How Well Do Automated Linking Methods Perform?

Evidence from the LIFE-M Project

Martha Bailey[1,2], Connor Cole[1], Morgan Henderson[1], and Catherine Massey

[1]University of Michigan and [2]NBER

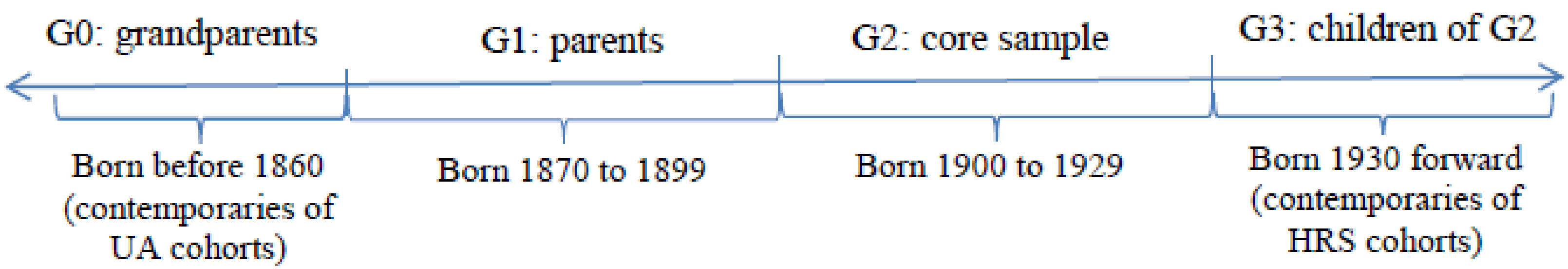
LIFE-M Objectives

- Combine digitized vital records (birth, marriage, & death) with Census
- Create longitudinal, 4-generation dataset span the late 19^{th} and 20^{th} centuries
- Enable high impact research on social and economic outcomes
- Funding from the National Science Foundation and 2 grants from the National Institutes of Health

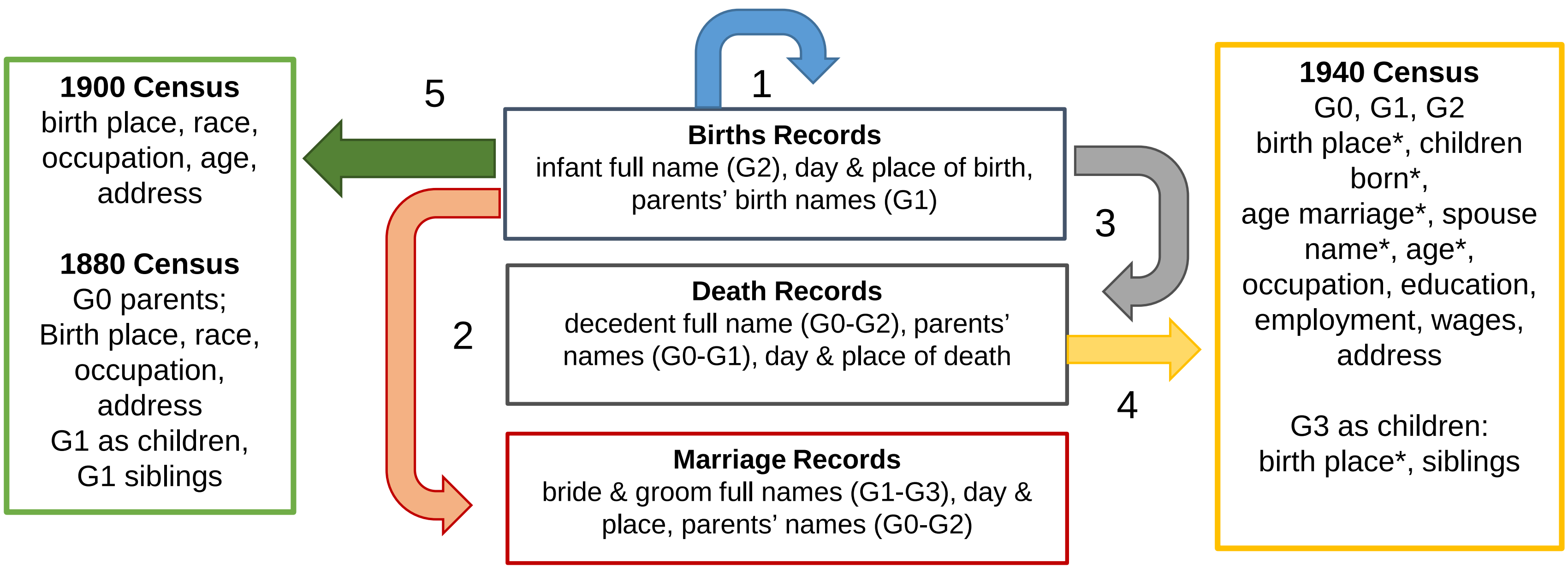
LIFE-M's Contributions

1. Large-scale dataset to provide longitudinal and intergenerational information for health and economic outcomes
2. Unprecedented coverage of women and large samples of racial minorities and immigrants
3. Geographic information facilitates linkages to other datasets

LIFE-M's Contributions



LIFE-M 's Linking Process



Key: G0 born <1860 (~UA cohorts); G1 born 1870-1899; G2 born 1900-1929; G3 born 1930- (~HRS cohorts)

Hand-Linking Process

- Semi-automated: Blind, independent review process

- Two highly trained individuals choosing from a *set* of computer-generated, probabilistic candidate links using name, date of birth (or age), and birth state

- In the three percent of cases where the two initial reviewers disagree, the records are *re*-reviewed by an additional three individuals to resolve these discrepancies

- We also use weekly meetings to discuss difficult linking cases and random "audit batches" to monitor the quality of data links for each trainer

Automated Linking is Crucial to Creating Large Samples

- Automated linking forms the basis of many on-going “big data” projects
  - Hand linking is cost prohibitive

- But…lack of “ground truth” limits evidence on the performance of different automated linking methods in historical settings and samples

This Paper’s Contribution

- Use 2 new high quality samples+synthetic data
  - LIFE-M: Birth certificates for Ohio boys born 1909-1920 linked to 1940 Census; double clerical review with discrepancy resolution
    - 96% of links agree with genealogical sample links
  - Oldest Old Union Army vets: Dora Costa (2016)

- Evaluate the performance of different (implicit) assumptions in linking methods and variations on them using hand-linked data
  - 4 automated linking methods in current practice
  - Variations on deterministic algorithms
    - 2 phonetic name cleaning: NYSIIS and Soundex
    - Using common names
    - Weighting ties

Prominent Algorithms for Linking Historical Data

<u>Deterministic</u>

- Ferrie (1996) tries to link names that appear less than 10 times (cleans name and uses age differences to choose best link)
- Abramitzky, Boustan, and Eriksson (2012, 2014) implement a similar algorithm but search for matches before dropping common names
  - Extension: even common names may have matches if we include multiple dimensions (like age and birth place)

<u>Probabilistic</u>

- Feigenbaum (2016) supervised method fitting a regression of record features to classify matches (uses training data)
- Abramitzky, Mill, and Perez's (2018) unsupervised method uses Expectation-Maximization algorithm (Fellegi and Sunter 1969, Winkler 2006, Dempster, Laird, and Rubin 1977) to classify records (no training data)

Data: Ohio and North Carolina boys hand-linked to 1940 Census

Births Records
random samples of birth certificates

→

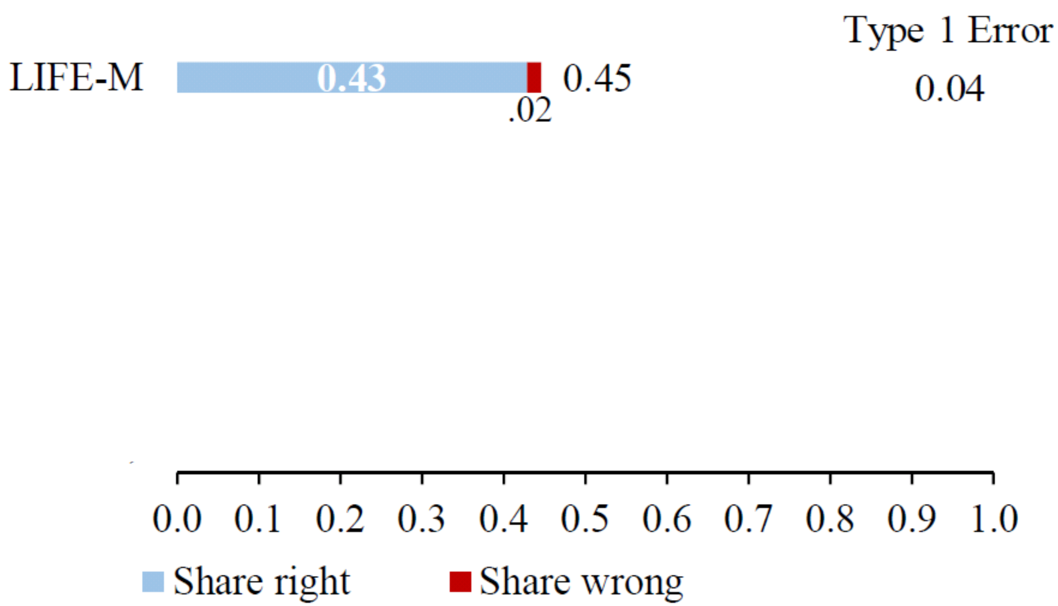
1940 Census
birth place*,
children born*,
age marriage*,
spouse name*,
age*, occupation,
education,
employment,
wages, address

- ~42,000 birth certificates which we try to link to the 1940 Census

- Vetted against genealogical method:
    1. Joe Price at BYU used family history students to hand link 1000 of our boys to the 1940 census
    2. 96 percent of links agree (4% disagreement)

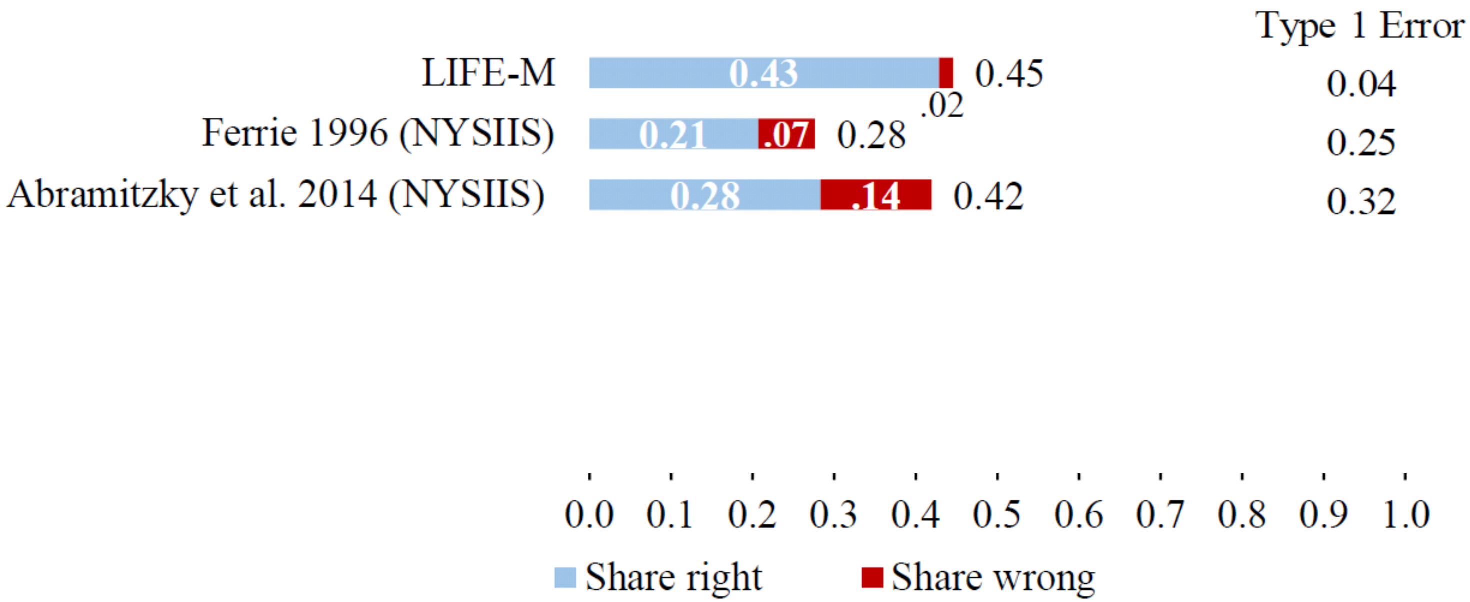
False links: Police Line Up

Performance of Prominent Methods

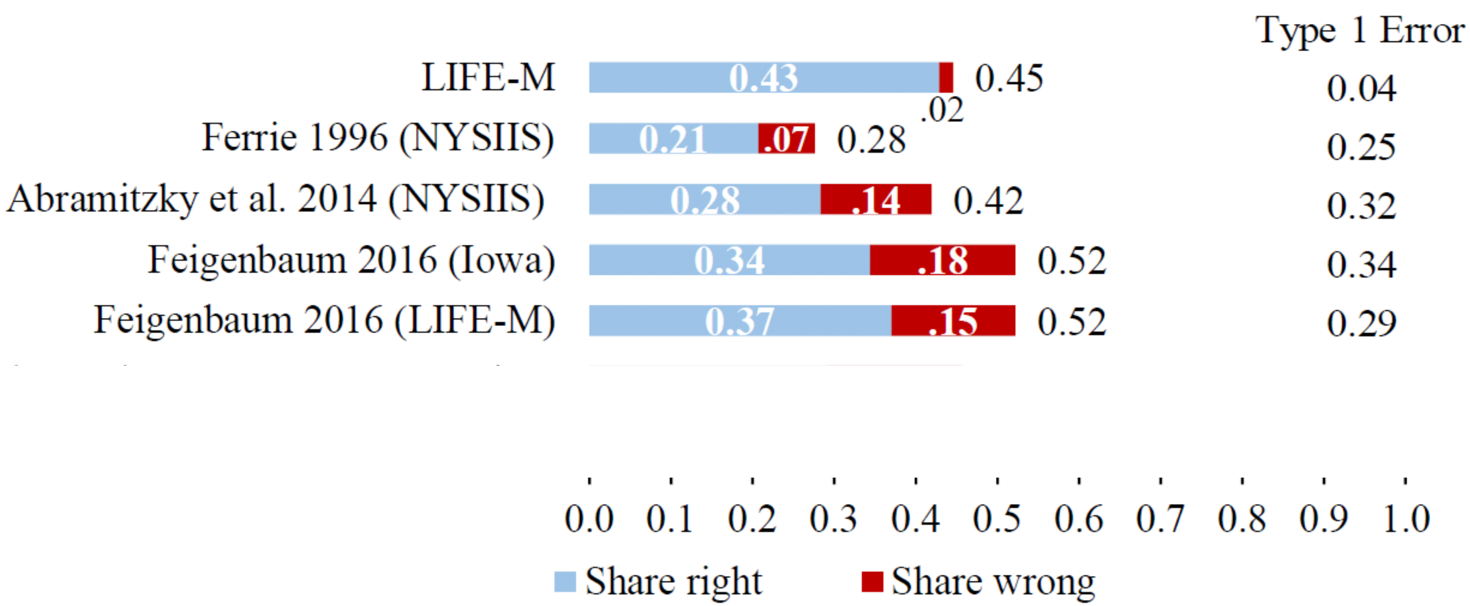


Performance of Prominent Methods

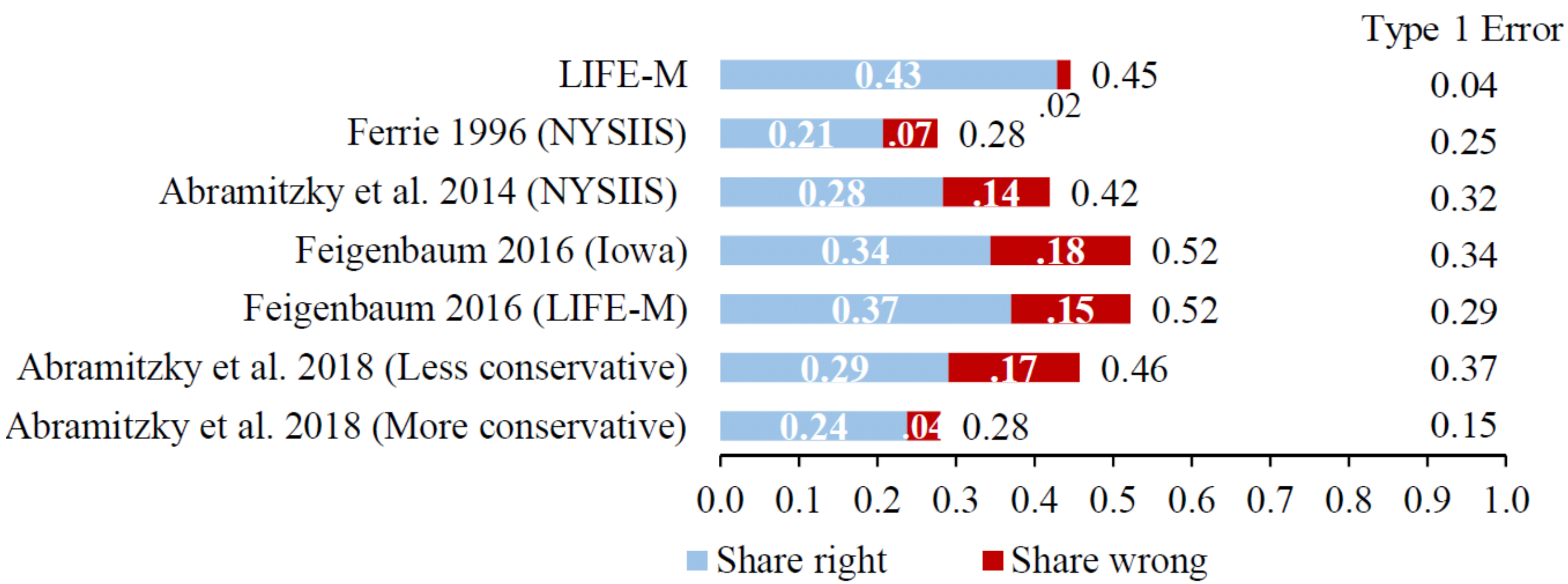
Performance of Prominent Methods



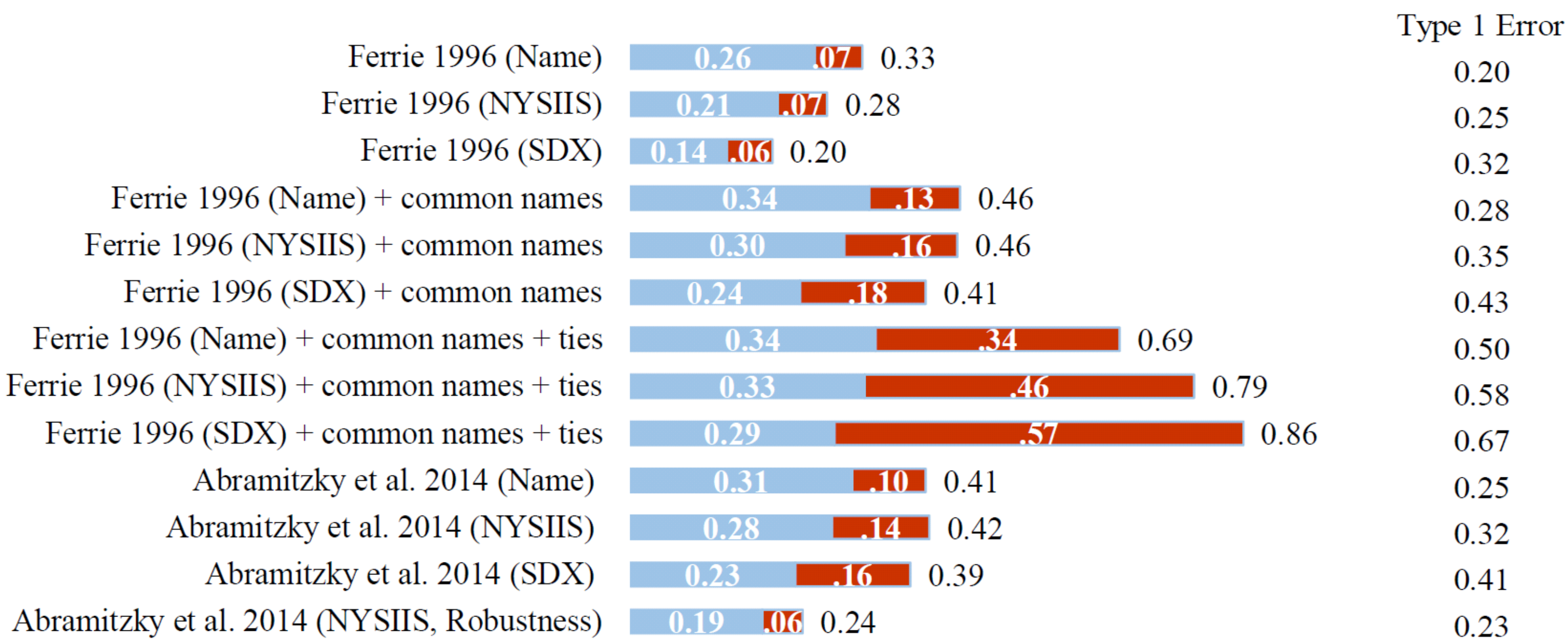
Performance of Prominent Methods



Performance of Prominent Methods

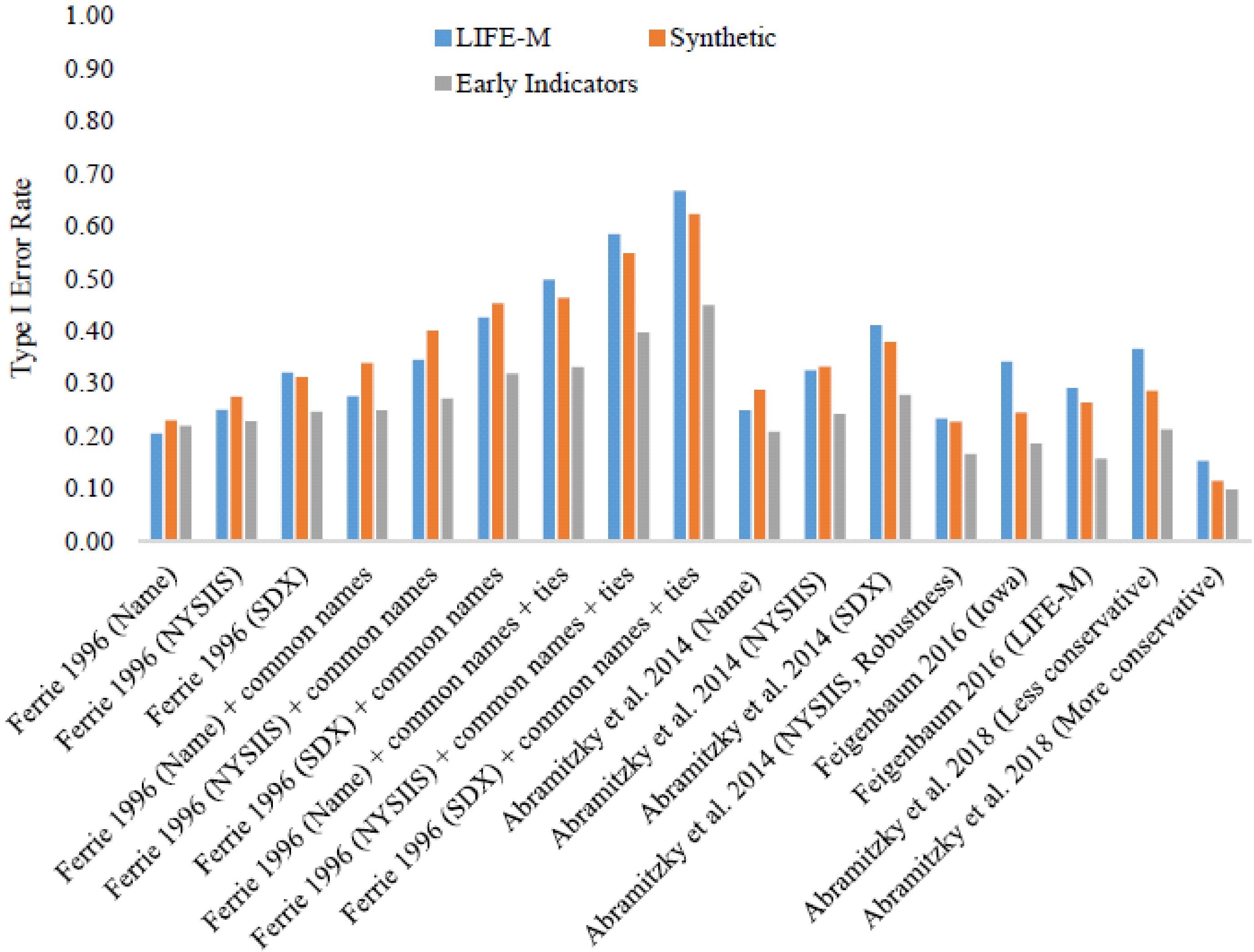


Variations: Phonetic cleaning, common names, and ties



1. Phonetic name cleaning increases Type I errors and does not necessarily increase true links
2. Linking common names doubles Type I errors but does increase true links
3. Using ties dramatically increases Type I errors with little effect on true links

Validate Conclusions using Synthetic Ground Truth and Early Indicators Sample



Performance Summary

1. True matches
   - Between 24 and 43 percent
2. False positives (Type I errors): bad links
   - Between 15 and 41 percent
3. Representativeness
   - No method achieves this
4. Representativeness of false links
   - No method achieves this, suggesting linking algorithms introduce complicated forms of selection bias and measurement error

Intergenerational Income Elasticities

- How does linking affect social science inferences?
- Depends crucially on how it error is related to the underlying observed and unobserved characteristics as well as composition of final sample

IGEs for 1920-1940

- Is the U.S. the land of opportunity? How economically mobile are people?

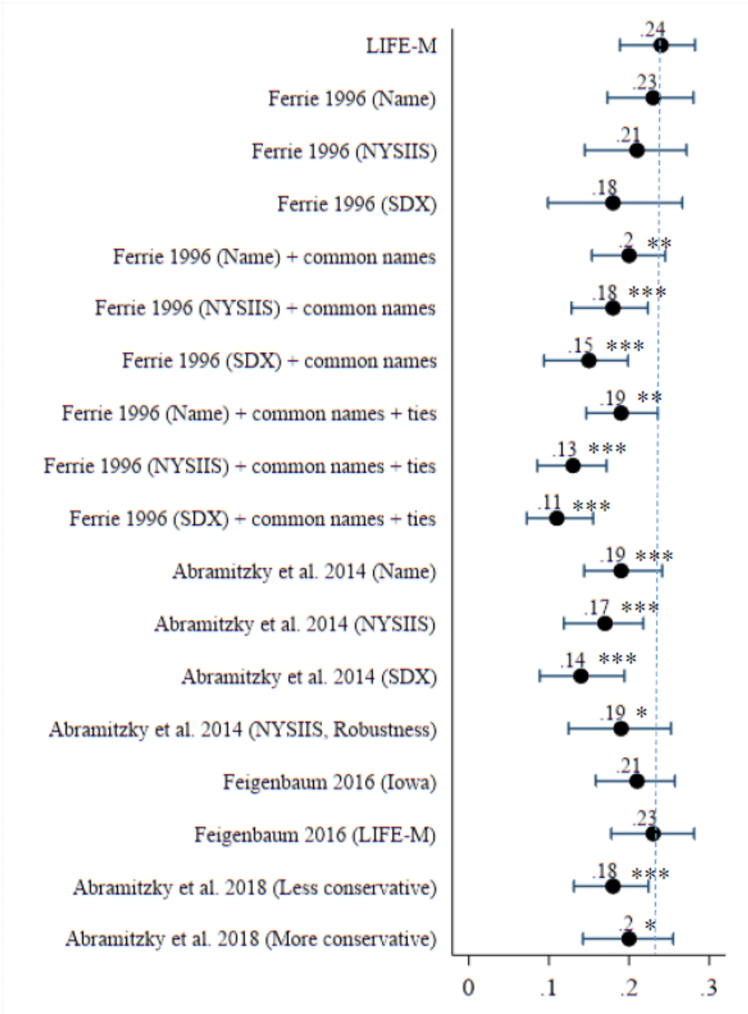
- Standard IGE regressions

$$\log (y_2) = \beta \log (y_1) + \varepsilon,$$

β is interpreted as the intergenerational earnings elasticity (IGE)
(intergenerational mobility is often measured as 1-beta)

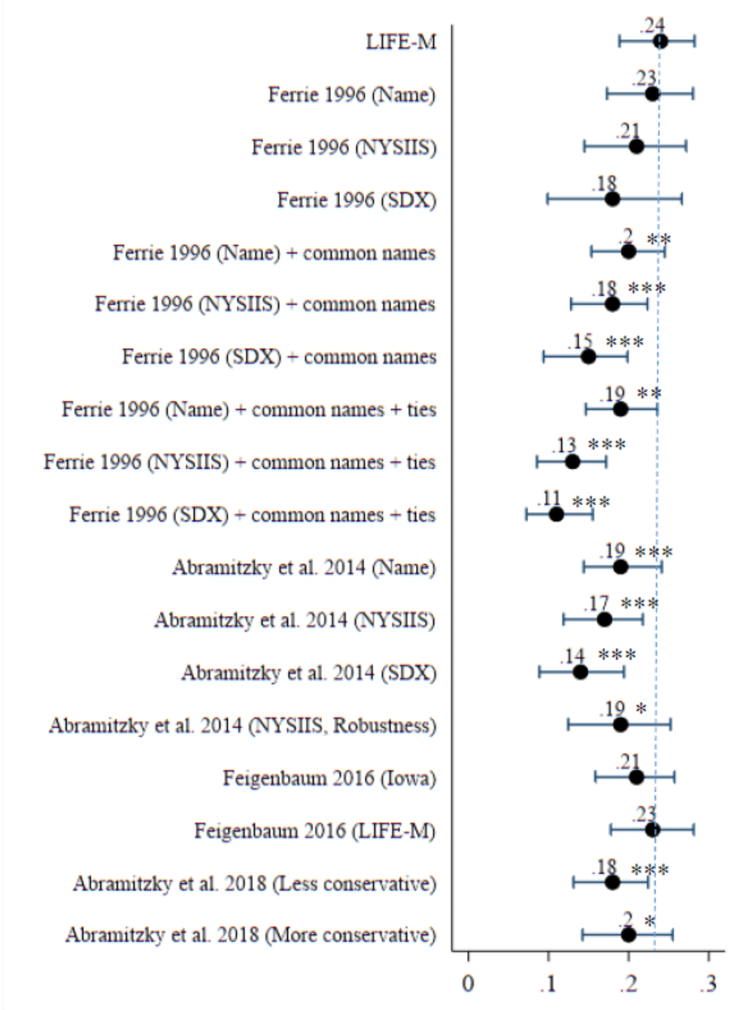
Measurement Error Attenuates Results

A. Unweighted Linked Samples

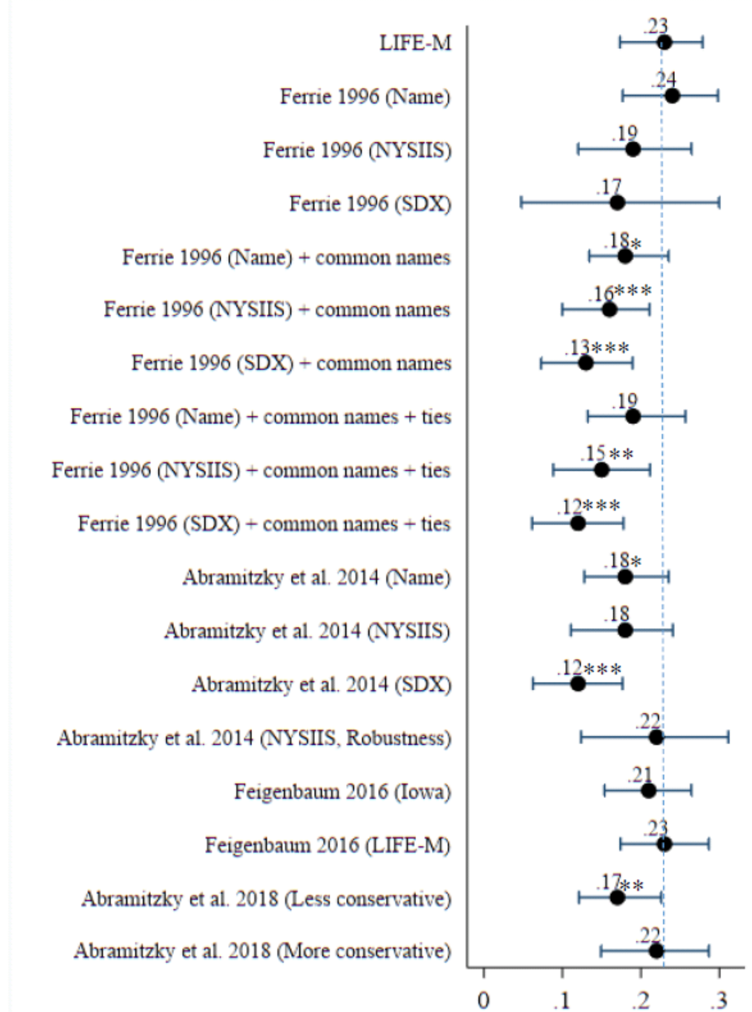


…But Sample Composition Matters Less

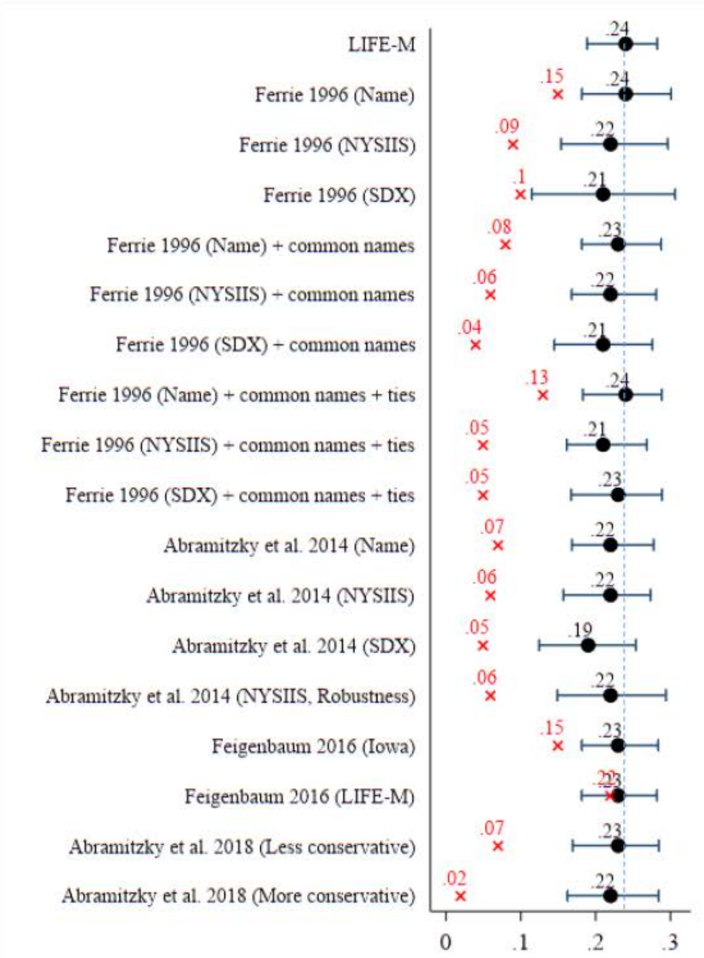
A. Unweighted Linked Samples

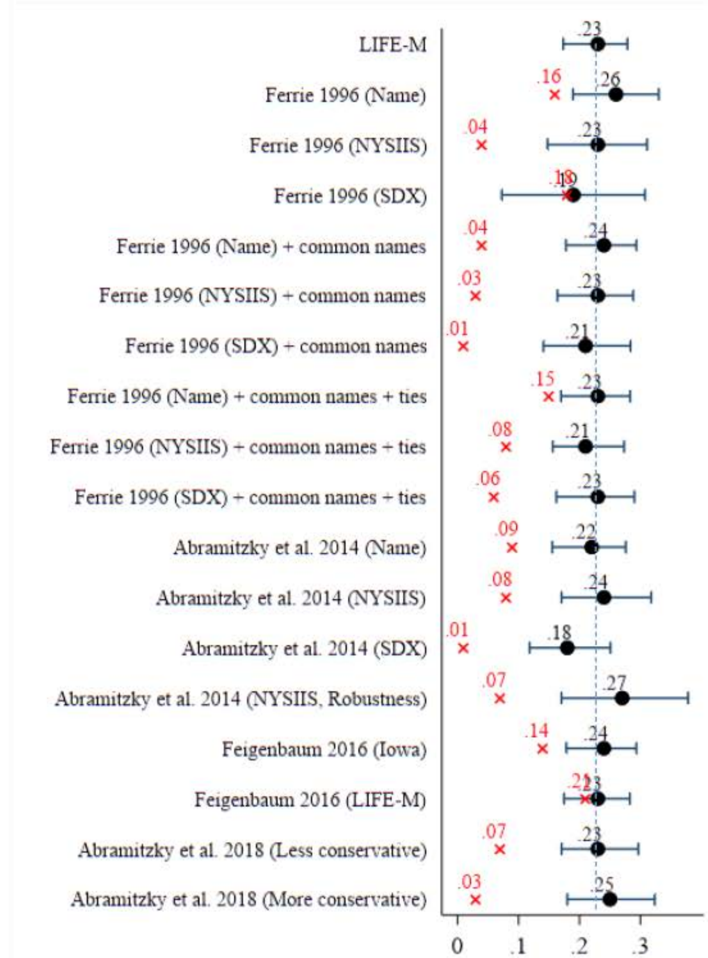


B. Inverse Propensity-Score Weighted Linked Samples



Incorrect v. Correct Links

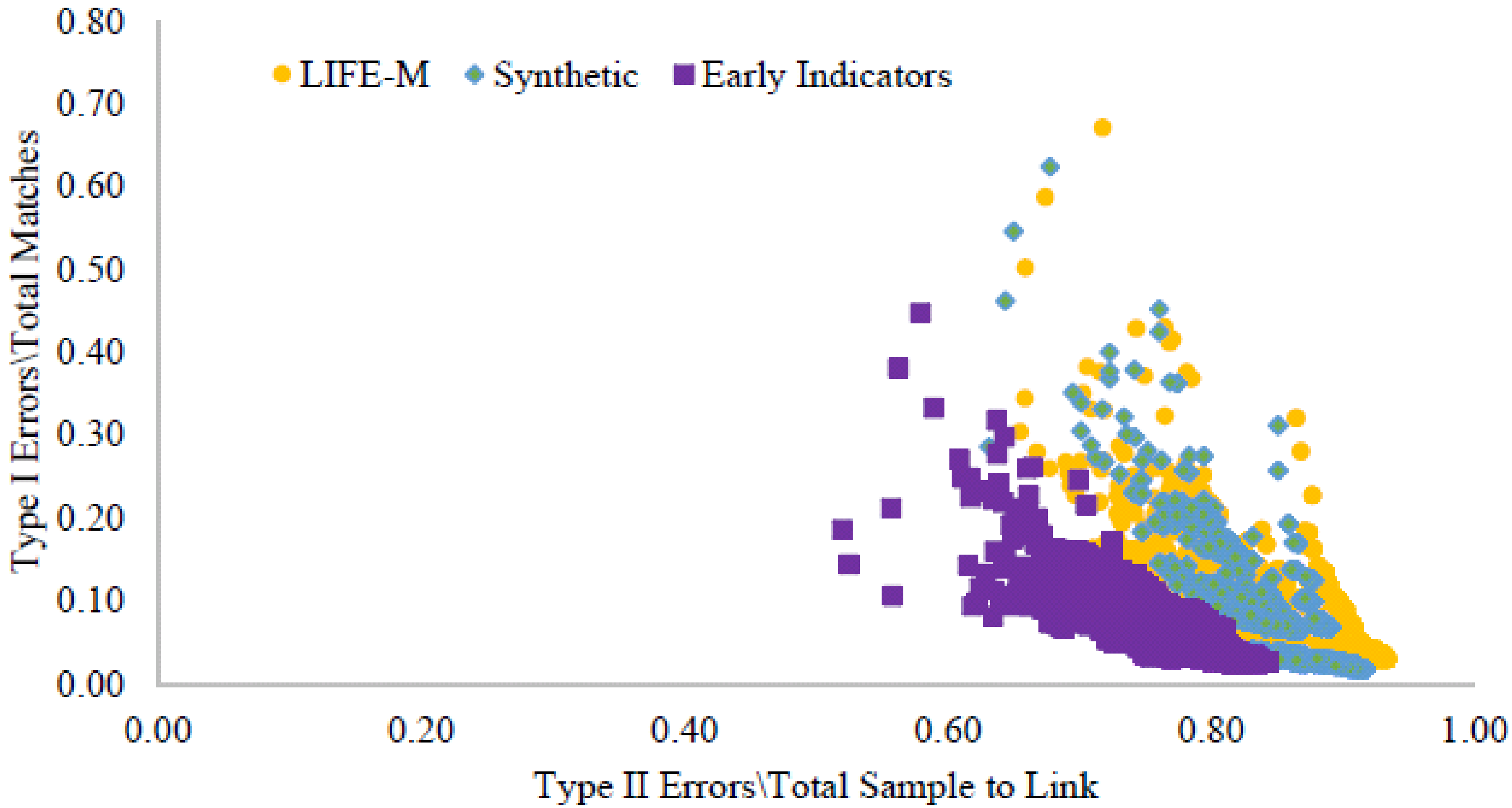




Bottom line: measurement error matters a lot!

Recommendations

1. Combine multiple methods



Constructive Suggestions

1. Combine multiple linked methods
    - Stata do-files are available: autolink.ado
    - discard problematic cases
    - diagnose type I errors and their causes
    - combine to reduce errors

2. Do not use NYSIIS and Soundex as a blocking strategy in deterministic algorithms.
    - Errors arising from these name-cleaning algorithms appear systematically related to a number of record characteristics, making it unclear how they should affect inferences

3. Consider many record features to assess sample representativeness and create weights
    - Make greater use of common record features such as name length or exact day of birth (when available) may provide important information about sample representativeness.
    - Use inverse-propensity weights for linked samples to help balance both observed and potentially unobserved characteristics (DiNardo et al. 1996, Heckman et al. 1998)