I. EXECUTIVE SUMMARY

Race is a "power construct of collected or merged differences that lives socially."[1] Race is neither a biological fact[2] nor a "fixed trait,"[3] and conceptions of race change over time.[4] In this sense, race is not "real." However, racism is real. The collection and study of data concerning racial and ethnic[5] inequities and disparities allows us to better understand experiences of racism. Racial and ethnic data collection is thus "a crucial step in fighting racism and structural racial inequality."[6]

Comprehensive demographic data on racial and ethnic disparities and inequities across critical policy areas allow us to see more clearly how and where racism manifests, including in access to resources and exposure to harms.[7] Studying the effects of racism, in turn, allows us to identify the policies that create and perpetuate inequities and to craft antiracist interventions and alternatives. Alongside racial and ethnic data, we need data concerning many other characteristics and variables that are connected to experiences of oppression and subordination, such as gender identity, disability, and socioeconomic status. Data collection across variables such as these provides a better understanding of the nuances of existing inequities and enables robust intersectional research.

Unfortunately, the current state of racial and ethnic data collection and reporting in the United States reflects many gaps and deficiencies that hinder antiracist policymaking. The efforts of the Boston University Center for Antiracist Research (the "Center") to collect race and ethnicity data across key policy areas confirms that existing data sources are inadequate. State and local sources of racial and ethnic data often vary in standards for how to report, "what to report, when to report it, and even whether to report it at all."[8] Moreover, lack of coordination regarding data collection at the federal level "can result in contradictory guidance to local and state agencies,"[9] further compounding the challenges of inconsistent, incomplete, and inaccurate race and ethnicity data collection and reporting practices at the local and state levels. Additionally, in the rare instances when states are required to report racial and ethnic data to the federal government—for instance, as part of their involvement in Medicaid—such requirements have historically not been well enforced.[10]

The lack of uniform and robust standards for racial and ethnic data collection and reporting has meant that existing data repositories are incomplete, contain errors, are usually incompatible with each other, and are often internally inconsistent across years, jurisdictions, subjects, and levels of analysis.[11] Moreover, a great deal of race and ethnicity data are not publicly available or easily accessible. Some entities are reluctant to make any changes to their own practices until a more centralized and standardized system is in place.[12]

In early Spring of 2020, Center staff[13] and *The Atlantic's* COVID Tracking Project team saw that the poor state of race and ethnicity data collection and reporting in the United States was hindering meaningful evidence-based policy responses to the COVID-19 pandemic. Among other things, the lack of robust data on COVID-19 cases, hospitalizations, and deaths by race and ethnicity was preventing policymakers from identifying and responding to resource inequities for their jurisdictions. The two entities quickly collaborated to fill this gap through the COVID Racial Data Tracker (CRDT), which collected race and ethnicity data on COVID-19 cases, hospitalizations, deaths, and to a lesser extent, testing (hereinafter collectively referred to as "COVID-19 outcomes") in the United States from April 2020 to March 2021.[14] While the CRDT was active, it was the most comprehensive and timely source of COVID-19 racial and ethnic data in the country, and was frequently cited by news outlets, government representatives, advocates, and scholars.

The Center then expanded its race and ethnicity data collection efforts through its Racial Data Tracker (RDT), which collects data on other key issue areas such as houselessness, criminal arrests, and police violence. This data collection work builds upon decades of effort by scholars and advocates from a variety of disciplines to obtain accurate and complete racial and ethnic data to inform evidence-based policy making. Public health scholars and advocates have been at the forefront of the call for better racial and ethnic data,[15] but the need extends to all issue areas in which people of color have been pushed to the margins. Indeed, race and ethnicity data are often strategically and deliberately omitted in order to preserve existing inequities, or neglected out of convenience or indifference—all of which works to perpetuate racism.

The CRDT and RDT teams' experiences provide great insight into the deficiencies of the racial and ethnic data collection and reporting methods used by local, state, and federal agencies, as well as the variety of methods they employ. These teams' work illustrates the need for a single, standardized, nationwide system of data collection and reporting by race and ethnicity.[16] Below, we summarize the challenges the teams faced and offer policy recommendations, which are discussed in depth in the following sections of this Report.

A. The CRDT Team's Findings Regarding Racial and Ethnic Data Collection and Reporting

The CRDT team's experience manually collecting state-reported racial and ethnic data on COVID-19 outcomes demonstrates that such state-reported data suffer from deficiencies that can cause errors and underestimations of racial and ethnic inequities. The CRDT team encountered the following challenges:

1. Data were incomplete. Many states failed to report any racial and ethnic data on COVID-19 outcomes for several months after the outbreak, and some states *never* reported such information for the duration of the CRDT collection period (April 2020 to March 2021). States that did report racial and ethnic data were not consistent about whether and how they did so from one month to the next. When states did report, race and ethnicity data were often only available for some of the COVID-19 outcomes (cases, hospitalizations, tests, or deaths). Additionally, for each of these outcomes, the data were often incomplete. This means, for example, that the race and ethnicity of every known person who was hospitalized for COVID-19 were not necessarily recorded. Likewise, different states failed to collect data concerning various racial or ethnic groups. For example, some states did not report any data concerning Hispanic/Latino/a/e/x[17] people. Such incomplete race and ethnicity data within and across states impede thorough analysis of national trends and preclude comparisons across jurisdictions.

2. States' data reporting methods varied. States varied not only in terms of what race and ethnicity data they reported and whether they reported, but also in *how* they presented their data to the public. This lack of uniformity across jurisdictions presents a barrier for those who are studying national trends. Moreover, some of the states' reporting conventions did not provide sufficient information. For example, some states presented the data *only* as percentages (rounded to certain decimal points or to the nearest whole number) without making raw numbers available. Reporting methods that do not make raw numbers available to the public can cause estimation and calculation errors that make it impossible for researchers, policymakers, and advocates to accurately interpret the full extent of racial and ethnic inequities.

3. States did not adequately account for the ways that race and ethnicity can intersect. Many states' reporting methods did not account for the fact that some people fall into both a racial *and* an ethnic group, or into multiple racial or ethnic groups. Those that did had a variety of approaches in how they handled this complexity, with varying degrees of effectiveness. Failure to carefully present race and ethnicity as intersecting measures when appropriate can lead to errors such as double counting individuals or undercounting group membership. This was particularly the case for the Hispanic/Latino/a/e/x and multiracial groups.

4. States used non-standard racial and ethnic categories. States were not consistent in how they categorized race and ethnicity. Some failed to count certain racial and ethnic categories outright, and others lumped categories together in nonstandard and overly broad ways, both of which obscure experiences of racism and subordination.

5. States reported the bare minimum, or less. Most states, at best, provided information about racial and ethnic categories that are included in the federal Office of Management and Budget (OMB) racial and ethnic categories ("OMB categories"),[18] which are used for certain federal data collection purposes such as the Decennial Census. Only a few jurisdictions departed from the OMB to provide more detailed race and ethnicity breakdowns, while many jurisdictions provided even fewer race and ethnicity categories than those outlined by the OMB.

6. States infrequently updated their data. States varied in terms of when and how often they updated their race and ethnicity data, making it difficult to compare trends across states. Many states updated data infrequently, preventing real-time assessments of health inequities.

B. The RDT Team's Findings Regarding Racial and Ethnic Data Collection and Reporting

Complementing the work of the CRDT, the RDT's examination of datasets regarding houselessness, criminal arrests, and police violence[19] confirms that many existing public datasets that report information by race and ethnicity are insufficient to inform policy choices. Indeed, despite ostensibly "national" data collection, these data—like COVID-19 information—are collected in a highly decentralized and uncoordinated way by state and local entities. The datasets examined for this study contained the following problems:

1. Data are incomplete. National datasets that provide race and ethnicity information frequently rely on local, regional, and state reporting. Since participation in such reporting schemes is voluntary, many entities do not contribute data to these sources, or fail to include racial and ethnic information when they do report data. As a result, existing datasets with race and ethnicity information are often incomplete, especially when assessed longitudinally (across time). Furthermore, when data are missing (or inaccurate) at the state, regional, or local level, their aggregation results in data that are inaccurate and potentially at odds with data from higher-level sources.

2. It is often not possible to obtain data that are disaggregated at the local or state levels. National datasets often fail to disaggregate data at local or state levels. Moreover, some race and ethnicity data, such as data concerning houselessness, are gathered and organized by non-governmental organizations or federally-designated entities with arbitrary geographical boundaries that do not correspond to the boundaries of cities, towns, and the like. For example, such an entity's jurisdiction might cover multiple cities, or even parts of cities. These entities often cannot or do not disaggregate data at the city level, where important policy decisions are often made. As a result, policymakers cannot access data for their specific jurisdiction. The lack of disaggregated data can obscure important disparities that may exist at state or local levels, or between states and locales.

3. Longitudinal data are largely unavailable. Datasets with racial and ethnic information have varied and arbitrary time periods. This temporal patchwork prevents analysis of trends over time or the impacts of policy changes.

4. Methodologies are varied and incomplete. Existing datasets vary too much in their methodologies to be used side-by-side in an attempt to overcome the above-mentioned challenges.

C. Policy Recommendations

Based on the data deficiencies and challenges described above, we make the following policy recommendations,[20] which are discussed in more detail in Section V of this Report:

1. Centralize and standardize racial and ethnic demographic data collection and reporting across critical issue areas. Federal leadership is needed to create a single standardized, nationwide system of data collection and reporting by race and ethnicity in key policy areas including, but not limited to, health, housing, employment, education, the criminal legal system, and the environment.

2. Use existing federal race and ethnicity standards as a starting point, and regularly reevaluate and amend such standards. Existing federal race and ethnicity data standards should be the starting point for a centralized and standardized race and ethnicity data collection system. The federal government should review and amend these standards immediately, and continue to do so regularly with input from scholars, community members, and advocates. This process should include a close examination and reevaluation of the OMB racial and ethnic categories.

3. Create an oversight board to monitor how the government uses race and ethnicity data. An external oversight board, which should include community partners and advocates, should monitor the creation of a centralized data collection system to ensure that the government's use of the data complies with existing laws and is used responsibly, ethically, and equitably.

4. Incentivize consistent and timely state and local participation through adequate funding. The federal government should create funding incentives that encourage states and localities to collect and report racial and ethnic data to the centralized system in a consistent and timely manner. This funding should include conditional funds to incentivize participation, and additional funds to support data administration infrastructure.

5. Monitor noncompliance. States and other entities that choose not to participate in a centralized and standardized system of racial and ethnic data reporting should be monitored by the aforementioned oversight board. Data reporting dashboards, websites, reports, and similar platforms should explicitly note which entities have chosen not to participate in order to increase public visibility and accountability.

6. Make data available at the national level, and also disaggregate at the state and local levels. Race and ethnicity data should be collected and reported such that researchers and policymakers have the option of obtaining data aggregated at the national level and disaggregated at state and local levels.

7. Incentivize more granular race and ethnicity reporting. State and local entities should be incentivized to collect data at more granular levels than the OMB racial and ethnic categories to reflect the ethnic and racial makeup of their particular jurisdictions, and to reveal inequities between subpopulations. These granular data should be collected and disaggregated such that they can be collapsed into future permutations of the OMB categories as needed.

8. Report race and ethnicity data as intersecting measures. Detailed race and ethnicity data should be collected and reported as intersecting characteristics in order to provide the most clarity and flexibility regarding the breakdown of the data. For example, reported data should make clear how many people experiencing houselessness who fell within the Hispanic/Latino/a/e/x group also fell within the Black group, and how many did not. This system should not limit the number of ethnicities or races that may apply to each person.

9. Collect and report data across additional characteristics and variables. Experiences of racism are multidimensional. In order to better understand the nuances of racism and enable robust intersectional research, race and ethnicity data should be disaggregated by additional characteristics and variables, including, but not limited to, sex, gender identity, sexual orientation, age, disability, religion, income, linguistic usage, educational attainment, socioeconomic status, "street race,"[21] and national origin.

10. Tailor privacy-related data suppression practices to the realities of each dataset. Small numbers may need to be suppressed for privacy reasons in local or state databases, but may not need to be suppressed when they are aggregated at the national level. State and local entities that collect race and ethnicity data should remove any identifying information, such as names and birth dates, before reporting the data to a national database. Those entities should then report *all* disaggregated data to a national database through an encrypted platform, and the national database should decide whether suppression of small numbers is necessary to protect privacy at the national level. State and local entities should, of course, continue to suppress small numbers for their own, publicly-facing reporting when necessary for privacy reasons.

11. Make data publicly available and accessible. Apart from data that must be suppressed or protected due to privacy or serious confidentiality concerns, racial and ethnic data should be made freely, publicly, and easily accessible for use by advocates, scholars, policymakers, and others. All such data should contain explicit disclaimers and guidance on best practices concerning any gaps and limitations associated with them.