shared, "[For Asian American and Pacific Islander communities] I think the benefits of data disaggregation are so overwhelming. And there's been such a call [for disaggregated data] among…those groups that are most disenfranchised within our community that I really don't see an ethical concern…I don't think the potential harms outweigh the amazing benefits: the substantive and material benefits that will flow to the least advantaged in our communities when it comes to data disaggregation."

I don't think the potential harms outweigh the amazing benefits: the substantive and material benefits that will flow to the least advantaged in our communities when it comes to data disaggregation. —Janelle Wong, AAPI Data

When asked about the risks and benefits of imputation, several interviewees said they considered imputed race and ethnicity data an imperfect substitute for the gold standard of self-reported race and ethnicity data. Preferably, trusted data collectors and data owners would collect race and ethnicity data directly through a community census or survey and make them widely accessible to researchers and the public, but this approach can be time-consuming and expensive. And it is impossible to collect information on race and ethnicity for historical data that lack it. In the absence of self-reported data, many in the field make a strong case for imputing race and ethnicity data, despite technical and ethical challenges.[18]

Importantly, we found in our literature review and interviews that some equity-focused stakeholders caution against the growth of big data and warn of harms it is doing to communities of color. A broader body of literature and advocacy at the intersection of racial equity and data outlines the risks that data-based algorithms pose to communities of color and the history of discrimination by collectors and users of big data.[19] For example, the Detroit Community Technology Project has protested the use of facial recognition software and predictive policing and worked to organize communities at risk of being undercounted in the 2020 Census, complementary efforts to protect against high-tech and data-driven abuses of power and achieve more accurate visibility in datasets like the census that are critical for allocating political and fiscal resources equitably.[20] Yeshi Milner of Data for Black Lives has explained how the development of big data traces a history of racialized social and economic control and segregation that originated from slavery and Jim Crow.[21] Milner and others in the big-data abolition movement criticize the large-scale accumulation and use of data for corporate interests without people's consent, data that often harm historically marginalized groups. "A bank, a college application, a patient algorithm used by doctors to determine who gets care, a risk assessment that determines your prison sentence, or surveillance system," she writes, "do not need to know your race, as long as they have your zip-code."[22]

Although these perspectives apply to data writ large, imputation in particular inevitably intersects with big-data systems, as big data are often used as the foundation for, or to train inputs into, predictive statistical models. The concern is that analysts using advanced analytics and statistical

methods such as imputation may not be mindful of how big data can reproduce biases and reflect disparities. And without sufficiently testing for bias to contextualize the structural factors that may cause or facilitate the disparities, the data produced may have limited ability to support disparity-reducing policies and actions and could be especially harmful to targeted communities. Analysts using imputation must grapple with imputation's potential similarities to and intersections with "facial recognition, biometric data collection, credit scoring, risk assessments," and other big-data systems that Milner and others have identified as upholding systems of oppression.[23]

Ethics Best Practices Are Underdeveloped amid a Focus on Technical Application

The literature on imputation focuses largely on technical questions and techniques and often fails to position those techniques within a broader ethics framework.[24] We found few sources offering approaches for mitigating ethical risks when using imputation to fill race and ethnicity data gaps. Several sources focus primarily on the benefits of imputation as an analytic approach for generating race and ethnicity data and thus for empowering important analysis on discrimination or racial disparities.[25]

Broadening our search to include fields like "open data," "disaggregated data," and "data privacy" uncovers examples of ethical frameworks that could be applied to or adapted for imputation.[26] But many of the specific recommendations in these frameworks cover topics in data governance, collection, and sharing that don't apply to imputing data with an existing dataset. Zhang and coauthors, for example, make recommendations for how big-data practices can be adapted and improved to reduce health disparities, including by "investing in data collection on small sample populations, building a diverse workforce pipeline for data science, actively seeking to reduce digital divides, developing novel ways to assure digital data privacy for small populations, and promoting widespread data sharing to benefit under-resourced minority-serving institutions and minority researchers" (2017, 95). They briefly discuss the limitations of imputation, but do not tailor their recommendations to improving imputation methodologies. Suggestions for expanding original data collection or deriving novel ways to protect data privacy for populations whose data are being collected do not necessarily apply to imputation as a data-expansion tool. Though necessary for a broader conversation on ethics in data, existing frameworks and suggestions do not specifically address the history of imputation and its risks and benefits.

There is little guidance for researchers and data scientists about how to ethically use imputation to create disaggregated data. When critiques and cautions from the literature *do* apply to imputation,

actionable steps and best practices are often missing. For example, some literature suggests that any method that involves building out and applying new data from existing sources (as does imputation) can limit resulting policy recommendations to existing (often racialized and oppressive) frames of understanding, but provides no avenue for avoiding this potential pitfall.[27] One study, for example, explains how an algorithm that estimated health risk for patients used health care cost as a proxy for risk. Because of this choice, Black patients were judged as being at lower risk than white patients because they had been systemically provided cheap, lower-quality care than white patients (Benjamin 2019). This is an example of how using a proxy for race when analyzing or imputing data can build on existing inequalities. There is an appetite for additional tools and guidance that allow researchers to audit their predictive models for bias.

When we asked our interviewees about the specific risk areas we identified in previous research (see box 1), few offered imputation-specific solutions to help mitigate those risks. Our interviewees had varying familiarity with the technical aspects of the methodology. Most were eager to discuss larger principles around ethics and empathy in data work instead, suggesting that when talking to stakeholders without technical expertise, more granular best practices for ethically applying imputation techniques may need to follow a pointed and focused discussion on what ethics and empathy mean to practitioners. Interviewees wanted further discussion on how the need for disaggregated data can be balanced with risks and methodological drawbacks.

Getting imputed values wrong and then using them to inform decisions that affect people's lives and well-being can cause significant harm. As Kingsley shared, "It's one thing to [use imputed data] for statistical or analysis purposes. It's another to [use] it for service delivery purposes or to inform an algorithm that might actually make a decision about somebody—about whether child welfare services visits your family or not." Ethical application starts at the beginning, when projects are being conceived, but it carries through the process and applies to how data are eventually used.

It's one thing to [use imputed data] for statistical or analysis purposes. It's another to [use] it for service delivery purposes or to inform an algorithm that might actually make a decision about somebody—about whether child welfare services visits your family or not. —Chris Kingsley, Annie E. Casey Foundation

---

As Hawn Nelson put it, "You can't do [imputation] without the legal and governance work first. It has to be administrative first. It's not ethical to start with just a pure dataset and figure it out. You have to have the legal framework and oversight." Imputation and other analytic methods do not typically require the input of the people whose data are being combined or augmented, creating ethical risks and a potential lack of empathy for people whose data are used in the process. Although it is possible to begin an imputation without legal guidance in place, particularly when conducting the imputation on secondary data sources with no personally identifying information, it is still necessary to set up accountability and governance structures before the project begins. In fact, given imputations appending race and ethnicity can be done without legal considerations, the need for ethics guardrails and outside oversight is all the more important.

To help ensure accountability, researchers could, in line with insights gleaned from the literature and expert interviews, consult many community stakeholders throughout the research process when imputing race and ethnicity; this is important given imputing race and ethnicity involves assigning probabilistic race and ethnicity identifiers to people from multiple communities, including racial and ethnic subgroups across different geographies.[28] This involves challenges though, in that identifying the correct community of stakeholders to represent large groups of people is not straightforward. As we noted in our discussion of ethics gaps in the imputation literature, little practical guidance is available to researchers on how to incorporate principles for robust community engagement into wide-reaching analytics projects where data cover many communities in many geographies.

You can't do [imputation] without the legal and governance work first. It has to be administrative first. It's not ethical to start with just a pure dataset and figure it out. You have to have the legal framework and oversight.
—Amy Hawn Nelson, Actionable Intelligence for Social Policy

A strong path for the field is to develop data governance practices and institutional standards for defining and employing ethical frameworks and best practices in community engagement, individual privacy protection, and other areas of ethical concern. As a foundation for ethical imputation, data producers and users can collaborate to develop data ownership and governance infrastructure centering communities of color; industry standards and best practices, including data-quality benchmarks; and community-engagement principles.

Imprecision Produces Disparate Benefit and Risk across Subgroups

Many interviewees questioned whether imputation can accurately represent racial distributions for less populous or more widely dispersed subgroups, such as American Indians, Alaskan Natives, Native Hawaiians, Pacific Islanders, and Asian subgroups (e.g., Indian, Cambodian, Vietnamese, Japanese).[29] High-quality, accurate benchmarks are needed to effectively "train" an imputation model to accurately predict race and ethnicity and to conduct accuracy checks on those predicted values. Several interviewees expressed concern over whether existing sources of data can effectively perform these critical functions. Many sources of statistical bias in existing data disproportionately impact less populous geographies and racial and ethnic groups because of their smaller sample sizes and greater estimate uncertainty. Chang explained how the census's lack of granularity masks diversity within the AAPI category: "Unfortunately…the AAPI label still is still being used regularly. So, we unfortunately continue to face the issue not only of *not* having our data but, if imputation is taking place, assuming that NHPIs [Native Hawaiians and Pacific Islanders] share similar characteristics to Asian Americans." The tension expressed here suggests that less populous communities could stand to benefit most from data disaggregated through imputation methods, but smaller sample sizes and uncertainty in the data mean the risk of misrepresenting these communities is often higher.

Elements of imputation that appear purely technical thus do pose an ethical and equity-related challenge in that they confer disparate benefits to racial and ethnic subgroups while exposing them to all of the risks of imputation. In fact, if proper steps are not taken to obscure people's identities, people in less populous racial and ethnic subgroups may be *more* vulnerable to risks like reidentification and other privacy violations while standing to benefit less from the data generated by that process.[30] This is especially true if the imputed race and ethnicity variable turns out to be accurate, although because it functions as an assigned probability, imputation is more subject to error or misidentification at the individual level.

Inaccurate race and ethnicity predictions, if used to tailor and craft policy, may result in choices that do not serve communities' needs, introducing a pragmatic policymaking risk that is not shared equally across all racial and ethnic subgroups. Wheat shared, "That is the platform upon which I think all ethical questions [sit]: Are people *worse off* because of a choice you made about the way you impute?" Analysts undertaking imputation encounter technical checkpoints with ethical implications, points at which they can assess whether their technique is mitigating or exacerbating statistical uncertainty for different subgroups (Stern and Narayanan 2021).

That is the platform upon which I think all ethical questions [sit]: Are people worse off because of a choice you made about the way you impute?
—Chris Wheat, JPMorgan Chase Institute

Interviewees commonly shared that analysts should approach imputation with methodological proficiency *and* awareness of community impact in mind. Our scan raised important questions about whether imputation can provide the information that less populous communities need to understand, and advocate for policymakers to redress, the disparities they face.

BOX 2

Using Imputation or Alternative Methods

The bulk of the evidence from our landscape scan suggests that imputation is a valuable tool that can produce more disaggregated data with which to understand racial disparities quicker and more affordably than fielding new data collection. But in some cases, such as when statistical bias in training data produces inaccurate race and ethnicity predictions, interviewees said imputation may cause more harm than good.

To avoid harm, interviewees suggested that analysts need to first identify a use case (the "why" behind an imputation project) that is connected to an expressed community need. Only after identifying a central policy or research question can an analyst make an informed decision about whether imputation is the best tool for meeting that need. For example, when asked about weighing the risks and benefits of imputing data for Native Hawaiian and Pacific Islander groups as a more granular disaggregation of Asian American and Pacific Islander data, Richard Chang of the UCLA Center for Health Policy was skeptical about imputation's accuracy. He shared, "There would have to be such absurdly large, detailed sources of [AAPI subgroup] data to be able to accurately impute subgroups of our communities…that I'm still skeptical [of imputation's] accuracy and wonder whether it…would be a better use of time to just oversample [collect disproportionately more data for target subgroups]."

The decision of *whether or not* to impute requires analysts to proactively examine the relative risks and rewards that imputation will confer to different racial and ethnic groups and subgroups within those communities. It also requires analysts to consider what alternative approaches might be available to help meet the expressed community need, along with their respective risks and benefits.[a] Other approaches that may be available include collecting self-reported race and ethnicity data from the community in question, conducting interviews to complement quantitative sources, and fielding

supplemental community surveys. For further considerations around about deciding whether and how to impute, see Stern and Narayanan (2021) and Brown, Ford, and Ashley (2021).

Note:

[a] Stern and Narayanan (2021) discuss different use cases and checkpoints that analysts can refer to when deciding to proceed with an imputation project.

Empathy Is a Critical but Often Missing Ingredient

Empathy is a crucial but often missing component of data work. Empathy requires recognizing that data represent real people at potentially vulnerable moments and an approach rooted in historical knowledge about the people represented in the data, how data have affected them, and how that history has affected their relationship to big data and their desire for representation in data.[31] As Hawn Nelson said during our interview, "Every single number is a person, and every number is often someone's worst day. For a lot of human services data—evictions, homelessness, abuse, etcetera—the data [that analysts are] looking at *every* day was someone's *worst* day."

Every single number is a person, and every number is often someone's worst day. For a lot of human services data—evictions, homelessness, abuse, etcetera—the data they're looking at every day was someone's worst day.
—Amy Hawn Nelson, Actionable Intelligence for Social Policy

Ideologically, data projects should be conceived from a robust understanding of people's lived reality and their communities' historical relationship with data. This requires that the work be conducted by people who are committed to humanizing those whose lives are reflected in the data they are manipulating and respecting their stated needs and boundaries.[32] Rooting a project in the needs of the community rather than of the researcher is a critical component of empathy. As Jayaraman shared, researchers need to be asking, "How does this become practically useful for people in their lifetimes?" She continued, "That's where the value lies. Thinking about, 'What is the value of doing this?' It is for human beings in their lives to live."

[Asking], "How does this become practically useful for people in their lifetimes?" That's where the value lies. Thinking about, "What is the value of doing this?" It is for human beings in their lives to live. —Shruti Jayaraman, Chicago Beyond

Empathy should also show up in **action**: researchers and data analysts need to ensure that people reflected in the data are part of the initial decisions about whether imputation is right approach to disaggregating data as well as ensuing data analysis, and that they are fully aware of progress, results, and eventual uses of data and decisions to share it (box 3). As Chang shared, "Empathy is necessary but insufficient, in and of itself. There also needs to be action." Building in community-review processes for data work, including imputation projects, is one action that supports empathy. For surname databases and other databases that may be used to predict race, ethnicity, or other variables, community members can help cocreate lists or double-check existing lists for accuracy.

BOX 3

Community Engagement in Data Work

Our interviewees and the literature routinely identified community engagement as an important practice in racially equitable data work, including imputation. Community engagement is a core element of equity and empathy in that, if approached robustly and authentically, it helps return ownership and control of community data to those whose lives are reflected in those data. As a result, community engagement makes data and analysis more accurate by more fully representing people and their concerns.

As Chris Kingsley of the Annie E. Casey Foundation shared, "What empathy would mean is trying to give more discretion and control back to some of the key advocacy groups that represent the interests of Asian Americans, Black Americans, or Native Americans." He suggested creating a council of representatives from impacted communities that "first and foremost honored their own decisions" about how an imputation (or any other research project) might proceed and what is permissible in that project, rather than leaving those decisions to teams of researchers or managerial government staff that are predominantly white.

Although the literature on how to approach community engagement in large-scale data projects is scarce, Yeshi Milner and Data for Black Lives, Our Data Bodies and the Detroit Technology Project, Chicago Beyond, and other community organizations have made a powerful case for returning big data back to the people for community control and auditing. For a truly equitable engagement process, it is important that analysts meet with community members not simply as an obligation to fulfill, but as a generative and critical element of the research process. As Shruti Jayaraman of Chicago Beyond

shared, "There is infrastructure around self-reflection and internal bias and knowledge that you can equip people with. Just literally [asking], 'Who am I bumping into and whose voices am I hearing?' because if I'm only hearing my own [voice] and [that of] people who think and look like me, then I'm probably not developing empathy with somebody else. So, there are things one can do, and it is very important to do them. But if it becomes [purely] mechanical, it misses the point of what empathy is."

Sources: Urban Institute stakeholder interviews conducted between February 2021 and April 2021.

Community-review processes are critical for mitigating harm because they raise flags and issues that data scientists might otherwise overlook because of their own lived experience and position outside the community. As Wong shared, "Keeping updated through conversations [with community members and organizations] about potential concerns is really the best [way to mitigate] potential harms. Hearing from the community, 'What are the potential harms?' and figuring out [whether those are] things we can address or not. People within the community usually can do a better job of that kind of identification. Community partnerships, I think, have been critical to being able to make those lists more accurate and comprehensive." Ensuring that people reflected in data can participate in data work allows a community to be more invested and involved in the problem-solving process, helps produce more accurate data that serve the community's needs, and encourages ownership over a shared process and output.

Examining one's own process and considering improvements that take into account relationships to privilege and power is a form of empathy. As researchers, we must prompt *ourselves* to remember how powerful institutions and people have used data and research to exploit communities, how data can be flawed, and how racist and classist foundations of thinking and power have historically gone unchecked. In combination with a methodology that incorporates community review and input, this self-critical lens will ensure that data work is more equitable and robust.[33] In the absence of self-reported race and ethnicity data, imputed data enable users to more clearly understand disparities. But it is important to remember that race is a social and historical construction, meaning researchers and policymakers need to be clear about its connection to the correlated structural factors that likely drive the racial disparities they are examining. This includes thinking critically about how structural racism can be embedded in the data sources used in analysis (like imputation or algorithmic systems) to avoid reproducing racial biases in imputed data, and acknowledging the historical origins of disparities when analyzing imputed data to avoid ascribing structural issues to individuals' racial identities.[34]

Conclusion

In policy areas such as health care, wealth, and the justice system, critical gaps exist between the data we have and the data we need to identify and address racial disparities. Multiple approaches can be used to address that need, including data linkage, machine learning, and even the direct collection of race and ethnicity information from people and communities. Imputation can be more cost- and time-efficient than those approaches and can provide more information about racial disparities from datasets for which directly collecting self-reported race and ethnicity is especially challenging or impossible (e.g., credit bureau data, tax data, and historical data).

The analyses and impacts that can result from imputed race and ethnicity data can produce important insights for advancing policy and equity. But adding race and ethnicity to a dataset can also harm people and communities, including by violating privacy and confidentiality (which can cause people to be reidentified), by producing misleading conclusions from inaccurate estimates, and by excluding impacted communities from designing analyses and applying findings.

In our landscape scan, we found that specific and actionable guidance on how to incorporate community engagement into imputation and other analytic projects is especially lacking. The literature we examined consistently recommends that analysts incorporate community feedback into their projects and engage the people and communities reflected in the data,[35] but concrete methods for engagement are missing.[36] Some literature identifies steps of the imputation process where stakeholders should be engaged (e.g., releasing findings to the affected community or people) but does not specify a preferred method of engagement, how feedback should be collected, or how it should affect the project or research conclusions.[37] Specific guidance usually extends only to identifying types of stakeholders and community organizations researchers should engage (e.g., community leaders, activists, tribal governments),[38] not to recommendations on how or at what points in the research process analysts should engage.

In a related report, we discuss standards and recommendations for addressing the risks involved in imputation. The primary way to make imputation more ethical and empathetic is to engage the communities whose race and ethnicity will be imputed and who will be most directly impacted by policymaking resulting from the imputation. Researchers can and should involve community stakeholders throughout the process, be they local community members or representatives from national organizations involved in racial equity advocacy and policymaking. It is important to involve them throughout the process, from deciding whether imputation is the right approach to determining use cases and applications to accessibly sharing results. That partnership allows the people who need

the imputed data to better understand the issues in their own communities to determine how the data are developed and used in ways that benefit them. Their meaningful involvement alleviates several of the risks of imputation, particularly those around ownership of data, doing harm to communities, and acknowledging the personhood of people and communities in developing and deploying imputed data. And their participation and engagement not only strongly increases the likelihood of an ethical and empathetic approach to analysis, it also increases the accuracy of the data and analysis by more fully incorporating the insights and reality of their lived experience.

In addition to community involvement, external accountability around methodological rigor and risk mitigation remains necessary. Some accountability can be built into enhanced or more vigorous institutional review board review, which would be especially helpful to researchers at institutions covered by such boards. But all researchers, whether at an institution with an institutional review board or not, who want to impute race and ethnicity data to understand and address disparities should seek out peers in the same or closely related research and policy areas who can review methodological decisions and techniques, weigh the fitness of the imputed data for potential uses, and warn against potential privacy violations and risk of bad-faith misuse.

Importantly, researchers should collaborate with community members and other accountability partners to decide whether imputation is the right approach to filling missing data and whether the risks outweigh the benefits. Those risks include the risk of maintaining the status quo; for instance, in some situations, a lack of disaggregated data could inhibit understanding of or could even harm communities, particularly relatively less populous communities such as Native American, Asian American, and Pacific Islander communities, for which data are often missing even in analyses that include other racial and ethnic groups. But if the risks are weighed and accounted for, imputation can be a powerful tool for the necessary disaggregation of data that are missing from important datasets.

Imputation has important limitations, especially compared with self-reported race and ethnicity information that can be directly collected or appended from other sources. But the advantages of being able to do imputation more efficiently than collecting new data offer an important opportunity to inform policymaking more quickly in areas where the scale and reach of racial disparities are insufficiently understood.

Notes

[1] In theory, for instance, property taxes are race neutral, levied equally on all homeowners in a jurisdiction through an assessment process that does not take race into account. In reality, Black and Latine homeowners pay significantly more in taxes relative to fair market value than white homeowners; see Avenancio-León and Howard's *The Assessment Gap: Racial Inequalities in Property Taxation*.

[2] Ibram X. Kendi, "What the Racial Data Show," *The Atlantic*, April 6, 2020, https://www.theatlantic.com/ideas/archive/2020/04/coronavirus-exposing-our-racial-divides/609526/; Leana S. Wen and Nakisa B. Sadeghi, "Addressing Racial Health Disparities in the COVID-19 Pandemic: Immediate and Long-Term Policy Solutions," Health Affairs blog, July 20, 2020, https://www.healthaffairs.org/do/10.1377/hblog20200716.620294/full/.

[3] See "The Impact of COVID-19 on Black Communities," Data for Black Lives, accessed July 19, 2021, https://d4bl.org/covid19-data; and Aletha Maybank, "Why Racial and Ethnic Data on COVID-19's Impact Is Badly Needed," American Medical Association, April 8, 2020, https://www.ama-assn.org/about/leadership/why-racial-and-ethnic-data-covid-19-s-impact-badly-needed.

[4] For more on the racial homeownership gap, see Choi et al. (2019). For a discussion on the lack of race and ethnicity in tax data, see Bearer-Friend (2019). For more on problems with employer credit checks, see Traub and McElwee (2016).

[5] For instance, see President Biden's executive order on racial equity in the federal government, Advancing Racial Equity and Support for Underserved Communities Through the Federal Government, 86 Fed. Reg. 7009 (2021).

[6] For examples, see Dembosky et al. (2019) and Cook et al. (2021).

[7] We use Latine in this report because it is more gender inclusive than Latino/Latina, and because some Spanish speakers may find Latine more easily pronounceable than Latinx (Schwabish and Feng 2021). We recognize and appreciate that not all Latino/Latina or Hispanic people may identify with the term, that language is constantly evolving, and that our efforts to inclusively capture the collective identities and heritages of peoples and cultures will continue to improve and evolve.

[8] For more information on the Equal Credit Opportunity Act of 1974, see this blog post from the Consumer Financial Protection Bureau.

[9] For example, the Census Bureau's Household Pulse Survey collects race and ethnicity information, but the sample size limits any disaggregation lower than the metropolitan-areas geography, and even then is limited to only the 15 largest metropolitan areas in the United States.

[10] See the Census Bureau's definitions of racial categories.

[11] For an example of using zip code, see Urban's 2021 feature "Debt in America: An Interactive Map."

[12] See also Ziyuan Zhong, "A Tutorial on Fairness in Machine Learning," Towards Data Science, October 21, 2018. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.

[13] Data linkage and data integration, which involve attaching one dataset to another, are common approaches to generating data disaggregated by race and ethnicity, and the literature has been paying more attention to best practices for ethically applying them in racial equity analysis. See, for example, work from Actionable Intelligence for Social Policy at the University of Pennsylvania, including Hawn Nelson, Jenkins, Zanti, Katz, and Berkowitz (2020).

[14] We began by conducting a simple internet browser search for specific search terms overlapping with these categories, and by soliciting recommended names and organizations from Urban Institute researchers with expertise in data equity, racial equity analysis, data science, and community-engaged methods.

[15] For definitions of big data, see this resource from the University of Wisconsin and this article from Built In.

[16] While all informants have agreed to be named, and those that are quoted agreed to have their quotes be attributed, the analyses presented reflect the views of the authors and should not be directly attributed to the informants.

[17] For more on equitable data practices, see Gaddy and Scott (2020).

[18] See, for example Kim, Gao, and Rzhetsky (2018), who impute race and ethnicity data onto anonymized electronic medical records as a mechanism for further studying medical discrimination and health disparities.

[19] For more examples, see Benjamin (2016, 2019) and Jefferson (2018).

[20] See "Data Justice," Detroit Community Technology Project, accessed July 20, 2021, https://detroitcommunitytech.org/?q=datajustice.

[21] As Milner writes, "Big data was necessary to distance oneself from the violence and the gore capitalism of slavery." See Yeshi Milner, "Abolition Means the Creation of Something New," Medium, December 31, 2019, https://medium.com/@YESHICAN/abolition-means-the-creation-of-something-new-72fc67c8f493.

[22] Milner, "Abolition Means the Creation of Something New."

[23] Milner, "Abolition Means the Creation of Something New."

[24] See, for example, Lee et al. (2016) and Kennickell (1998).

[25] See, for example, Kim, Gao, and Rzhetsky (2018).

[26] See, for example, Hawn Nelson, Jenkins, Zanti, Katz, and Burnett et al. (2020), Hawn Nelson, Jenkins, Zanti, Katz, and Berkowitz (2020), NFES (2016), and Pike (2020).

[27] See, for example, Benjamin (2019), Jefferson (2018), and Hawn Nelson, Jenkins, Zanti, Katz, and Burnett et al. (2020), who critique various methods that rely on existing, often racially and statistically biased, data to make predictive choices. Although none of these papers focus specifically on imputation, their critique is applicable.

[28] See, for example, Benjamin (2016), Hawn Nelson, Jenkins, Zanti, Katz, and Berkowitz (2020), Leslie (2019), and Lucero and Roubideaux (2020).

[29] Zhang et al. (2017) also raise this concern in their discussion on statistical uncertainty.

[30] See, for example, Lee et al. (2016), Lee, Ramakrishnan, and Wong (2018), Pike (2020), and Zhang et al. (2017) which touch on reidentification and risks associated with anonymized data and weighing the benefits of imputation.

[31] For further discussion of the history of race and big data, see Milner, "Abolition Means the Creation of Something New."

[32] See, for example, Benjamin (2016), Knight et al. (2021), and Milner, "Abolition Means the Creation of Something New."

[33] See NFES (2016).

[34] For a recent discussion of problematic facial recognition software, see Lauren Palmer and Annie Feiner, "Rules around Facial Recognition and Policing Remain Blurry," CNBC, June 12, 2021, https://www.cnbc.com/2021/06/12/a-year-later-tech-companies-calls-to-regulate-facial-recognition-met-

with-little-progress.html. For a discussion of Amazon's sexist AI-driven recruiting tool, see BBC, "Amazon scrapped 'sexist AI' tool," October 10, 2018, https://www.bbc.com/news/technology-45809919.

[35] Nucera and Sonnenberg (2017) seek to return "ownership" of big data from exploitative institutions to communities and people through data education, advocacy, and organizing efforts.

[36] See, for example, Gaddis (2019) and Gibbs et al. (2017).

[37] See, for example, Gaddis (2019) and Leslie (2019).

[38] See, for example, Lucero and Roubideaux (2020).

References

Anson-Dwamena, Rexford, Priya Pattah, and Justin Crow. 2020. "Imputing Missing Race and Ethnicity Data in COVID-19 Cases." Richmond: Virginia Department of Health.

Bearer-Friend, Jeremy. 2018. "Should the IRS Know Your Race? The Challenge of Colorblind Tax Data." Tax Law Review 73 (1): 1–68.

Benjamin, Ruha. 2016. "Informed Refusal: Toward a Justice-Based Bioethics." *Science, Technology, & Human Values* 41 (6): 967–90. https://doi.org/10.1177/0162243916656059.

———. 2019. "Assessing Risk, Automating Racism." *Science* 366 (6464): 421–22. https://doi.org/10.1126/science.aaz3873.

Biden Jr., Joseph R. 2021. "Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government." Washington, DC: The White House. "Executive Order 13228 of October 8, 2001, Establishing the Office of Homeland Security and the Homeland Security Council," Code of Federal Regulations, title 3 (2001): 796-802, http://www.gpo.gov/fdsys/pkg/CFR-2002-title3-vol1/pdf/CFR-2002-title3-vol1-eo13228.pdf.

Brown, K. Steven, LesLeigh D. Ford, and Shena Ashley. 2021. *Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity: Recommendations and Standards Guide*. Washington, DC: Urban Institute.

Choi, Jung Hyun, Caitlin Young, Alanna McCargo, Michael Neal, Laurie Goodman, and Caitlin Young. 2019. *Explaining the Black-White Homeownership Gap*. Washington, DC: Urban Institute.

Cook, Michael B., Lauren M. Hurwitz, Ashley M. Geczik, and Eboneé N. Butler. 2021. "An Up-to-Date Assessment of US Prostate Cancer Incidence Rates by Stage and Race: A Novel Approach Combining Multiple Imputation with Age and Delay Adjustment." *European Urology* 79 (1): 33–41. https://doi.org/10.1016/j.eururo.2020.09.041.

Dembosky, Jacob W., Amelia M. Haviland, Ann Haas, Katrin Hambarsoomian, Robert Weech-Maldonado, Shondelle M. Wilson-Frederick, Sarah Gaillot, and Marc N. Elliott. 2019. "Indirect Estimation of Race/Ethnicity for Survey Respondents Who Do Not Report Race/Ethnicity." *Medical Care* 57 (5): e28–33. https://doi.org/10.1097/MLR.0000000000001011.

Derenoncourt, Ellora. 2019. "Can You Move to Opportunity? Evidence from the Great Migration." Working Paper.

Fremont, Allen, Joel S. Weissman, Emily Hoch, and Marc N. Elliott. 2016. "When Race/Ethnicity Data are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities."

Gaddis, S. Michael. 2019. "Understanding the 'How' and 'Why' Aspects of Racial-Ethnic Discrimination: A Multimethod Approach to Audit Studies." *Sociology of Race and Ethnicity* 5 (4): 443–55. https://doi.org/10.1177/2332649219870183.

Gaddy, Marcus, and Kassie Scott. 2020. "Principles for Advancing Equitable Data Practice." Washington, DC: Urban Institute.

Gibbs, Linda, Amy Hawn Nelson, Erin Dalton, Joel Cantor, Stephanie Shipp, and Della Jenkins. 2017. "IDS Governance: Setting Up for Ethical and Effective Use." Philadelphia: Actionable Intelligence for Social Policy.

Harris, Ada. 2020. "Using Bayesian Improved Surname Geocoding (BISG) to Classify Race and Ethnicity in Administrative Employment Data by Industry: A Validation Study." Alexandria, VA: American Statistical Association.

Hawn Nelson, Amy, Della Jenkins, Sharon Zanti, Matthew Katz, T. C. Burnett, Dennis Culhane, and Katie Barghaus et al. 2020. "Introduction to Data Integration and Sharing." Philadelphia: Actionable Intelligence for Social Policy. https://www.aisp.upenn.edu/wp-content/uploads/2020/06/AISP-Intro-.pdf.

Hawn Nelson, Amy, Della Jenkins, Sharon Zanti, Matthew Katz, and Emily Berkowitz. 2020. "A Toolkit for Centering Racial Equity Throughout Data Integration." Philadelphia: Actionable Intelligence for Social Policy.

Hill, Ian, Sarah Benatar, Brigette Courtot, Fredric Blavin, Embry M. Howell, Lisa Dubay, Bowen Garrett, … and Mark Rouse. 2015. *Strong Start for Mothers and Newborns Evaluation*. Washington, DC: Urban Institute.

Hwang, Jackelyn, Michael Hankison, and Kreg Steven Brown. 2015. "Racial and Spatial Targeting: Segregation and Subprime Lending within and across Metropolitan Areas." *Social Forces* 93 (3): 1081–108. https://doi.org/10.1093/sf/sou099.

Jefferson, Brian Jordan. 2018. "Predictable Policing: Predictive Crime Mapping and Geographies of Policing and Race." *Annals of the American Association of Geographers* 108 (1): 1–16. https://doi.org/10.1080/24694452.2017.1293500.

Kennickell, Arthur B. 1998. "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances." Dallas: Board of Governors of the Federal Reserve System.

Kim, Ji-Sung, Xin Gao, and Andrey Rzhetsky. 2018. "RIDDLE: Race and Ethnicity Imputation from Disease History with Deep LEarning." *PLOS Computational Biology* 14 (4): e1006106. https://doi.org/10.1371/journal.pcbi.1006106.

Knight, Hannah E., Sarah R. Deeny, Kathryn Dreyer, Jorgen Engmann, Maxine Mackintosh, Sobia Raza, Mai Stafford, Rachel Tesfaye, and Adam Steventon. 2021. "Challenging Racism in the Use of Health Data." *The Lancet Digital Health* 3 (3): e144–46. https://doi.org/10.1016/S2589-7500(21)00019-4.

Lee, Jennifer, and Karthick Ramakrishnan. 2020. "Who Counts as Asian." *Ethnic and Racial Studies* 43 (10): 1733–56. https://doi.org/10.1080/01419870.2019.1671600.

Lee, Jennifer, Karthick Ramakrishnan, and Janelle Wong. 2018. "Accurately Counting Asian Americans Is a Civil Rights Issue." *The ANNALS of the American Academy of Political and Social Science* 677 (1): 191–202. https://doi.org/10.1177%2F0002716218765432.

Lee, Katherine J., Gehan Roberts, Lex W. Doyle, Peter J. Anderson, and John B. Carlin. 2016. "Multiple Imputation for Missing Data in a Longitudinal Cohort Study: A Tutorial Based on a Detailed Case Study Involving Imputation of Missing Outcome Data." *International Journal of Social Research Methodology* 19 (5): 575–91. https://doi.org/10.1080/13645579.2015.1126486.

Leslie, David. 2019. "Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector." SSRN Scholarly Paper ID 3403301. Rochester, NY: The Alan Turing Institute.

Lucero, Julie E, and Yvette Roubideaux. 2020. "Holding Space for All of Us." *AMA Journal of Ethics* 22 (10): 882–87.

Maybank, Aletha. 2020. "Why Racial and Ethnic Data on COVID-19's Impact Is Badly Needed." American Medical Association. April 8, 2020. https://www.ama-assn.org/about/leadership/why-racial-and-ethnic-data-covid-19-s-impact-badly-needed.

Milner, Yeshi. 2019. "Abolition Means the Creation of Something New." Medium. *Medium* (blog). December 31, 2019. https://medium.com/@YESHICAN/abolition-means-the-creation-of-something-new-72fc67c8f493.

Nechamkin, Emma, and Graham MacDonald. 2019. "Predicting Zoned Density Using Property Records." Washington, DC: Urban Institute.

NFES (National Forum on Education Statistics). 2016. "Forum Guide to Collecting and Using Disaggregated Data on Racial/Ethnic Subgroups." NFES 2017-017. Washington, DC: National Center for Education Statistics.

Oglesby-Neal, Ashlin, Emily Tiry, and KiDeuk Kim. 2019. "Public Perceptions of Police on Social Media." Washington, DC: Urban Institute.

Nucera, Diana, and Kristyn Sonnenberg, eds. 2017. *Opening Data*, vol. 2. Detroit: Detroit Community Technology Project.

Palmer, Lauren, and Annie Feiner. 2021. "Rules around Facial Recognition and Policing Remain Blurry." *CNBC*, June 12, 2021, sec. Technology. https://www.cnbc.com/2021/06/12/a-year-later-tech-companies-calls-to-regulate-facial-recognition-met-with-little-progress.html.

Petty, Tawana, Mariella Saba, Tamika Lewis, Seeta Peña Gangadharan, and Virginia Eubanks. 2018. *Our Data Bodies: Reclaiming Our Data*. Our Data Bodies.

Pike, Elizabeth R. 2020. "Defending Data: Toward Ethical Protections and Comprehensive Data Governance." *Emory Law Journal 69* (4): 687–743.

Randall, Megan, Alena Stern, and Yipeng Su. 2021. "Five Ethical Risks to Consider before Filling Missing Race and Ethnicity Data." Washington, DC: Urban Institute.

Saleiro, Pedro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. 2019. *Aequitas: A Bias and Fairness Audit Toolkit*. Chicago: University of Chicago.

Schwabish, Jon, and Alice Feng. 2021. *Do No Harm Guide: Applying Equity Awareness in Data Visualization*. Washington, DC: Urban Institute.

Stern, Alena, and Ajjit Narayanan. 2021. *Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity: A Case Study Imputing Credit Bureau Data*. Washington, DC: Urban Institute.

Theodos, Brett, and Eric Hangen. 2017. "Expanding Community Development Financial Institutions." Washington, DC: Urban Institute.

Traub, Amy, and Sean McElwee. 2016. Bad Credit Shouldn't Block Employment: How to Make State Bans on Employment Credit Checks More Effective. New York: Demos.

Zhang, Xinzhi, Eliseo J. Pérez-Stable, Philip E. Bourne, Emmanuel Peprah, O. Kenrik Duru, Nancy Breen, David Berrigan, Fred Wood, James S. Jackson, David W. S. Wong, and Joshua Denny. 2017. "Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century." *Ethnicity & Disease* 27 (2): 95–106. https://doi.org/10.18865/ed.27.2.95.

About the Authors

Kreg Steven Brown is a senior research associate in the Center on Labor, Human Services, and Population and associate director of the Racial Equity Analytics Lab at the Urban Institute. He regularly leads and collaborates on research and policy advising projects that explore sources of and solutions to racial inequalities in economic opportunity. His primary research focuses on employment, particularly career pathways, earnings, and workplace protections; financial well-being; and economic mobility. Previous work includes federal evaluation and academic research on segregation and homeownership, access to affordable housing, and educational equity. Brown received his bachelor's from Princeton University and his master's in sociology from Harvard University.

Yipeng Su is a research associate in the Metropolitan Housing and Communities Policy Center at the Urban Institute. Her research interests include housing, economic development, and the intersection between urban planning and technology. She holds a master's degree in public administration from New York University's Wagner Graduate School of Public Service.

Jahnavi Jagannath is a policy analyst in the Justice Policy Center at the Urban Institute. Her work focuses on community safety to reduce need of police and correctional systems through arts and placemaking efforts, understanding pathways to incarceration for women, and strengthening family connections within correctional facilities. She has a BA in Policy Studies and Sociology from Rice University.

Jacqueline Rayfield is a policy assistant in the Center on Labor, Human Services, and Population at the Urban Institute, where she works on apprenticeships and labor force development. Her previous research analyzed education and job training programs for Syrians in Istanbul. She graduated from Boston University with a degree in international relations and minors in computer science and French.

Megan Randall is a research associate in the Urban Institute's Research to Action Lab, where she works on federal place-based programs, state and local economic development policy, and inclusive economic recovery. She also provides research and project management support to Urban's Racial Equity Analytics Lab. Before joining Urban, she worked on state health care policy and advocacy at the Center for Public Policy Priorities (now Every Texan), and state and local housing and disaster relief policy at Texas Appleseed. Randall graduated with a bachelor's degree in political science from the

University of California, Berkeley, and earned master's degrees in public affairs and in community and regional planning from the University of Texas at Austin.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org