URBAN INSTITUTE · ELEVATE · THE · DEBATE

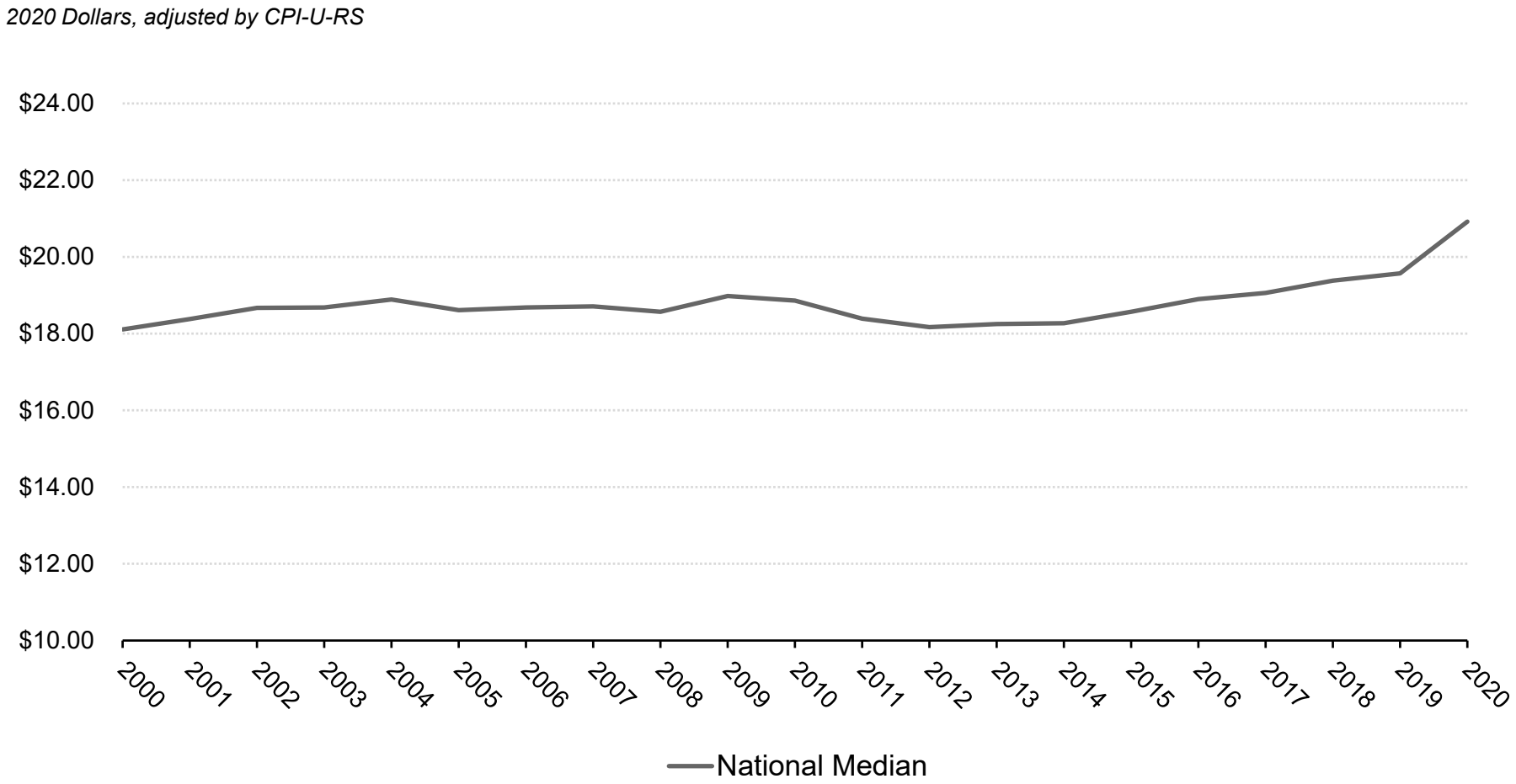
Approaches to Data Disaggregation to Advance Racial Equity

K. Steven Brown
National Academies of Science, Engineering, and Medicine
May 2022

Why is data important for racial equity?

Why is disaggregating data important?

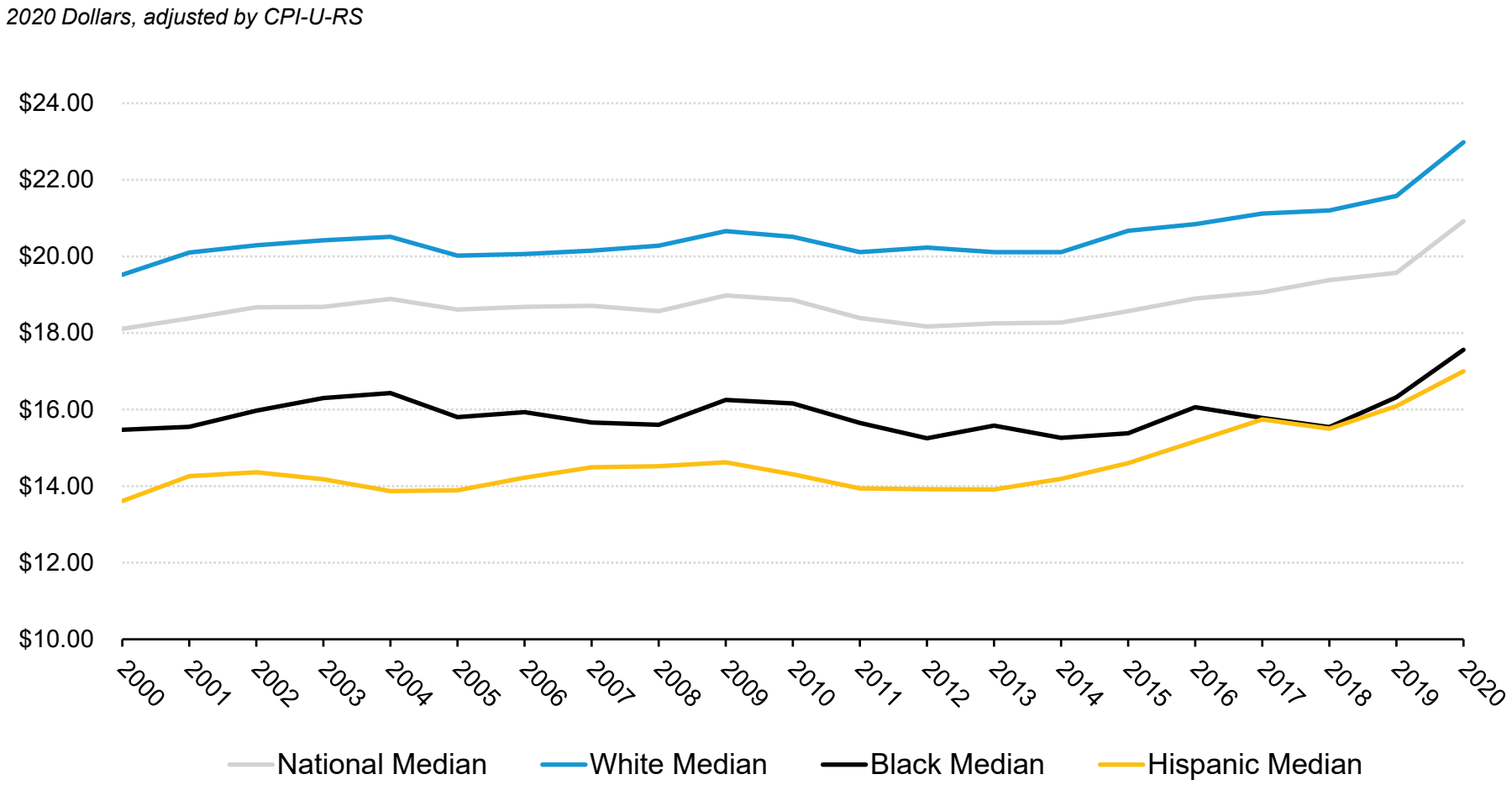
Example: Median Wages 2000-2020



Source: BLS

Why is disaggregating data important?

Example: Median Wages 2000-2020



Source: BLS

Data Landscape & Challenges

- “Nationally representative” data does not paint an accurate picture for counties and their residents
- To effectively measure racial disparities, **we need disaggregated data on people’s race**
- **Challenges:** disaggregated data are strong in some areas (employment, education) and lacking in others (health, wealth)
- There is never enough data that we want – **so how do we measure and make progress with what we have or can reasonably attain?**

Three alternatives for measuring structural racism when individual-level race is missing or insufficient

1. Consider the proximate "structure"

- Geographic proxies (e.g. zip code, neighborhood or tract) are valuable
  - **Strengths:** correlated highly with race due to persistent segregation; typically meets all the other markers of high-quality data
  - **Weaknesses:** works well for larger racial groups but can be lacking for smaller or more integrated ones (e.g. Native Americans or Asians); misses the geographic minorities (e.g. people of color in predominantly white neighborhoods)
- Other proxies include schools, labor market characteristics, housing tenure (renter vs homeowner)
- Historical events or prior policies that have created disparities
  - **Example:** redlining

2. Improve the existing data

- Tap into available administrative data
  - Does the county have an existing integrated data system?
  - Can it be used to link to other key data sets with necessary race information in a legal and ethical way?
- Data science techniques
  - Matching/linking to non-administrative data
  - Imputing the missing race data (BISG method)

3. Collect new data

- Can be challenging to get at scale, but important in the absence of reasonable alternatives
- Method of collection is critical, consider:
  - Quantitative vs. qualitative
  - Representativeness
  - Ability to be repeated
  - Who is involved in creating and framing the questions/study

Example of #1:
Considering the proximate “structure” via geography

February 2022

An introduction to the

Spatial Equity Data Tool

Spatial Equity Data Tool: What does it do?

- Assesses racial, economic, and geographic representativeness of user uploaded geographic point data
- Works at national, state, and county levels (current tool works at city level)
- Goal is to democratize quick data analysis capabilities

What can it be used for?

- Equity in allocation of place-based programs or resources (e.g. grocery stores, traffic accidents, funding allocation)
- Examine representativeness of program participants
- Identifying areas for future investment (e.g. safety data, previous investment)

Geographic distribution of your data compared with the **total population** in the state

Disparity Score

See which counties are over- and underrepresented in your data

KING COUNTY, WA

This county contains:

26.2% of your data points

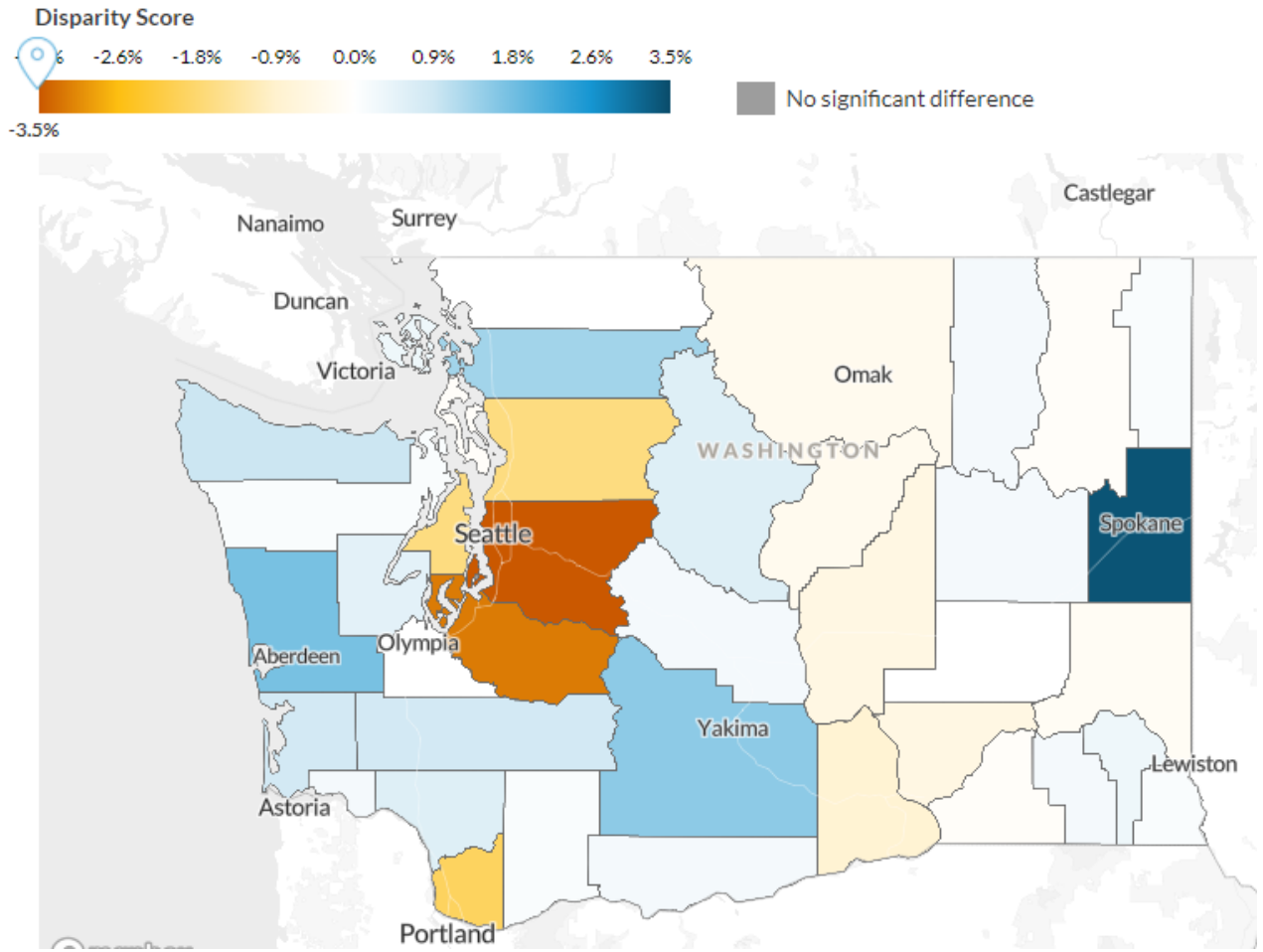
29.7% of the total population

-3.5% underrepresented

King County, WA has 3.5 percent fewer of the data points than we'd expect if the data were distributed in accordance with your baseline.

Data comparison

See your data side by side with the total population in the state



Limitations

- Can’t tell users **why** certain areas / groups are over/underrepresented
- Simplified measure of access to resources
- Only allows national, state, county, and city level datasets
- Built in baseline datasets may not match all use cases

Example of #2:
Improving the existing data through ethical and empathetic application of imputation

Project Overview

- Credit bureau data contains no racially identifying information
- Research finds many disparities in lending and access to credit
- Having disaggregated information on credit could be helpful…but it could also be harmful

Project Overview

- Credit bureau data contains no racially identifying information
- Research finds many disparities in lending and access to credit
- Having disaggregated information on credit could be helpful…but it could also be harmful
- Via imputation, we appended racial identifiers from the ACS onto credit bureau data
- Project goal was not just to see if we could do it accurately, but to see if we could handle technical process with sensitive information ***ethically and with empathy***.

What do we mean by ethics and empathy?

- **Ethics**
  - Balancing uses of disaggregated data with potential harms
  - Minimizing risks associated with disaggregation
  - More just, equitable distribution of resources
- **Empathy**
  - Thoughtfulness in engaging with and responding to communities represented in the data
  - Acknowledgement of the personhood of individuals and their experiences
  - Agency - efforts to make data available and useful to communities represented in the data

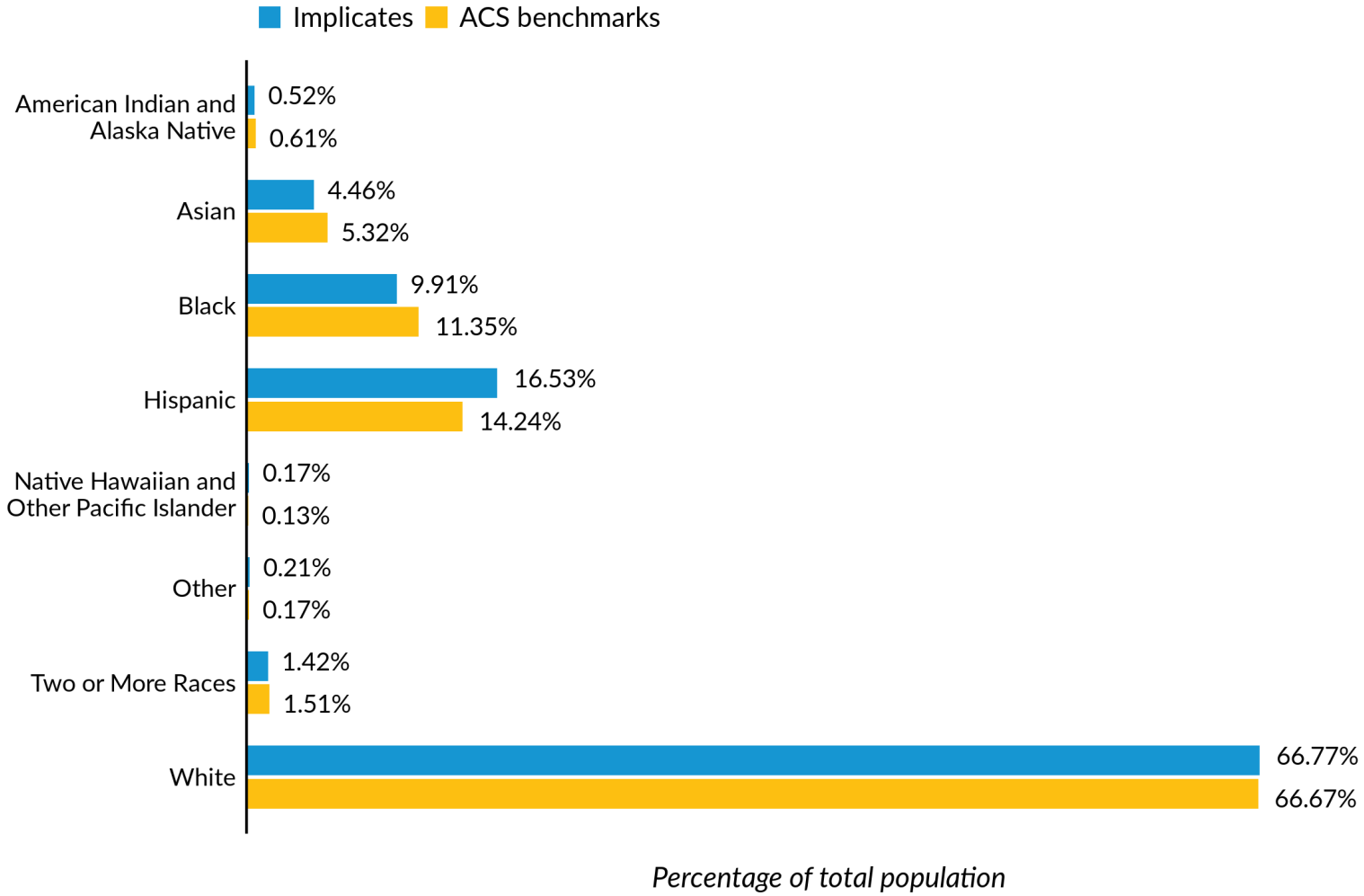
Asking whether imputation is the right approach

- Do the potential benefits of imputation outweigh the risks?
  - **Opportunity cost**: Would resources for imputation be better used to improve data collection? What is the next-best available data?
  - **Fitness for purpose**: How will the imputed data be used? Does the available data support the use case?
  - **Outcomes**: How do the applications advance racial equity? (e.g. identifying disparities, case-making and corrective action, targeting resources)

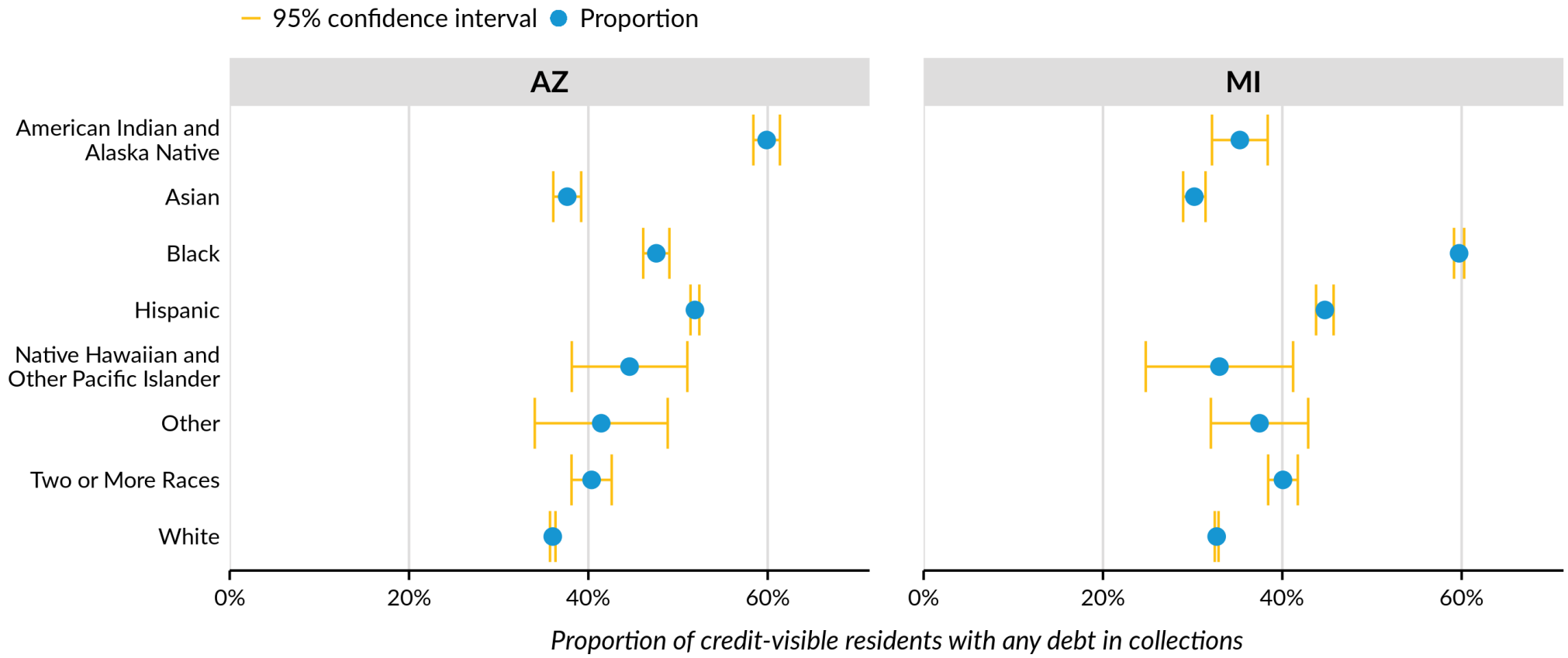
Case Study Process Checkpoints

- *Checkpoint 1:* Before imputation, audit input data for bias (e.g. “garbage in, garbage out.”)
- *Checkpoint 2*: During imputation, examine where bias could be introduced at each step. Communicate limitations and document uncertainty.
- *Checkpoint 3*: After imputation, assess whether imputed race/ethnicity data are accurate enough to be used ethnically for your analytic purpose.

Checkpoint 3: benchmark against trusted aggregate statistics



Checkpoint 3: Examine Fitness for Purpose



Standards and Recommendations Guide

- *Whether* imputation is the right approach for disaggregating data for a given use case
- *Who* should be involved in the process for review and accountability – with a particular emphasis on community partners
- *How* to develop community-led standards for data sharing that protect privacy and harm from use by bad actors

Standards and Recommendations Guide

- *Whether* imputation is the right approach for disaggregating data for a given use case
- *Who* should be involved in the process for review and accountability – with a particular emphasis on community partners
- *How* to develop community-led standards for data sharing that protect privacy and harm from use by bad actors
- Standards revolve around ***relevance***, ***accuracy***, ***privacy***, ***interpretability***, and ***institutional review/quality control***, and ***community engagement***

Conclusion

Concluding Thoughts and Key Takeaways

- Equity must be considered in every decision
- Examine differential outcomes by race and ethnicity
- Examine fitness for purpose in context of specific analytic case
- Clearly communicate limitations (including accurately estimating margins of error of statistics)
- Do the best you can with what you have access to, while constantly pursuing data improvements and new collection (when it makes sense)