Evaluating the Quality of a National Voter Registration File Using Administrative Records

By: Andres Felipe Mira, PhD, Research Economist, Center for Economic Studies, U.S. Census Bureau

AAPOR 81st Annual Conference
Los Angeles, CA
May 14, 2026

Motivation

- Commercial voter registration records are widely used for:
  - Survey sampling frames.
  - Research on turnout, participation, and representation.
  - Policy analysis.
  - A third-party source to include in administrative record composite files.

- Despite widespread use, **systemic evidence on voter registration quality remains outdated and insufficient** (Kim et al., 2025).

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Why Data Quality Matters

- Data quality issues can introduce several types of errors:
  - Coverage error: Inclusion of “deadwood” records (e.g., deceased or outdated records).
  - Measurement error: Misclassification of demographic characteristics.

- These errors can lead to:
  - Bias in inference: Distorted turnout estimates and biased subgroup analysis.
  - Administrative inefficiencies: Increased survey costs.

- Accurate measurement of these errors provides information on the potential magnitude of the potential bias and inefficiencies being incurred.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Contribution

- Provide first comprehensive nation-wide data quality analysis of 2020 commercial voter registration files.

- Use the Census Bureau's Person Identification Validation System (PVS) to assign registrations a unique Protected Identification Keys (PIKs).
  - PIKs enable record-linkage to federal and survey datasets available at the U.S. Census.

- Using PIKs, I produce high-quality measurements of:
  - Duplicate registrations (within and across state files).
  - Deceased records.
  - Demographic classification accuracy (race and Hispanic ethnicity).

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Data sources

- National Voter Registration File (NVRF):
  - A 2020 snapshot of all 50 state + DC commercial voter registration files.

- Administrative Records:
  - PVS to assign PIKs.
  - 2025 Census Numident provides mortality information.
  - 2020 Best Race and Ethnicity Administrative Records Composite (Best Race) file provides validated demographic characteristics from administrative and survey responses.

Together, these data enable person-level validation of almost all voter file records.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

The Commercial Voter Registration Data

United States® Census Bureau

The Data: From Registration to Commercial Vendor to Census

Decentralized:

- There is no single national voter registration file (NVRF).
- Voter rolls and access are managed differently in each state.
  - Different laws and policies regarding data disclosure, list maintenance regulations, etc.
- Potential for significant variation in quality of data.

Acquisition and processing by commercial data vendors:

- Once state files are acquired the data is standardized, cleaned, and processed.

Important to emphasize:

- Measuring the quality of the commercial voter registration files.
- NOT the quality of the voter rolls in local election offices.
- Files do not contain any political party status information.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Assigning Unique Person Identifiers

United States® Census Bureau

What is a PIK?

- A Protected Identification Key (PIK) is: A unique anonymized person identifier.

- Assigned using probabilistic matching on (Wagner and Layne, 2014):
  - Demographics: Complete name, date-of-birth, and sex history and variations.
  - Geographic: Address information from administrative records.

- This enables:
  - Linking individuals across administrative datasets.
  - Identifying multiple records belonging to the same person.

- Key Difference to standard methods used:
  - Identifies all reported name changes and name variations.
  - Address data provides additional strength to the accuracy of the match.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

PIK Assignment Rates

United States® Census Bureau

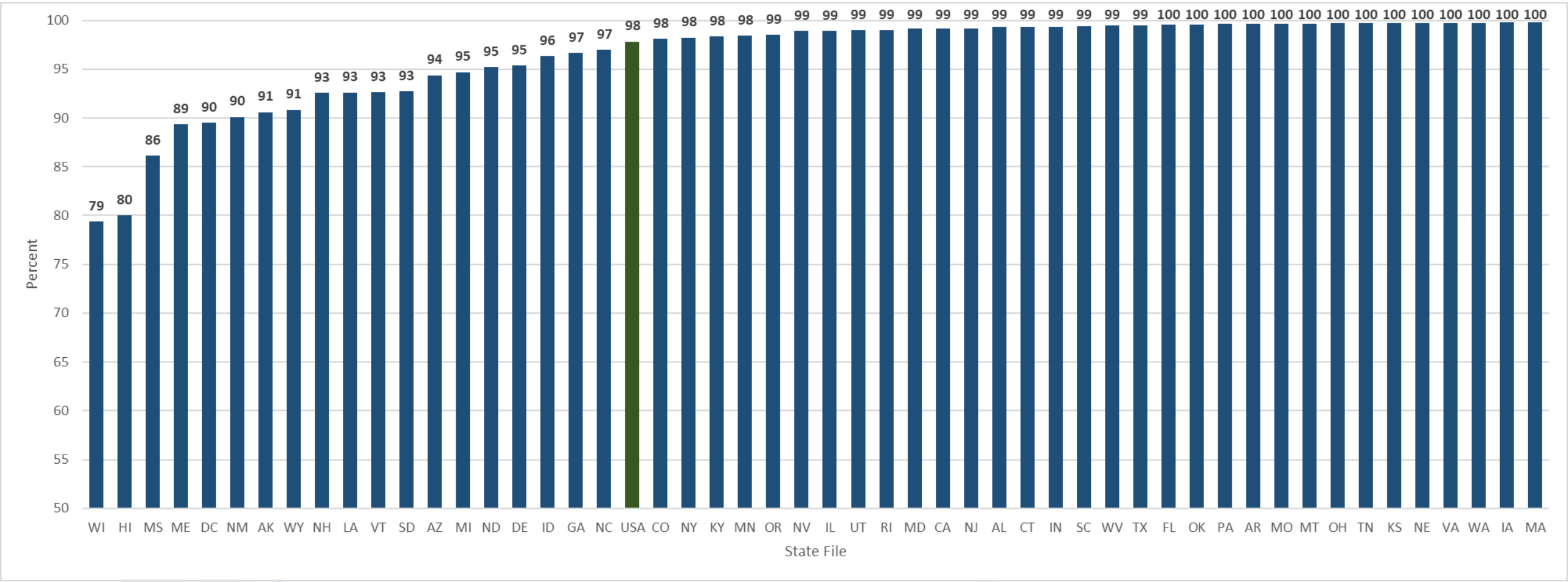
PIK Assignment Rate in the NVRF

	Number of registrations	Percent of registrations
Total Registrations	200,000,000	100.00
With PIKs	195,576,000	97.84
Without PIKs	4,324,400	2.16

- PIK assignment rate as high as federal administrative records (Layne, Wagner, and Rothhass, 2014).
- Indication that the quality of input data (name, date-of-birth, sex, address) is high.
- Masks considerable state-level heterogeneity.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Percent of Registrations Assigned a PIK, by State

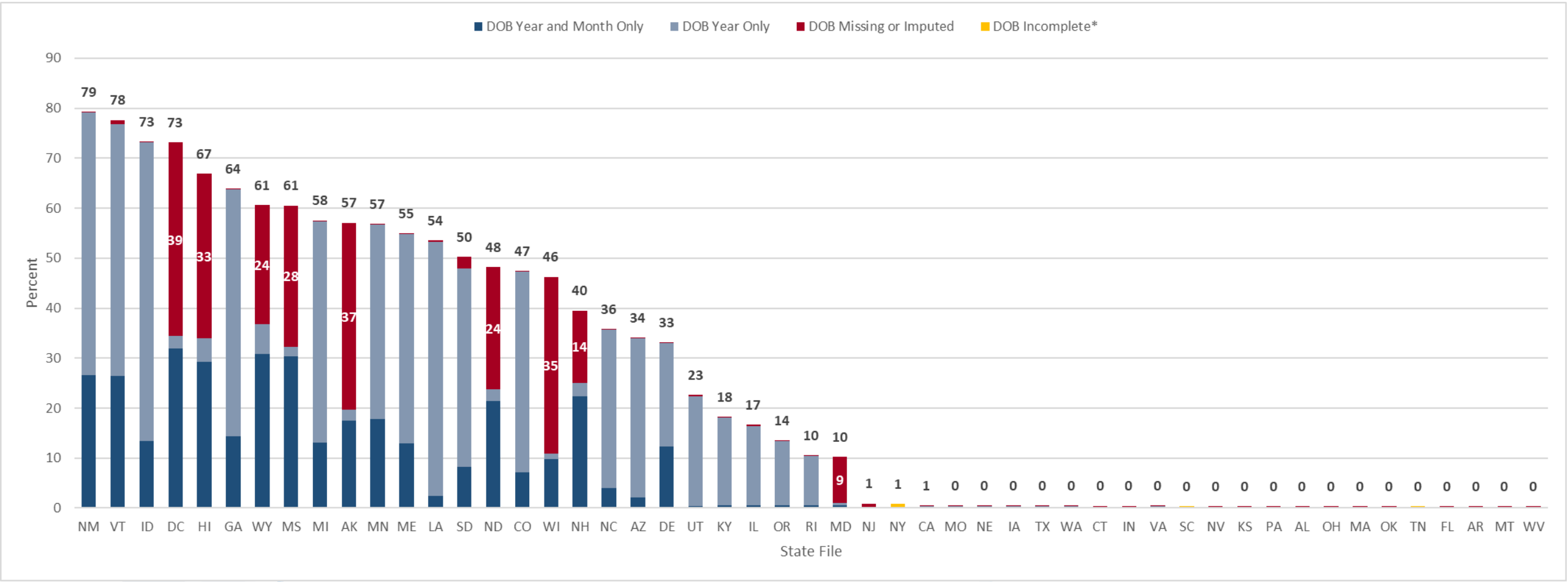


- High variation in PIK assignment (and quality) across states.
- Driven by data disclosure limitations.

United States® Census Bureau

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Percent of Registrations with Incomplete Date-of-Birth, by State



- States with high share of completely missing information more likely to have lower share of PIK assignment

United States® Census Bureau

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Measuring Duplicate Registration Records

United States® Census Bureau

Number of Persons in the NVRF

	Number of Persons	Percent of Persons
Total Persons	188,297,700	100.00
Unique PIKs	181,228,950	96.25
Unduplicated PIKs	7,069,050	3.75

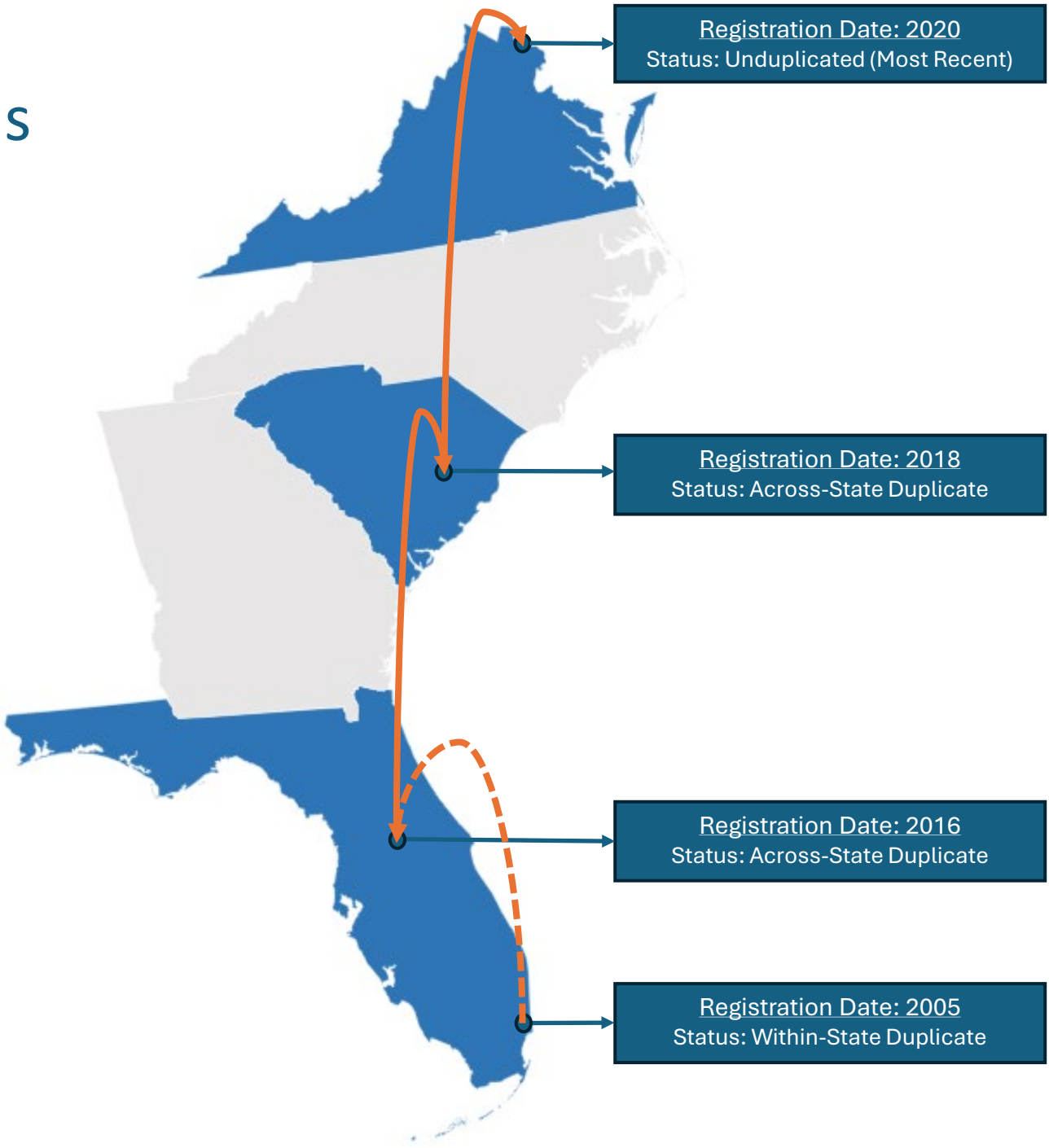
- Unique PIK: person was found in only one voter registration.
- Unduplicated PIK: person was found in more than one registration (within or across state files).

- Identify 188.3 million unique individuals in the NVRF.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Measuring Duplicate Registrations

- Duplicate registrations when a PIK appears more than once in the NVRF.
- Person may have multiple registrations within and across state files.
- Unduplicated registration: registration is the most recent registration.
- Across-state duplicate: Registrations are the most recent registrations found in another other states.
- Within-state duplicate: All other duplicate registrations.



United States® Census Bureau

Not real data, for illustrative purposes only.

Across-State Duplicate Registration Records

United States® Census Bureau

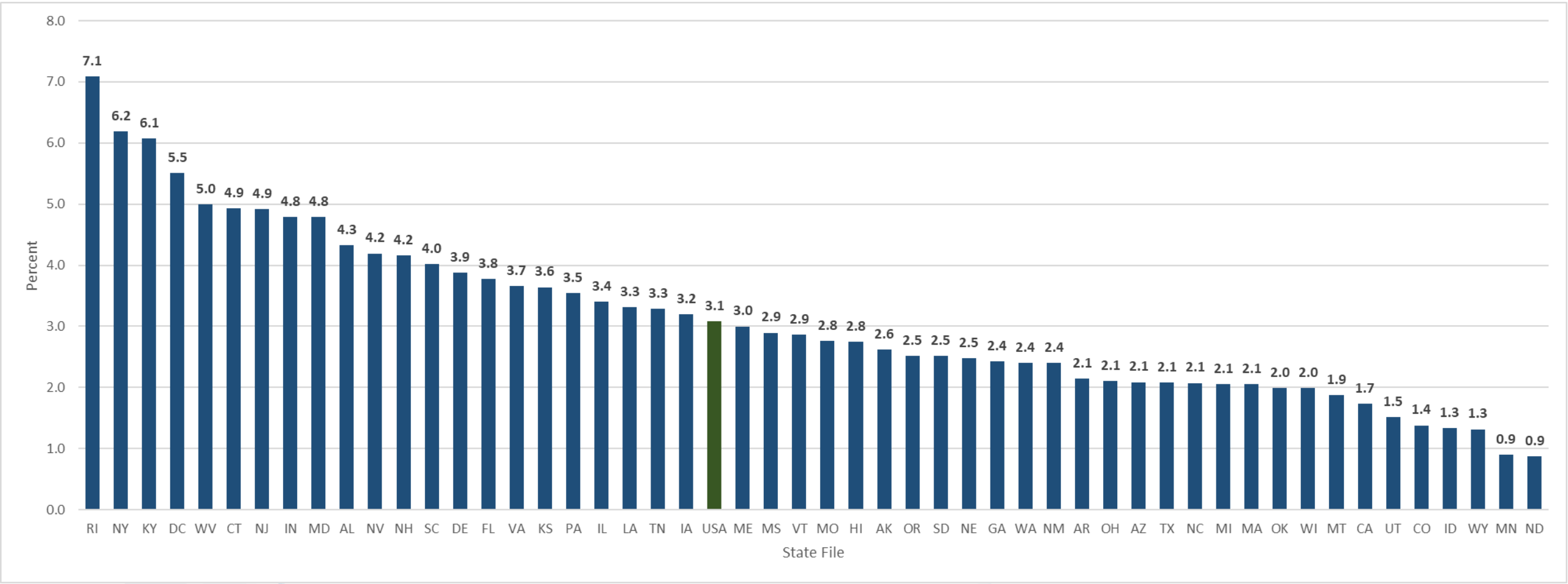
Distribution of States per PIK (Across State duplicates)

Number of States	Number of PIKs	Percent
1	182,300,000	96.84
2	5,744,000	3.05
3	190,000	0.10
4	11,000	0.01
5 or more	3,100	0.00

- Pew Center on the States, 2012: 2.7 million persons registered in more than one state in 2011.
- Dahl et al., 2023: Using Law of large numbers, predicts 3.1% of registered persons in 2020 have cross-state duplicates.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Percent of Across-State Duplicate Registration, by State



- Represents individuals who have moved but remain registered in their prior state(s).
- ~8X Difference in across-state duplicates between lowest and highest states.

United States® Census Bureau

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Within-State Duplicate Registrations

United States® Census Bureau

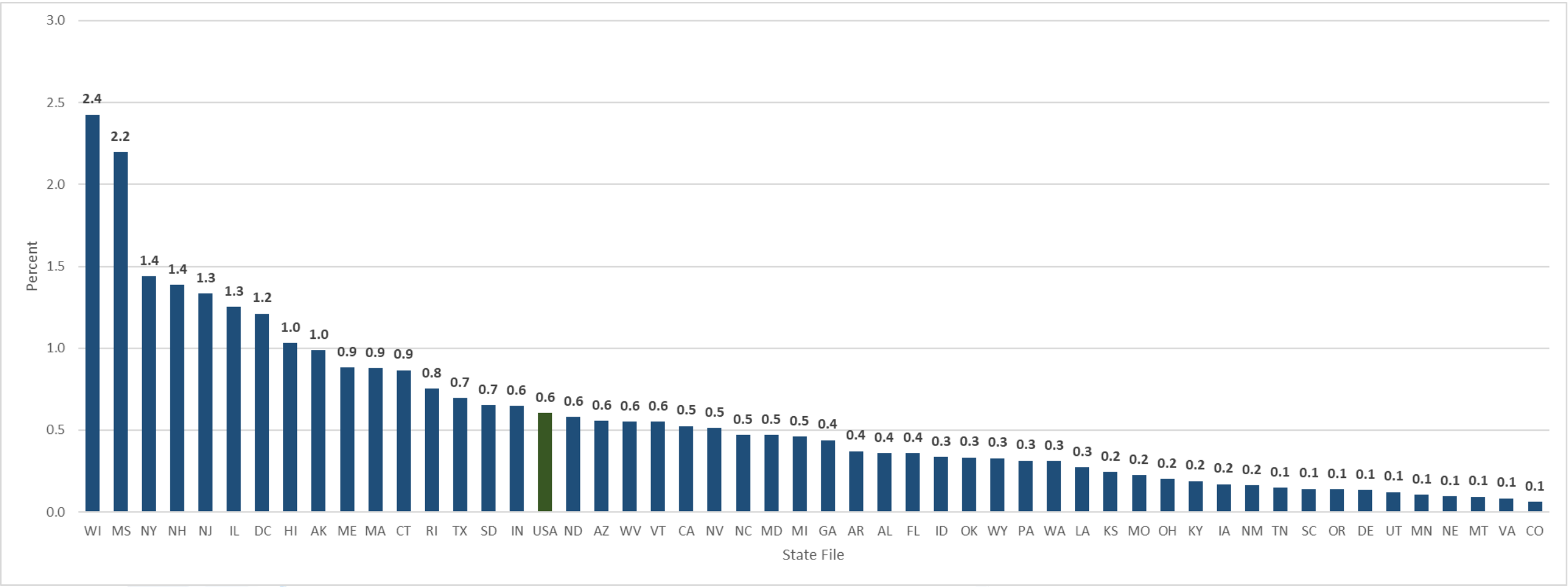
Distribution of Number of Registrations per PIK-State Pairs

Num. of Registrations within State	Num. of PIK-State Pairs	Percent
1	193,200,000	99.39
2	1,175,000	0.60
3	13,500	0.10
4	650	0.01
5 or more	300	0.00

- States are better at dealing with duplicates caused by within-state moves.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Percent of Within-State Duplicate Registration, by State



- Variation reflects differences in interstate mobility and list maintenance practices.
- Wisconsin and Mississippi state files are outliers.

United States® Census Bureau

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Deceased Registration Records

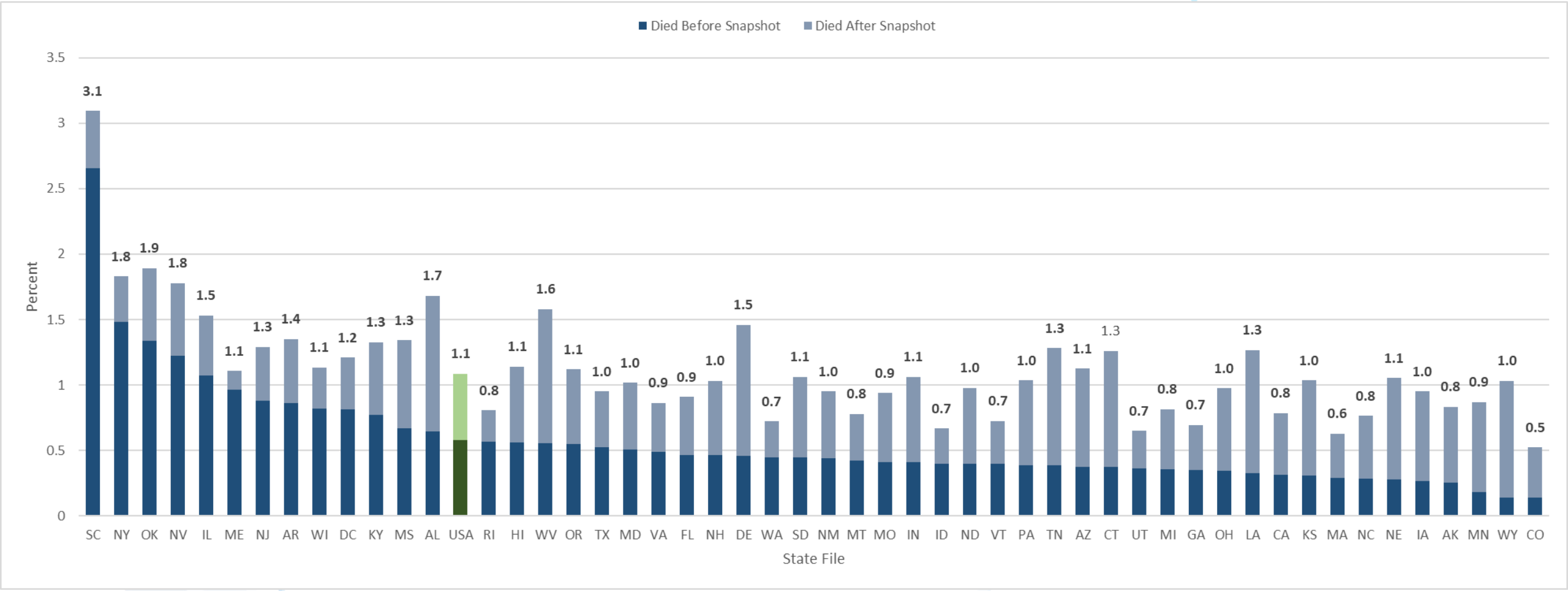
Deceased Registrations, by Period of Death

	Number	Percent of registrations
Total Registrations	200,000,000	100.00
Deceased as of 12/31/2020	2,176,000	1.09
Died before snapshot date	1,164,000	53.49
Died after snapshot date	1,012,000	46.51

- Snapshot date is based on the date the state file was pulled by the commercial vendor.
- Pew (2012) recorded 1.8 million records no linger living but still registered in 2011.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Percent of Deceased Registrations, by State



**South Carolina is a clear outlier in the share of records linked to deceased individuals.

Quality of Race and Ethnicity

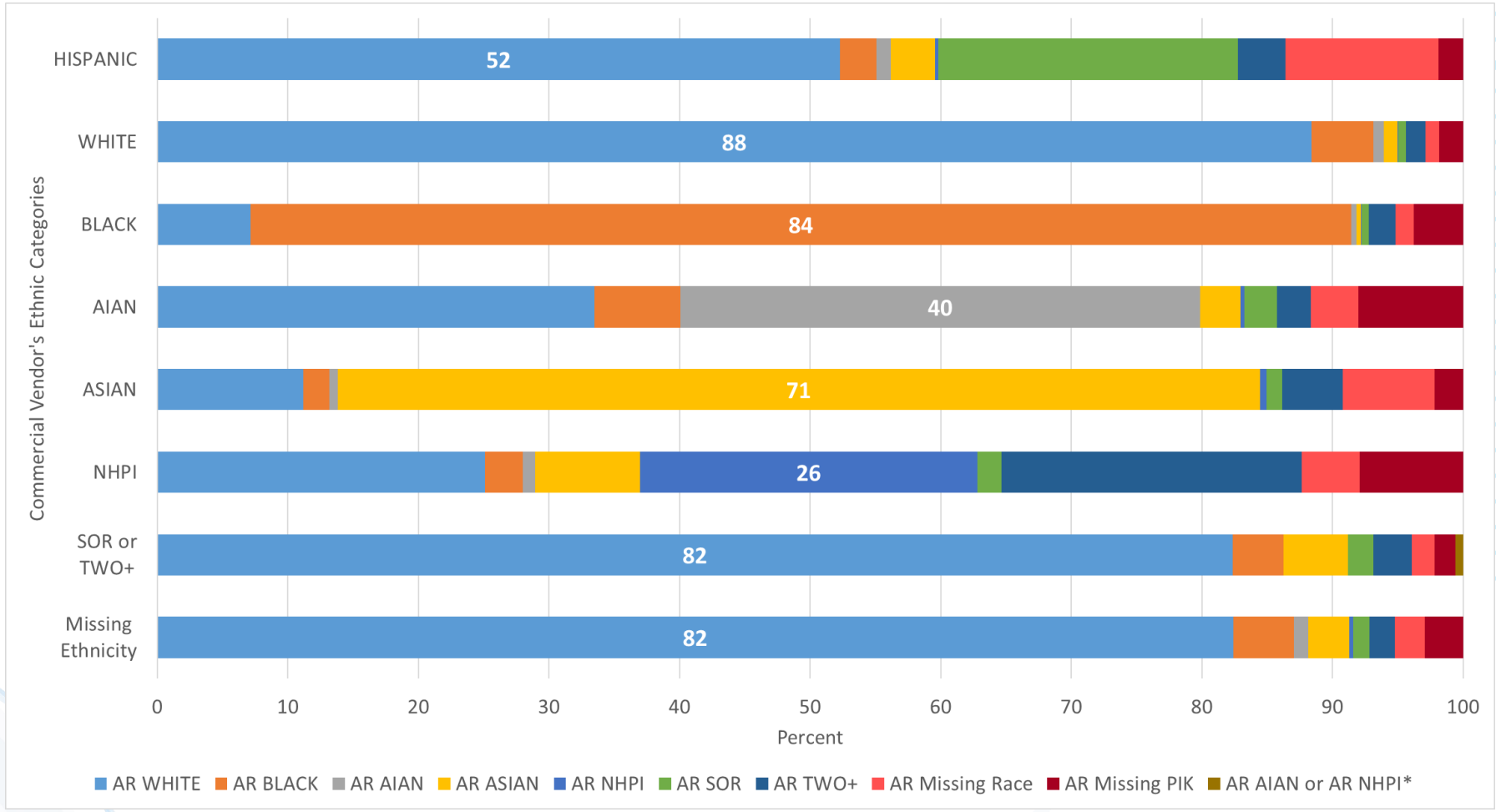
United States® Census Bureau

Quality of Vendor Ethnicity Variable in NVRF

- The vendor uses name and address information to assign one of a 100 different ethnic categories to each registration.
- To ease comparison, I crosswalk the 100 ethnic categories to match the Census racial and Hispanic ethnicity categories.
- Compare race and Hispanic ethnicity in the Census Best Race File separately.
- Critical variable used for;
  - Validating survey responses.
  - Producing subgroup analysis on turnout, participation, and representation.

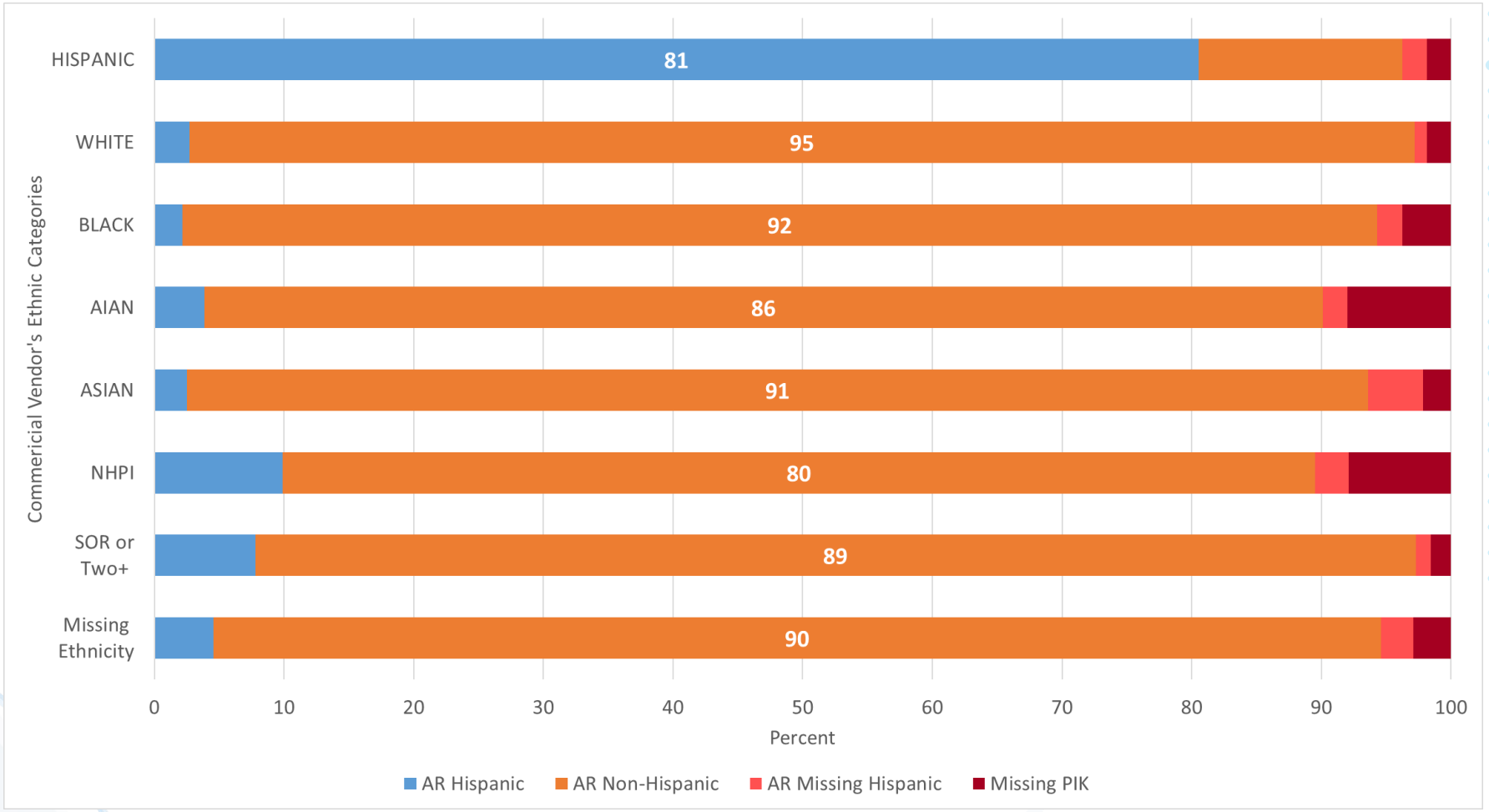
The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

“Best Race” breakdown by NVRF Ethnicity Categories



- Registrations with Vendor provided AIAN, ASIAN, NHPI, have low match rates with AR race response.
- AIAN and NHPI have lowest PIK assignment rates.

“Best Hispanic Ethnicity” breakdown by NVRF Ethnicity Categories



- Hispanic/Non-Hispanic match rates are broadly consistent across vendor provided ethnic groups.
- Yet, only 81% of Hispanics in registration file have a matching Hispanic ethnicity in the AR.

United States® Census Bureau

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Remarks and Implications for Research

This paper provides the first national evaluation of commercial voter file quality using linked administration records:

- Find considerable heterogeneity in duplicates and deceased records across state files.
- Vendor provided ethnic categories have considerable mismatch to AR-based race/ethnicity.

These findings have several implications:

- Duplicate and deceased registrations can reduce the efficiency of survey sampling frames.
- Misclassification of race and ethnicity can bias subgroup analysis.
- Gaps between record-based and survey-based estimates may be driven by ineligible records.
- Careful evaluation of voter registration data when used for research and surveys needed.

Limitations:

- High PIK assignment rate but still not complete (who are the un PIK'd registrations?)
- Focus on only 2020 snapshot. Next step is every year available.

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)

Thank you!

- Questions:
  - Email andres.f.mira@census.gov

The data presented in this slide are approved for dissemination by the DRB (CBDRB-FY26-CES014-010 and CBDRB-FY26-CES023-003)