# Subgroup Disparities in Automated Census Record Linkage

Hannah Postel (Duke Sanford)
Ran Abramitzky (Stanford and NBER)
Leah Boustan (Princeton and NBER)
Katherine Eriksson (UC-Davis and NBER)

NBER Linking Small Populations | 21 March 2025

# Motivation

Historical record linkage enables researchers to study important topics, e.g. intergenerational mobility and immigrant assimilation.

- No unique identifiers in historical records.

We often assess automated linking methods by analyzing match rates and accuracy of the full census population.

- These measures vary substantially by population subgroup, e.g. for immigrants and racial minorities.
  - Lower match rates *and* accuracy.

This project aims to understand why certain subgroups are more difficult to link than others and provide solutions.

# Match rates for population subgroups (1900-10)

| | Total Pop. | ABE-S | ABE-C | MLP | Full CT |
|---|---|---|---|---|---|
| **NB White** | 28,532,509 | 27.1% | 19.2% | 49.5% | 70.1% |
| **England** | 463,252 | 25.1% | 15.9% | 39.8% | 59.3% |
| **Germany** | 1,439,139 | 18.3% | 12.5% | 41.9% | 57.6% |
| **Russia** | 232,392 | 15.4% | 9.3% | 37.4% | 58.3% |
| **NB Black** | 4,321,182 | 14.9% | 8.3% | 24.9% | 45.2% |
| **Ireland** | 751,587 | 13.3% | 5.8% | 28.9% | 48.9% |
| **Sweden** | 324,411 | 13.2% | 7.5% | 34.7% | 58.9% |
| **Hispanic** | 71,515 | 7.9% | 4.0% | 14.8% | 28.9% |
| **China** | 96,319 | 7.0% | 3.2% | 3.5% | 22.3% |
| **Italy** | 316,824 | 6.2% | 4.4% | 17.2% | 37.8% |

# Why are subgroups more difficult to link? (Part 1)

Lower ground truth match rate?

- Immigrants may return to their country of origin; return migration rates high during this time period.

(Abramitzky, Boustan, & Eriksson 2019; Bandiera et al. 2013)

- Higher mortality?

Group-specific cohort effects?

- E.g. improved Black numeracy after Emancipation.

**We can't *fix* these, but we can *account* for them.**

# Do subgroups have a lower "ground truth" match rate?

Age group analysis to identify group-cohort effects.

- Calculate match rates by 10-year age band.
- Highlights both *intra-group* match differentials by age and *inter-group* age-match rate patterning.

"Fast forwarding" decennial census stocks.

- Exclude births, immigration.
- Remaining population captures deaths & emigration.

# Subgroup-cohort effects.

| Age band | NB white | NB Black | Chinese |
|---|---|---|---|
| **0-9** | 23.11% | 10.69% | 1.41% |
| **10-19** | 19.52% | 8.08% | 8.61% |
| **20-29** | 17.16% | 6.99% | 7.49% |
| **30-39** | 17.66% | 7.91% | 4.58% |
| **40-49** | 17.75% | 7.99% | 3.52% |
| **50-59** | 17.51% | 7.17% | 2.87% |
| **60-69** | 15.41% | 5.41% | 1.12% |
| **70+** | 7.34% | 2.64% | 0.33% |
| **Total** | 19.20% | 8.34% | 4.02% |

# Subgroup-cohort effects.

| Age band | NB White | Chinese | Ratio |
|---|---|---|---|
| **0-9** | 23.11% | 1.41% | 6.10% |
| **10-19** | 19.52% | 8.61% | **44.08%** |
| **20-29** | 17.16% | 7.49% | 43.66% |
| **30-39** | 17.66% | 4.58% | 25.92% |
| **40-49** | 17.75% | 3.52% | 19.82% |
| **50-59** | 17.51% | 2.87% | 16.40% |
| **60-69** | 15.41% | 1.12% | 7.30% |
| **70+** | 7.34% | 0.33% | 4.54% |
| **Total** | 19.20% | 4.02% | **20.93%** |

# Subgroup-cohort effects.

"Removing" group-cohort effects.

- I.e, fixing the ratio of the two groups' matched age structure at its maximum.

| | **Original MR** | **New MR** | **% increase** |
|---|---|---|---|
| **NB Black** | 8.3% | 8.9% | 6.7% |
| **Chinese** | 4.0% | 8.4% | 108.5% |
| **Italian** | 4.4% | 5.7% | 30.0% |
| **Swedish** | 7.6% | 12.3% | 63.0% |
| **Irish** | 5.8% | 13.3% | 129.3% |
| **English** | 16.1% | 21.6% | 34.2% |

# Modified population projections

| | % found 1900-10 | Scaled ABE-C |
|---|---|---|
| **NB White** | 89.7% | 21.3% |
| **NB Black** | 84.1% | 9.9% |
| **English** | 77.5% | 20.5% |
| **Irish** | 67.2% | 8.6% |
| **Swedish** | 84.2% | 8.9% |
| **Italian** | 85.1% | 5.2% |
| **Chinese** | 58.4% | 5.5% |

# Summary: lower ground truth match rate?

| | Base MR | Age-scaled | Remainder-scaled |
|---|---|---|---|
| **NB White** | 19.10% | 19.10% | 21.3% |
| **NB Black** | 8.30% | 8.84% | 10.5% |
| **Chinese** | 3.20% | 8.42% | 14.4% |
| **Italian** | 4.40% | 5.64% | 6.6% |
| **Swedish** | 7.50% | 12.26% | 14.6% |
| **Irish** | 5.80% | 13.26% | 19.7% |
| **English** | 15.90% | 21.53% | 27.8% |

# Why are subgroups more difficult to link? (Part 2)

Algorithmic challenges.

- Lack of birthplace granularity.
- Names:
  - Transcription/enumeration errors.
  - Name changes.
  - Phonetic standardization algorithms designed for English.

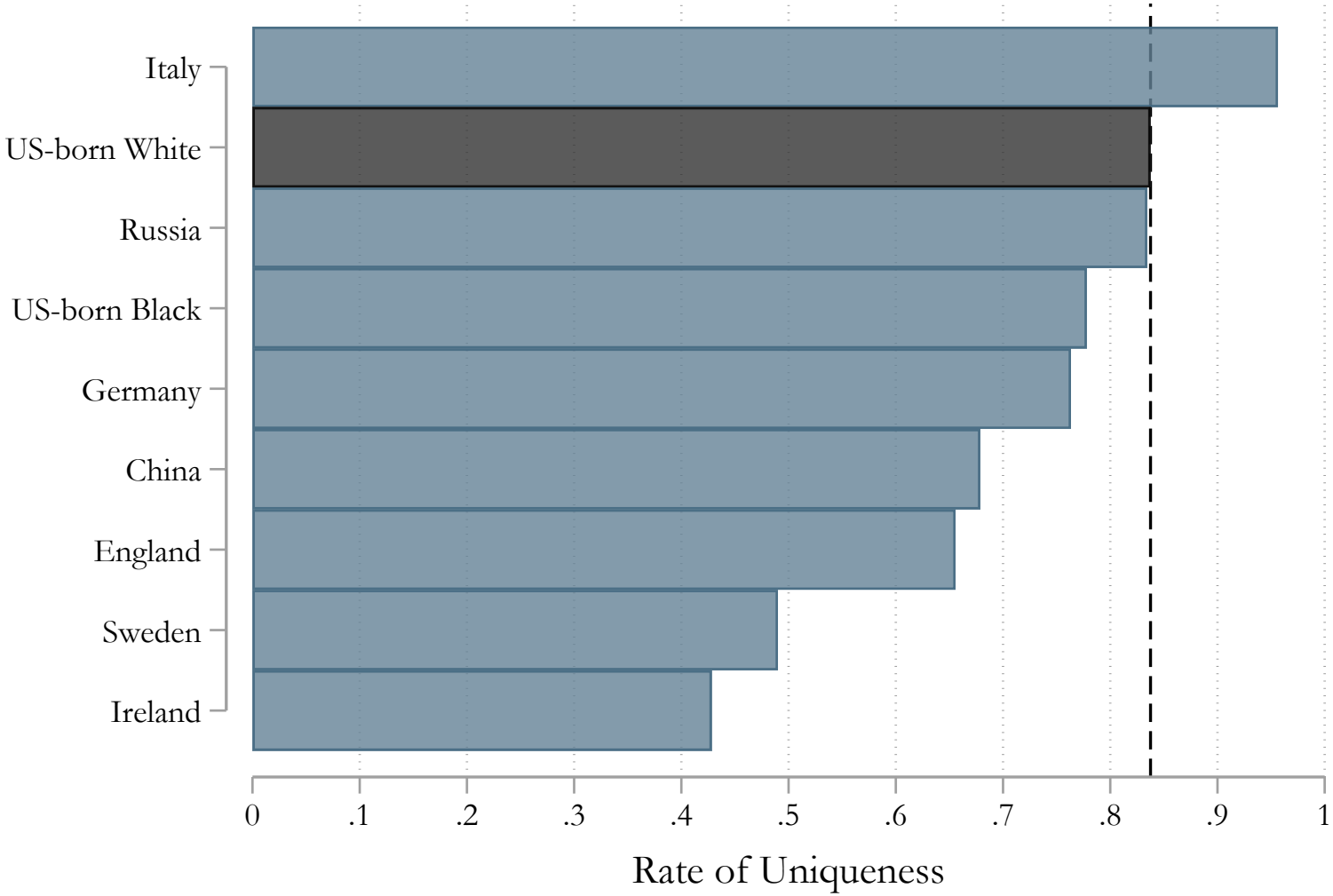
**We can *try* to fix these.**

# Recommendations

1. Add extra information to resolve low uniqueness (Swedish, Irish).
    - This makes name standardization (e.g. NYSIIS) possible.
2. Conduct specialized name processing (Chinese, Hispanic, Italian).

# Recommendation 1: add extra information

Sweden and Ireland have the lowest levels of name-age uniqueness.



# Recommendation 1: add extra information

Adding extra information "breaks ties" to distinguish between individuals who previously looked identical.

The ABE-Extra Information algorithm:

- Select additional variables. Ideally: immutable, always observable.
- Add: parents' names, middle initial, county of residence, spouse's name.
- Create candidate matches using each piece of extra information separately.
- Choose valid matches based on uniqueness criterion.

# Recommendation 1: add extra information

| | ABE (S) | | ABE-EI (S) | |
|---|---|---|---|---|
| | *Match Rate* | *Accuracy*[1] | *Match Rate* | *Accuracy* |
| **Swedish** | 13.2% | 80.6% | 21.9% | 92.6% |
| **Irish** | 13.3% | 71.1% | 18.0% | 91.0% |

| | ABE (C) | | ABE-EI (C) | |
|---|---|---|---|---|
| | *Match Rate* | *Accuracy* | *Match Rate* | *Accuracy* |
| **Swedish** | 7.5% | 90.9% | 18.5% | 95.7% |
| **Irish** | 5.8% | 88.9% | 16.6% | 94.9% |

[1] Comparing matches with "ground truth" Census Tree hand links.

# Recommendation 1: add extra information

EI to break ties and more precisely identify links enables use of phonetic standardization algorithms.

- Developed for English and should be used with caution.
- Sufficiently accurate *only* with conservative specification.

| | ABE1 E-C | ABE EI E-C | ABE EI NYSIIS-C |
|---|---|---|---|
| **NB White** | 19.1% | 28.7% | 35.4% |
| **English** | 15.9% | 26.9% | 31.4% |
| **NB Black** | 8.3% | 14.2% | 17.9% |
| **Swedish** | 7.5% | 18.4% | 23.8% |
| **Irish** | 5.8% | 15.6% | 18.3% |
| **Hispanic** | 5.2% | 4.2% | 6.9% |
| **Italian** | 4.7% | 3.3% | 9.3% |
| **Chinese** | 3.2% | 1.4% | 3.2% |

# Recommendation 2: specialized name cleaning

Three groups still lagging: Chinese, Hispanics, Italians.

Many more issues with name transcription for some immigrant groups.

- Also correlates with uniqueness; e.g. Italians are the highest.

Names aren't "organized" the same way.

- E.g. multi-part names.

# Recommendation 3: specialized name cleaning

1. Chinese: segmentation, name order, & standardization (Postel 2023)
2. Hispanics: segmentation & standardization

(Espín-Sánchez, Gil-Guirado, & Vickers 2021; Pérez 2017)

3. Italians: ?

# Conclusion

Historical record linkage algorithms yield both lower match rates and accuracy for population subgroups.

- For some groups (English, Irish), lower "ground truth" accounts for most of this differential.

Two areas for improvement:

1. Add extra information.
   - This enables out-of-the-box standardization approaches.
2. Conduct specialized name processing.

Need to look at these groups individually!

- We will publish new links.
- Still relevant today (see e.g. Bai et al. 2025)

**Thanks! All feedback welcome.**

hannah.postel@duke.edu

www.censuslinkingproject.org