Statistical Assessment of Record Linkage in the Estimation of the Citizen Voting-Age Population

J. David Brown

U.S. Census Bureau

August 9, 2023

Uses of Record Linkage in Census Bureau Data

- Characteristic imputation and item replacement in surveys
- Administrative record-based statistics
- Evaluation of survey and administrative record data
- Academic research using data linked across sources and time

United States® Census Bureau

Effects of Record Linkage Error

- Characteristic imputations for wrong person
- Undercoverage when some records cannot be assigned unique identifiers
- Overcoverage when same person is assigned multiple identifiers
- Discrepancies across sources could be due to linkage error rather than error in the survey or administrative records being evaluated
- Inferences in academic research affected by record linkage error and biased samples using only linked records

United States® Census Bureau

Person Identification Validation System (PVS)

- Assigns unique person identifier, called Protected Identification Key (PIK)
- Reference files
  - Social Security Administration NUMIDENT variables
    - SSN, date of birth (DOB) variants, name variants, gender
  - Addresses from government administrative records

United States® Census Bureau

PVS Search Modules

- Verification Module
  - Exact match on SSN
  - Checks for sufficient agreement by name and DOB
- GeoSearch Module
  - Blocks at housing unit level, then broadens geography up to first 3 digits of ZIP Code
  - Match variables include name variables, DOB, gender, and address fields
- NameSearch Module
  - Blocks on exact DOB and parts of names, then blocks on parts of name and DOB
  - Match variables include name elements, DOB, and gender

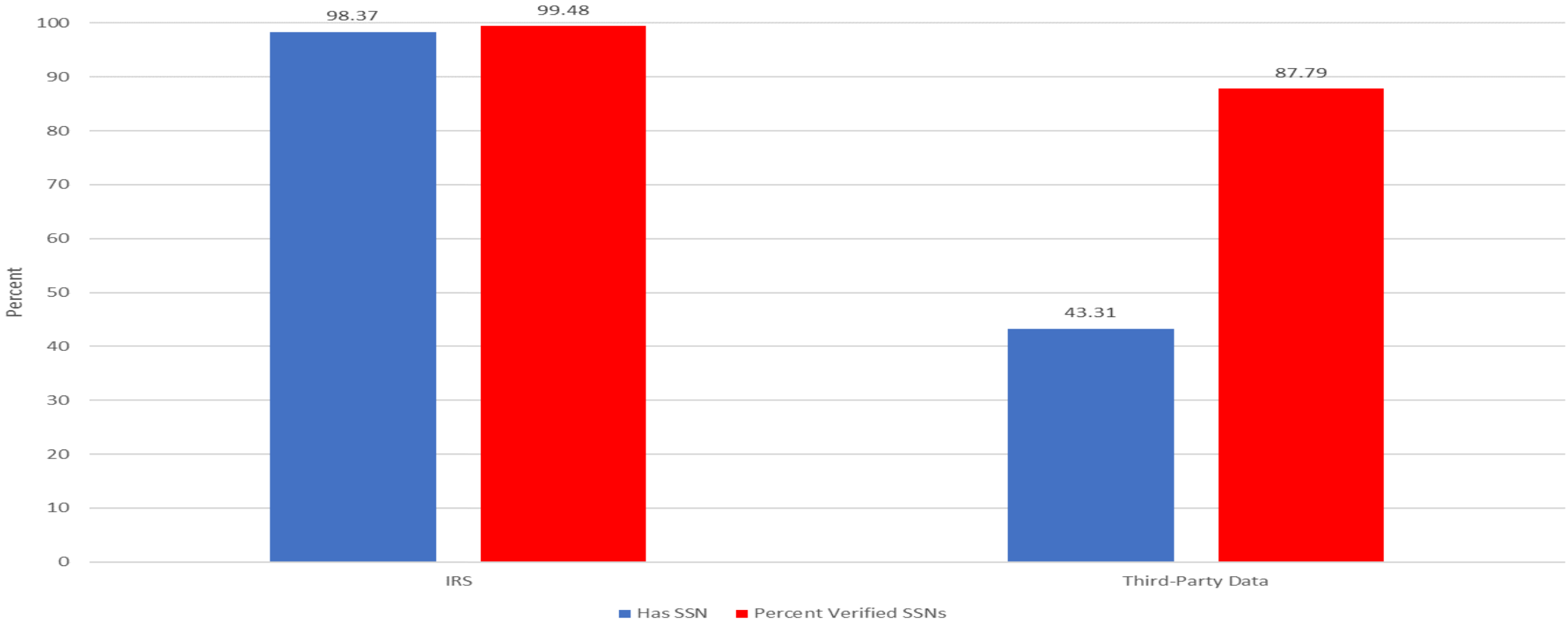
United States® Census Bureau

Main Questions

- **Why can some records be linked and others not?**
- How does linkage error vary by linkage variables used?
- How does linkage reliability vary by demographic characteristics?
- How do linkage rates vary by demographic, housing, and neighborhood characteristics?

United States® Census Bureau

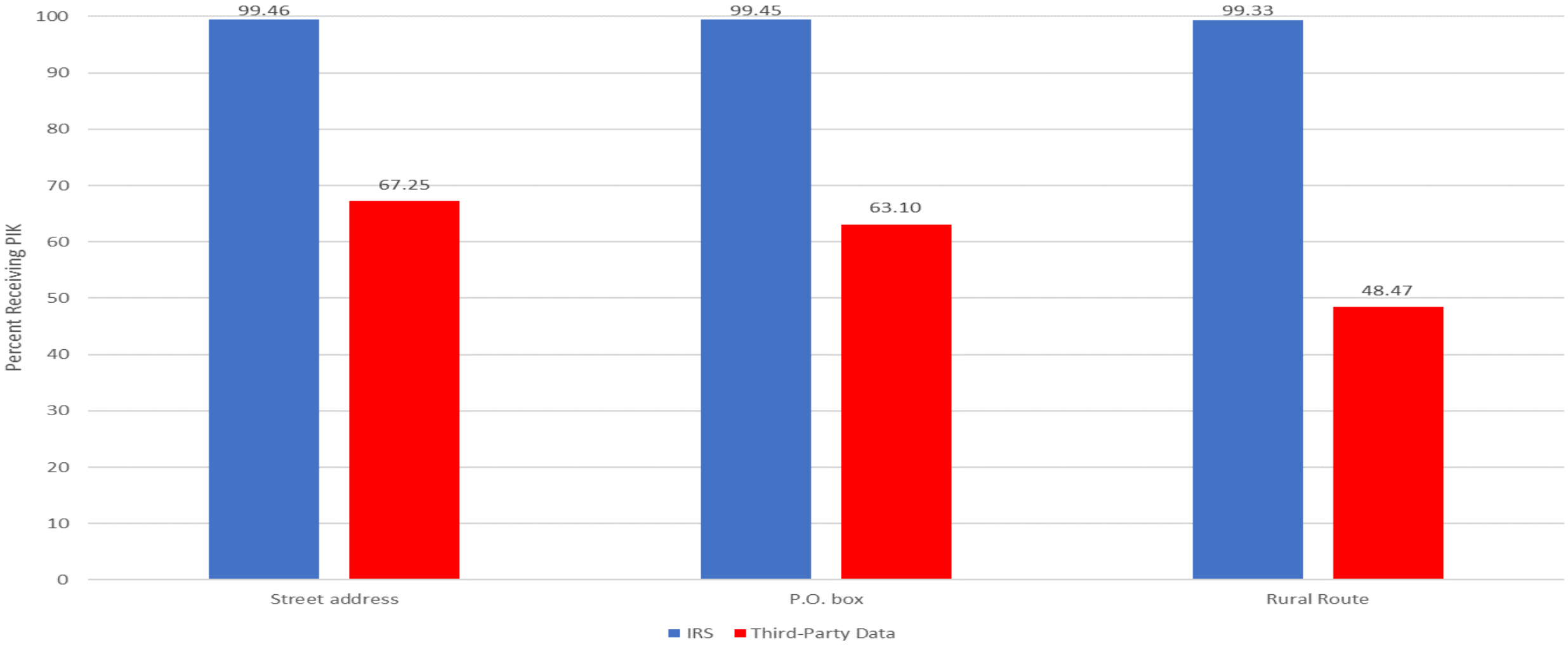
Percent with Social Security number (SSN) and of which, Verified SSN



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

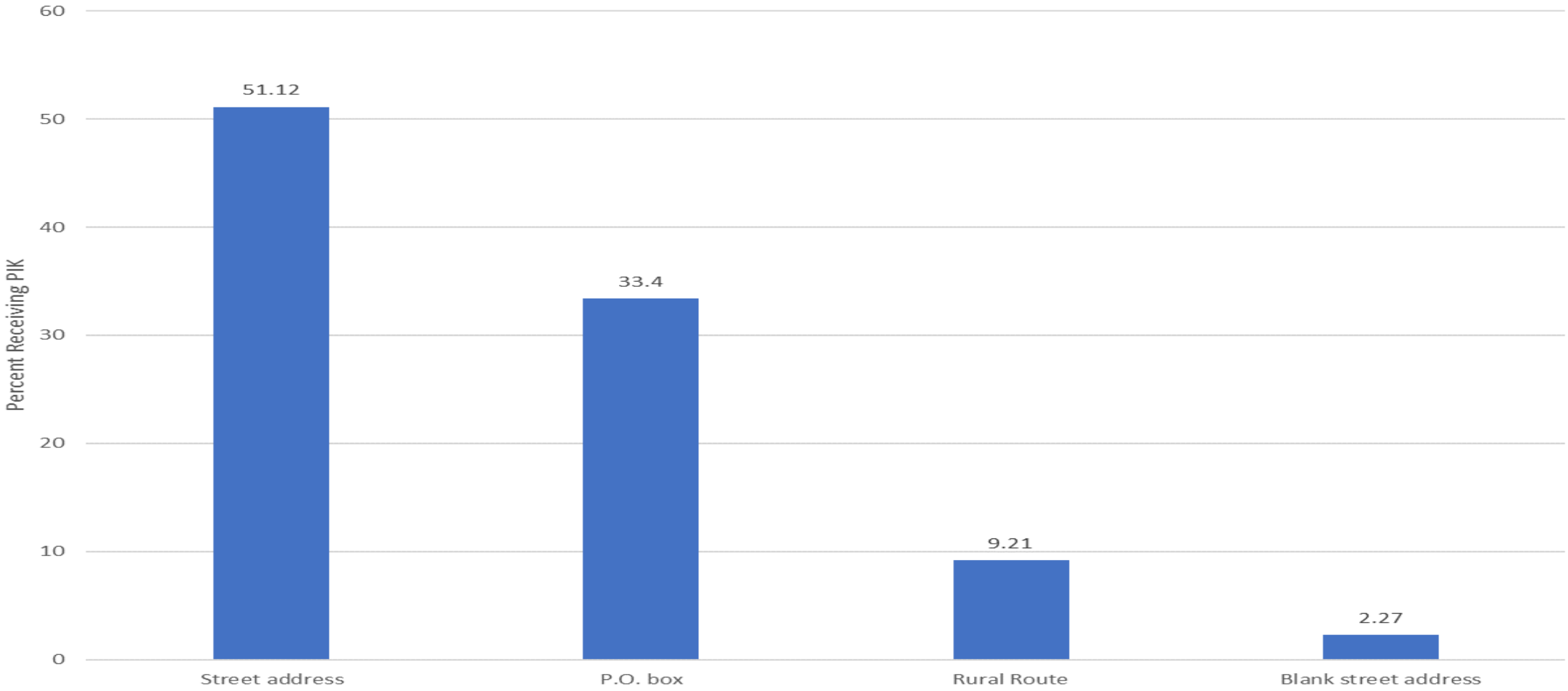
Percent Receiving Unique Person Identifier (PIK)



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

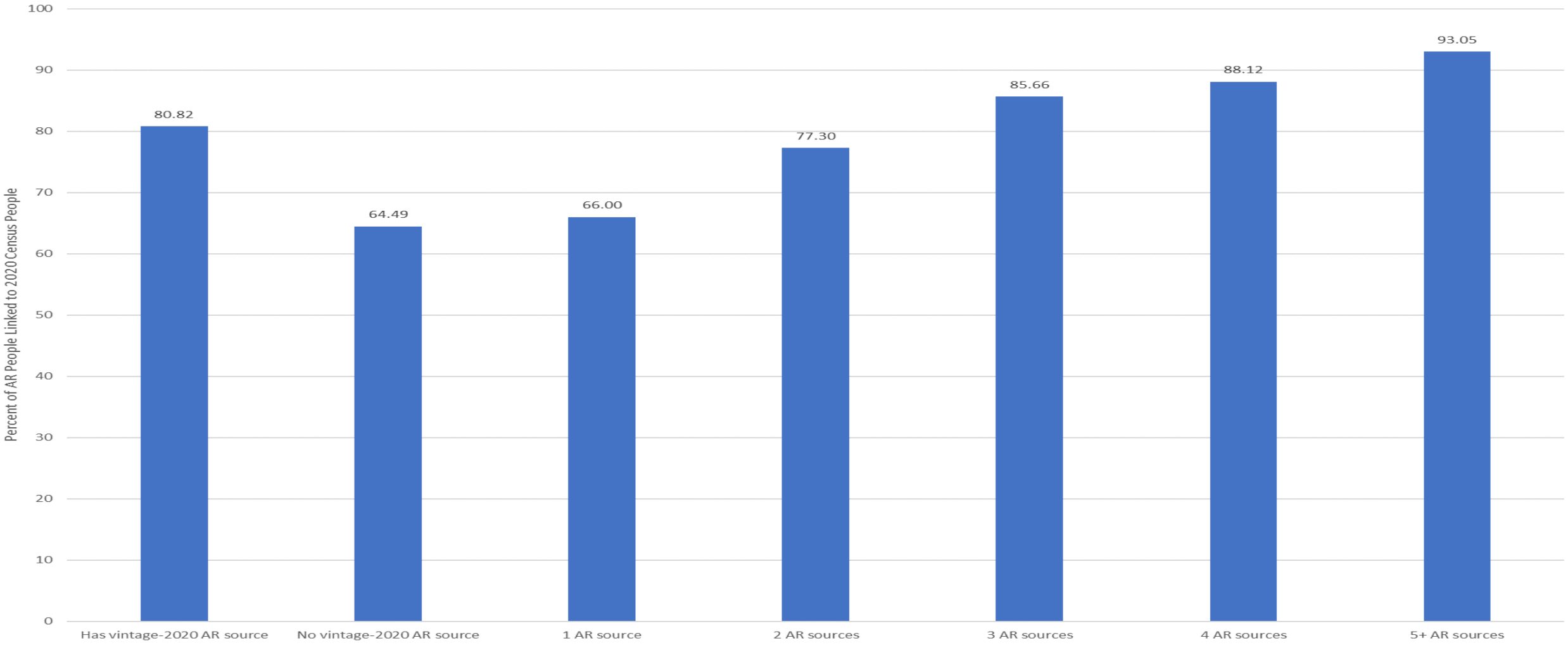
Percent Receiving PIK in Third-Party Data Lacking SSNs



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

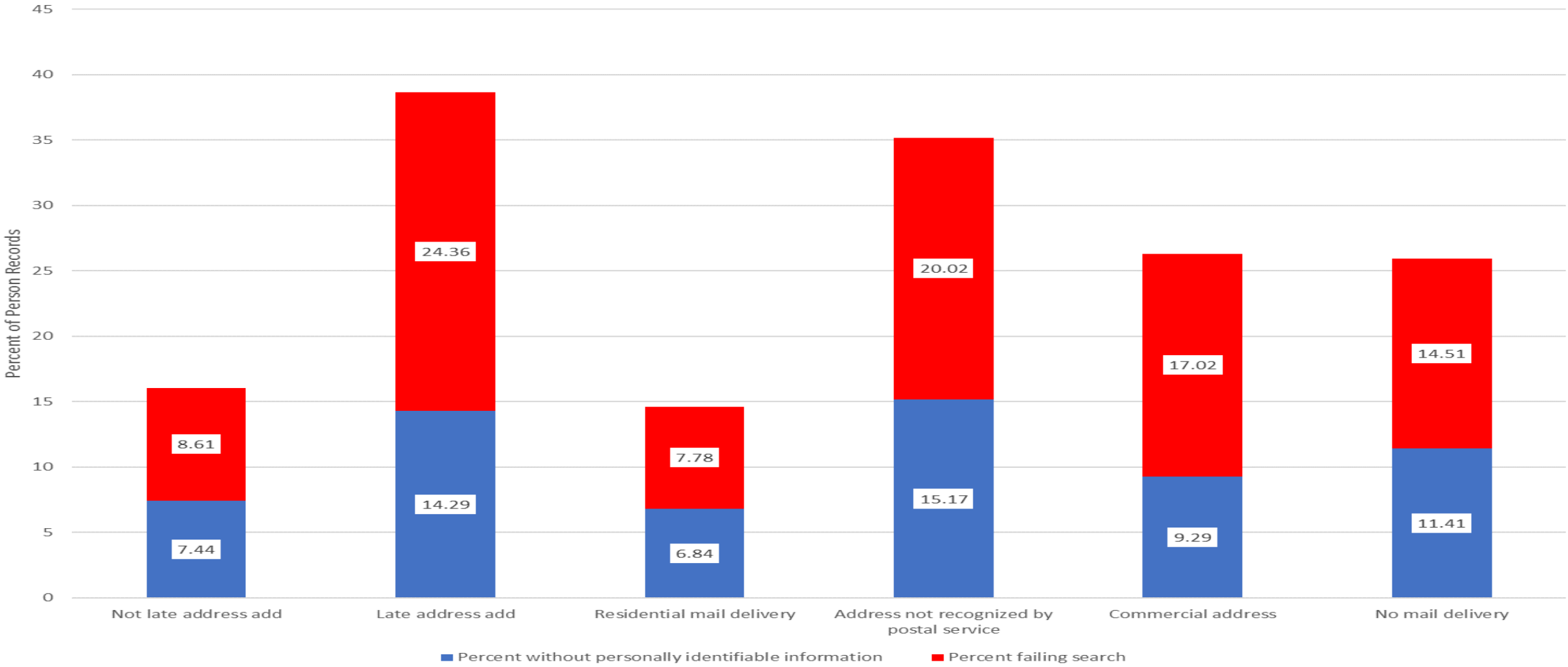
Percent of AR People Linked to 2020 Census People



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

Percent of 2020 Census Person Records without PIKs



United States® Census Bureau

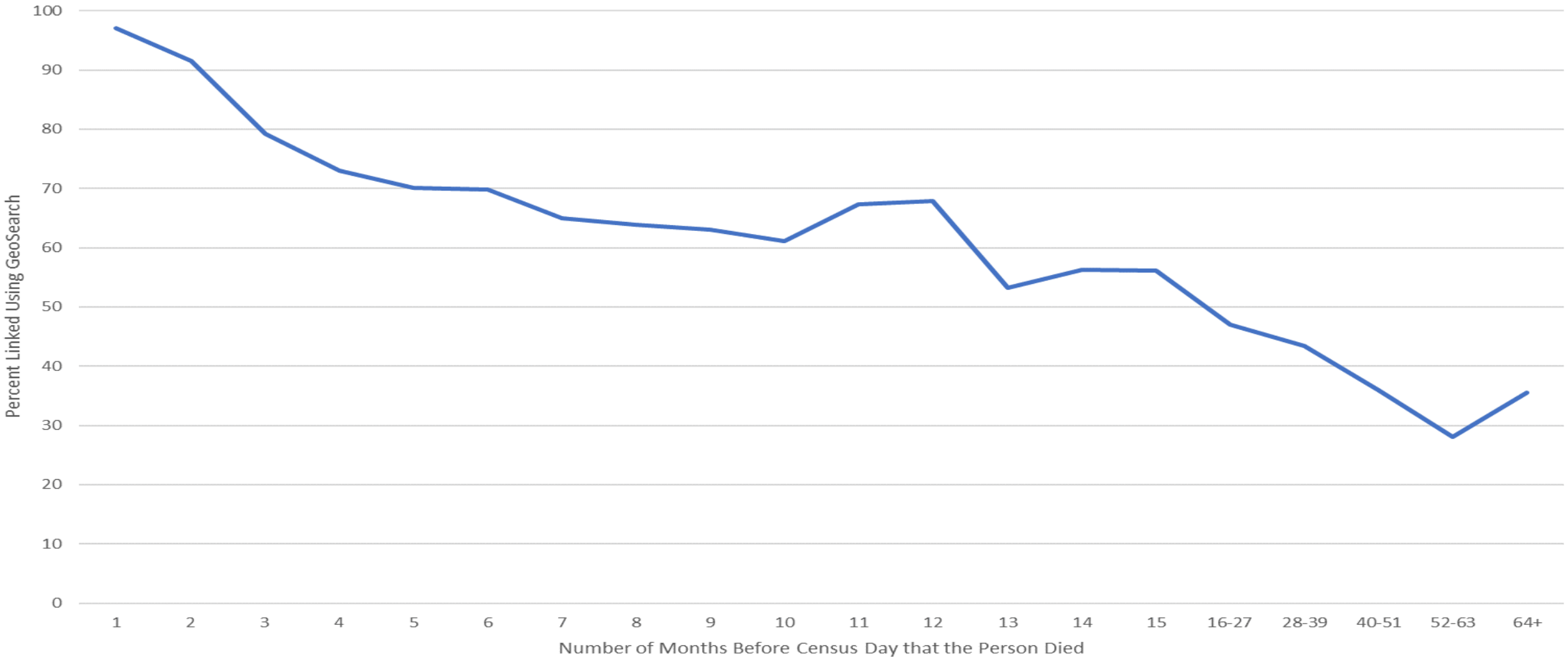
The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

Main Questions

- Why can some records be linked and others not?
- **How does linkage error vary by linkage variables used?**
- How does linkage reliability vary by demographic and housing characteristics?
- How do linkage rates vary by demographic, housing, and neighborhood characteristics?

United States® Census Bureau

Among 2020 Census People Recorded as Deceased in Linked AR, Association Between Percent Linked using Location and Death Timing Relative to Census Day



United States® Census Bureau

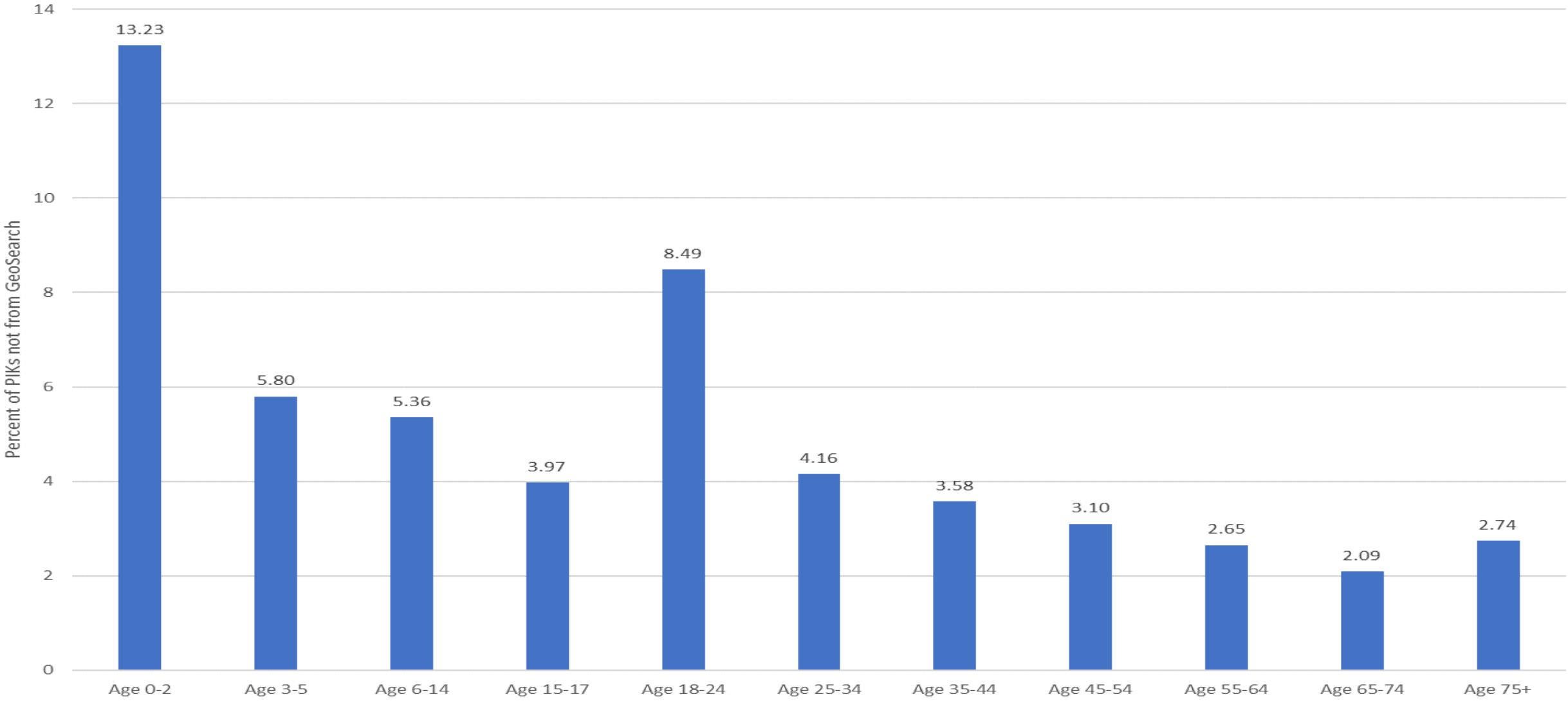
The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

Main Questions

- Why can some records be linked and others not?
- How does linkage error vary by linkage variables used?
- **How does linkage reliability vary by demographic characteristics?**
- How do linkage rates vary by demographic, housing, and neighborhood characteristics?

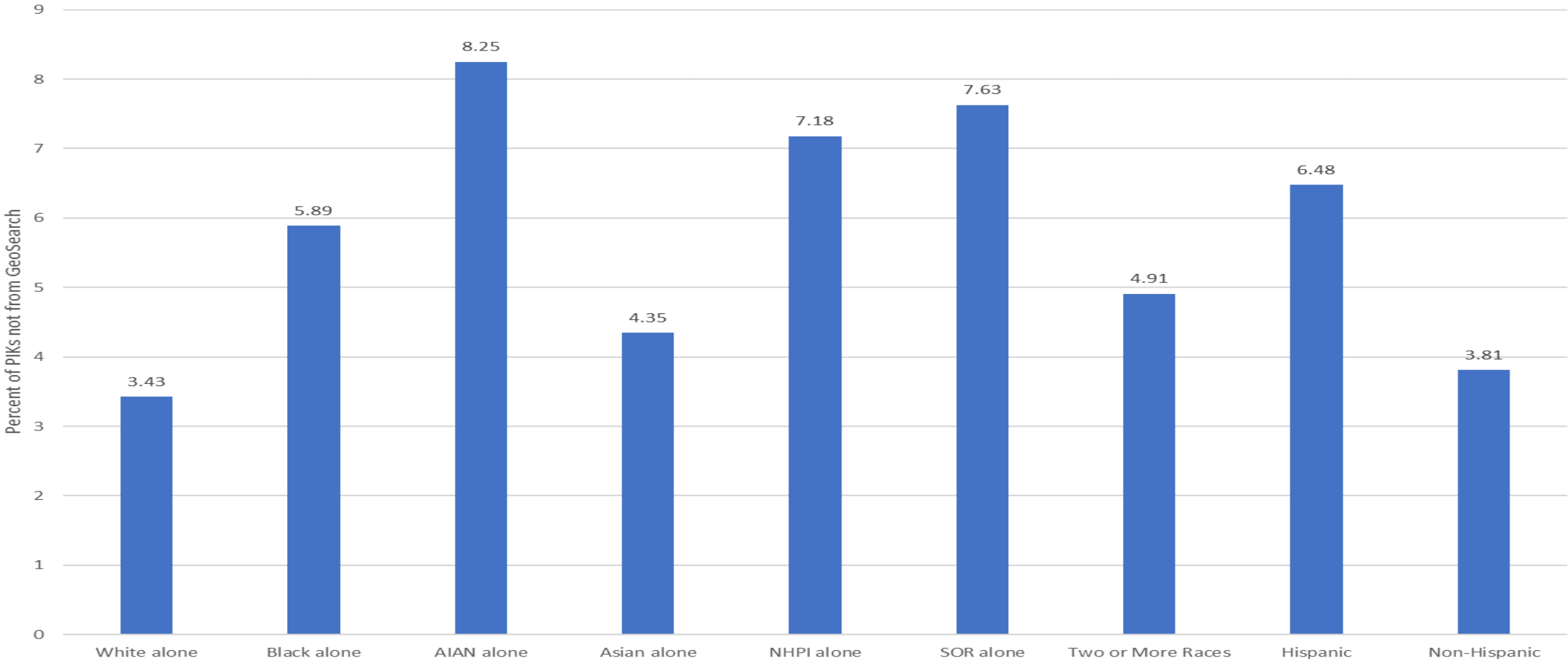
United States® Census Bureau

Among 2020 Census Person Records with PIKs, Percent not using Location in Linkage



The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

Among 2020 Census Person Records with PIKs, Percent not using Location in Linkage



United States® Census Bureau

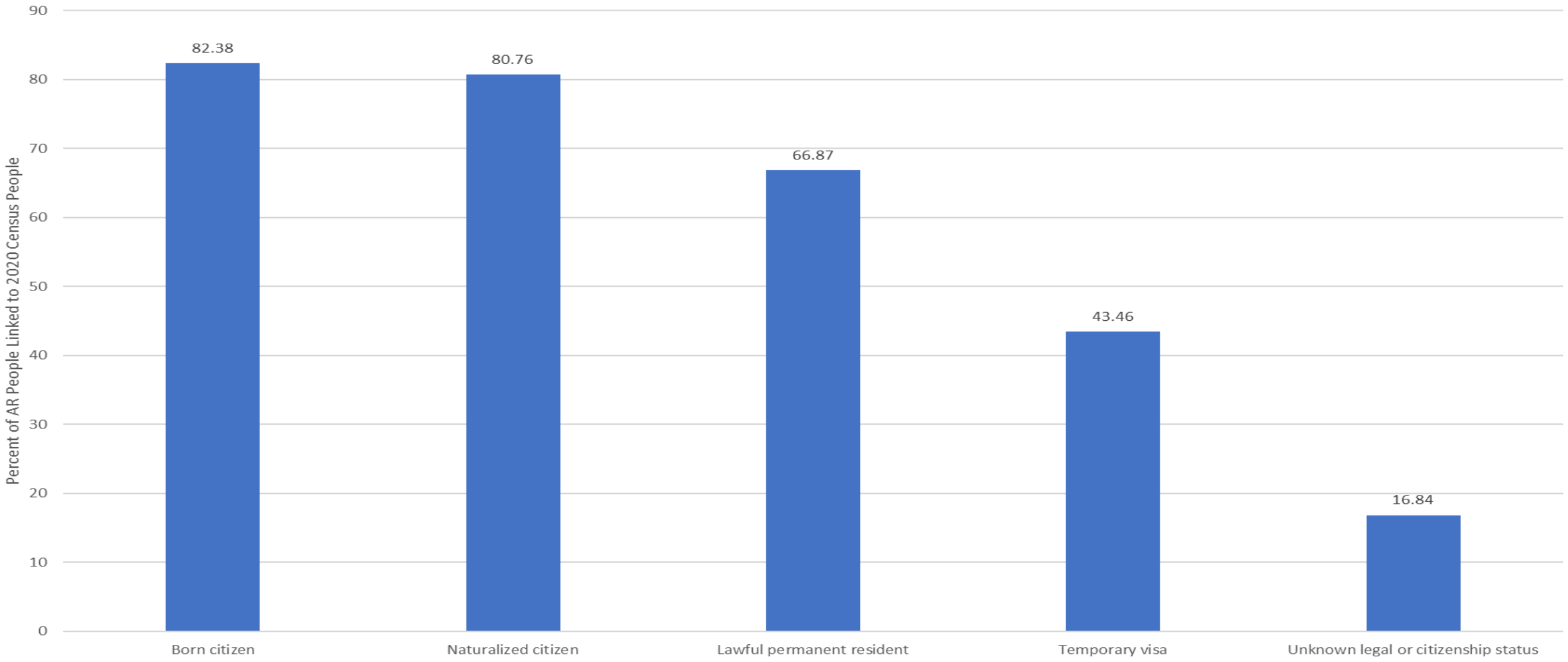
The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

Main Questions

- Why can some records be linked and others not?
- How does linkage error vary by linkage variables used?
- How does linkage reliability vary by demographic characteristics?
- **How do linkage rates vary by demographic, housing, and neighborhood characteristics?**

United States® Census Bureau

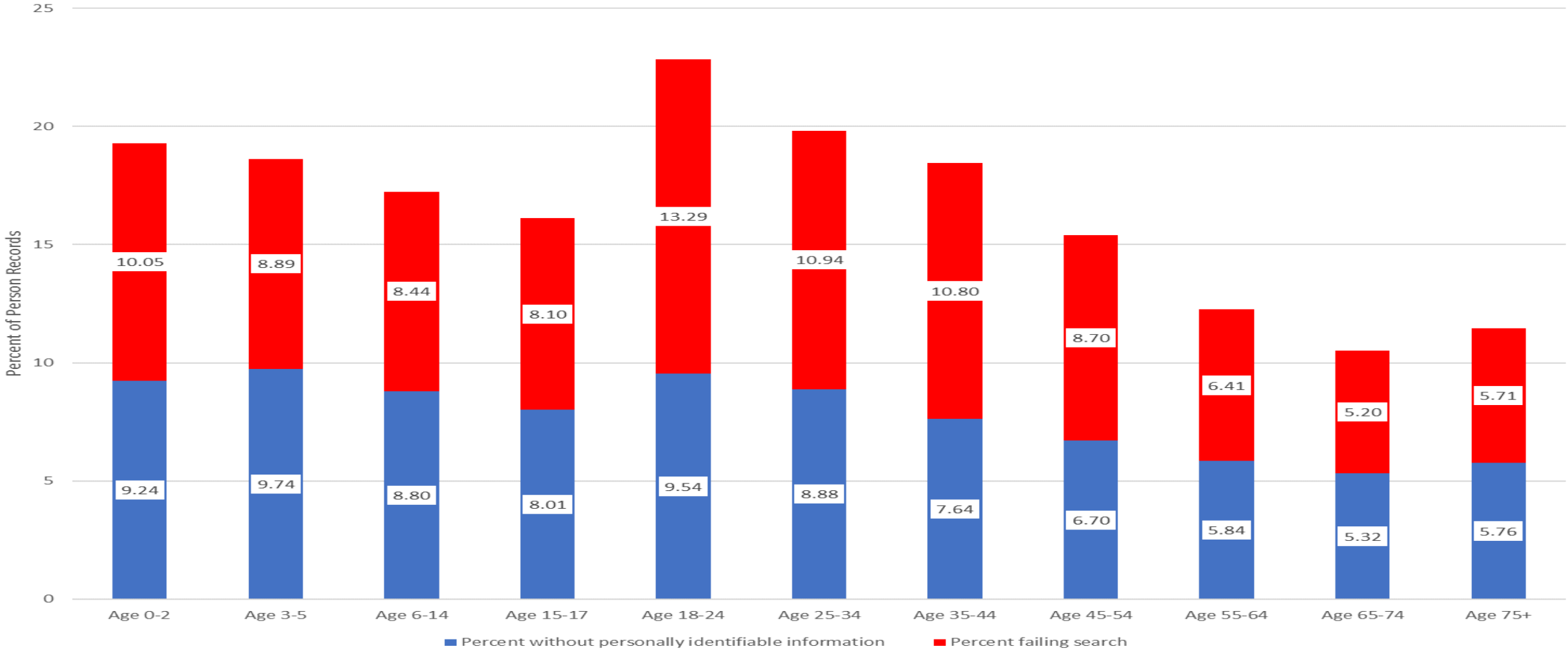
Percent of AR People Linked to 2020 Census People



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

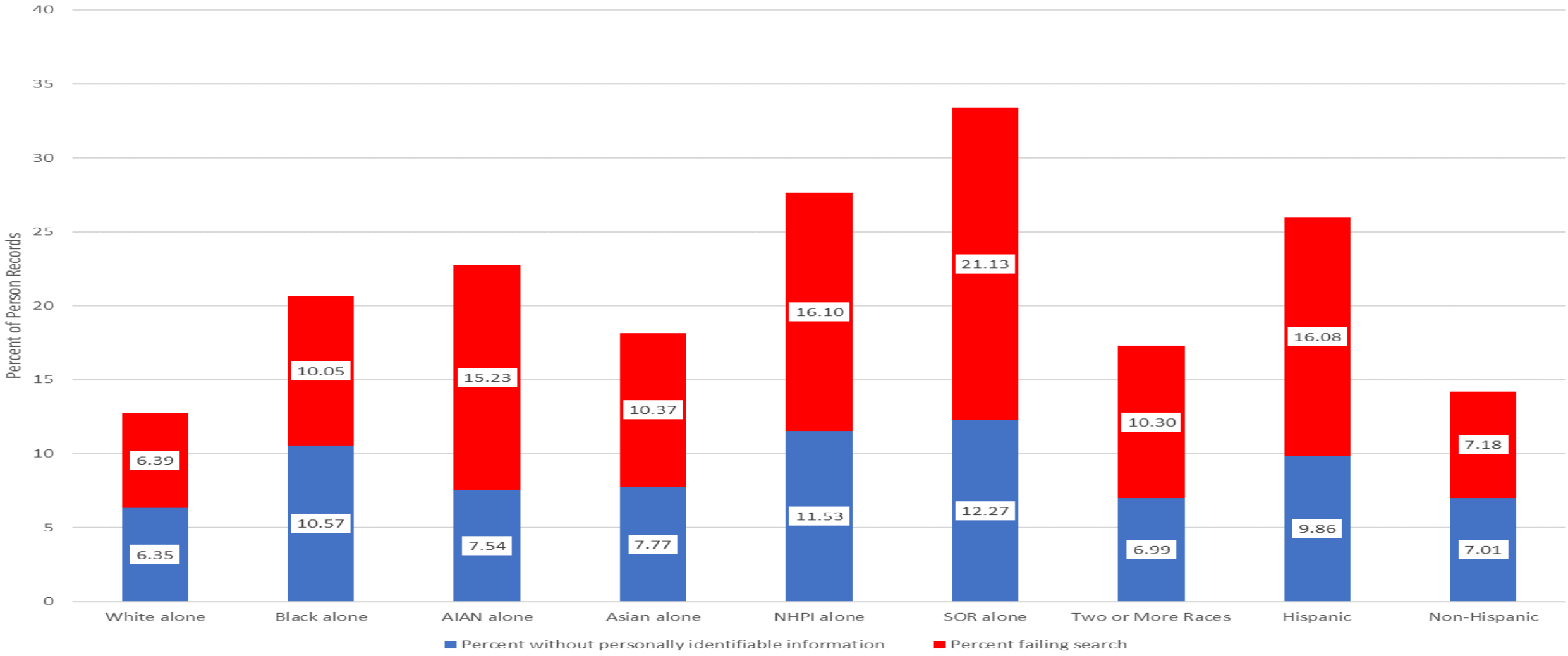
Percent of 2020 Census People without PIKs



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

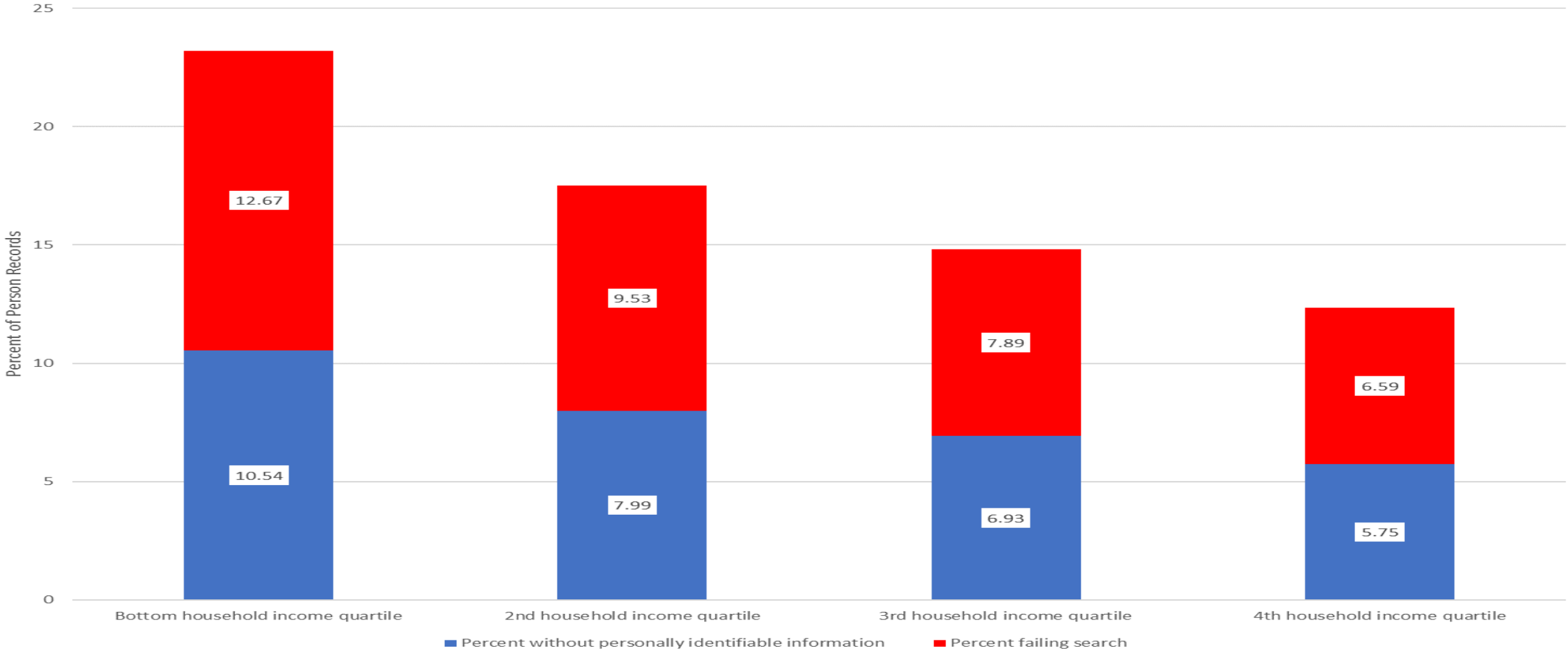
Percent of 2020 Census People without PIKs



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

Percent of 2020 Census People without PIKs



United States® Census Bureau

The data presented in this figure are approved for dissemination by the DRB (CBDRB-FY23-0253).

Conclusions

- Address is key for reliable record linkage when SSN not available
  - Records with city-style residential addresses with mail delivery have best linkage
  - Recent vintage
  - Error rates are lower when these characteristics are present
- Linkage and its reliability vary with demographic, housing, and local area characteristics
  - Whites, Non-Hispanics, older people have higher linkage rates and reliability
  - Citizens have higher linkage rates
  - Linkage positively associated with neighborhood income
  - These characteristics are associated with lower mobility

United States® Census Bureau

Additional Slides

United States® Census Bureau

Other PVS Search Modules

- DOBSearch Module
  - Blocks on first name in incoming file to last name in reference file, last name in incoming file to first name in reference file
- HHCompSearch Module
  - At least one person in the household must have received a Protected Identification Key (PIK)
  - Tries to link persons without PIKs in household to persons in the reference file at that household who haven't been linked to the incoming file
  - Blocked by Master Address File Identification Key (MAFID), name, DOB, and gender

United States® Census Bureau

PVS Process

- Matching based on Fellegi-Sunter (1969) probabilistic record linkage method
- Records receive a PIK in a module and pass if the PVS score (weighted average of closeness of matching variables) is above a threshold
- Records not receiving a PIK in a module and pass are sent through the next module/pass combination for which they are eligible
- For records with a PIK, the PVS crosswalk contains the module and pass in which a record received a PIK
  - Except in SSN verification, the crosswalk contains the PVS score from that module and pass
- For records not receiving a PIK, the PVS crosswalk distinguishes between those that were not sent to search, failed search, or had multiple matches

United States® Census Bureau