How Well Does Automated Linking Perform (in Historical Samples)? Lessons for Modern Practice

MARTHA BAILEY, UNIVERSITY OF MICHIGAN AND NBER

JOINT WORK WITH CONNOR COLE, MORGAN HENDERSON, AND CATHERINE MASSEY

Dynamic Questions

How has the human experience evolved and why?

What factors—environmental or human made—have interacted to improve well being or hold back economic development?

What have been the long-run effects of policies, innovations, environmental factors, and public health efforts?

Need for Dynamic Data

Dynamic questions relate to lives and experiences vary *across time*

Until recently, most U.S. data spanning the late 19th and 20th centuries were cross-sectional—individuals *at one point in time*

Need for Dynamic Data

Dynamic questions relate to lives and experiences vary *across time*

Until recently, most U.S. data spanning the late 19th and 20th centuries were cross-sectional—individuals *at one point in time*

New data include

- 1940 full-count census
- 1850-1930 IPUMS Linked Historical Samples (linked to 1880)
- Possible links of historical to modern data with ALIRA, CLIP, AoS
- LIFE-M links from vital records to 1880-1940 (coming 2020)

New Data Require New Tools

Management of (very) large and complex data

Tools to link data

Theoretical and econometric tools to *use* linked data *wisely*

Outline of Talk

Overview of historical linking methods

Summarize method performance in four datasets

Case study: IGE estimates for the 1940s

Suggestions for modern practice

Measuring intergenerational mobility c. 1940

$$\log(y_1) = \pi \log(y_0) + \varepsilon$$

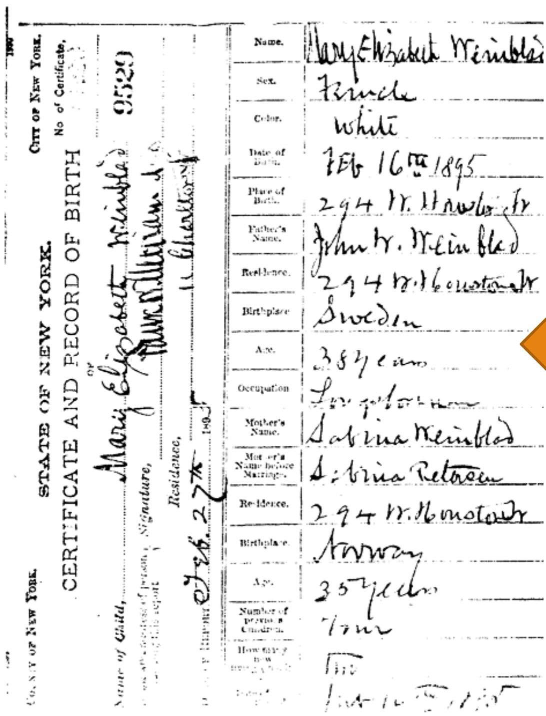
π is interpreted as the intergenerational earnings elasticity

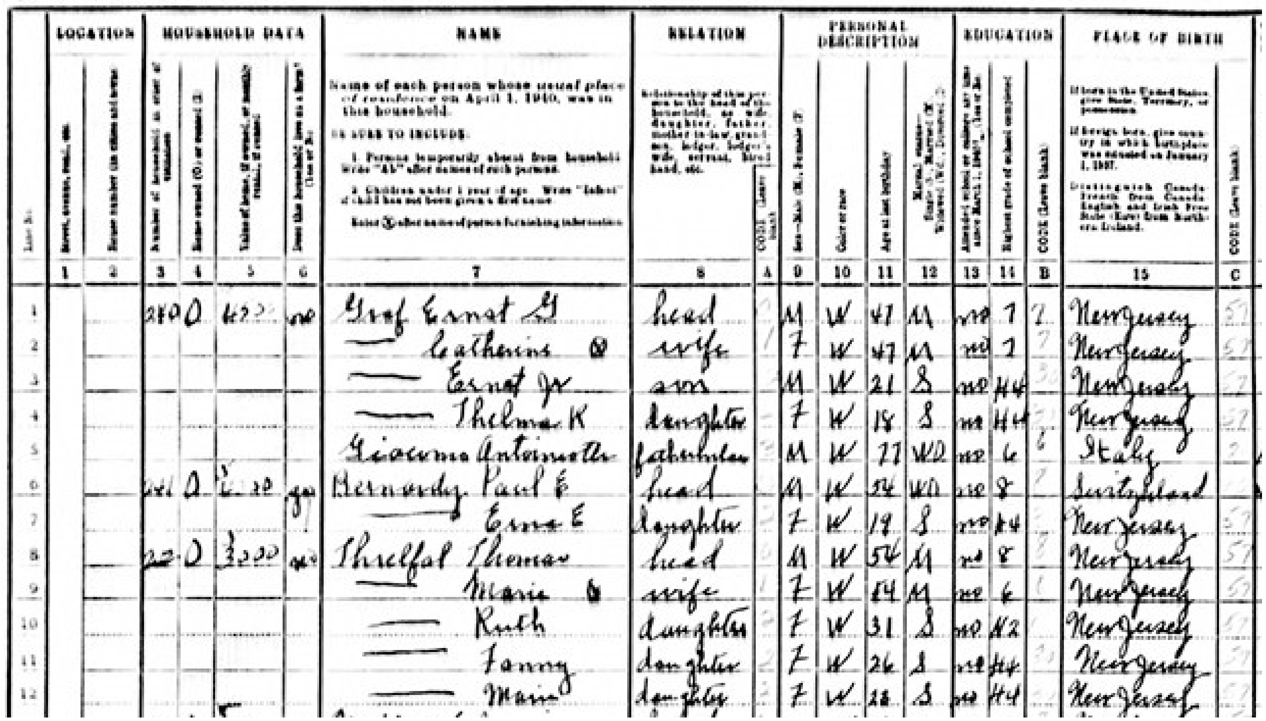
The larger π, the more persistent social class

and the less equal economic opportunity

($1-\pi$ is interpreted as the intergenerational mobility)

US Record Linkage (Early 20th Century)

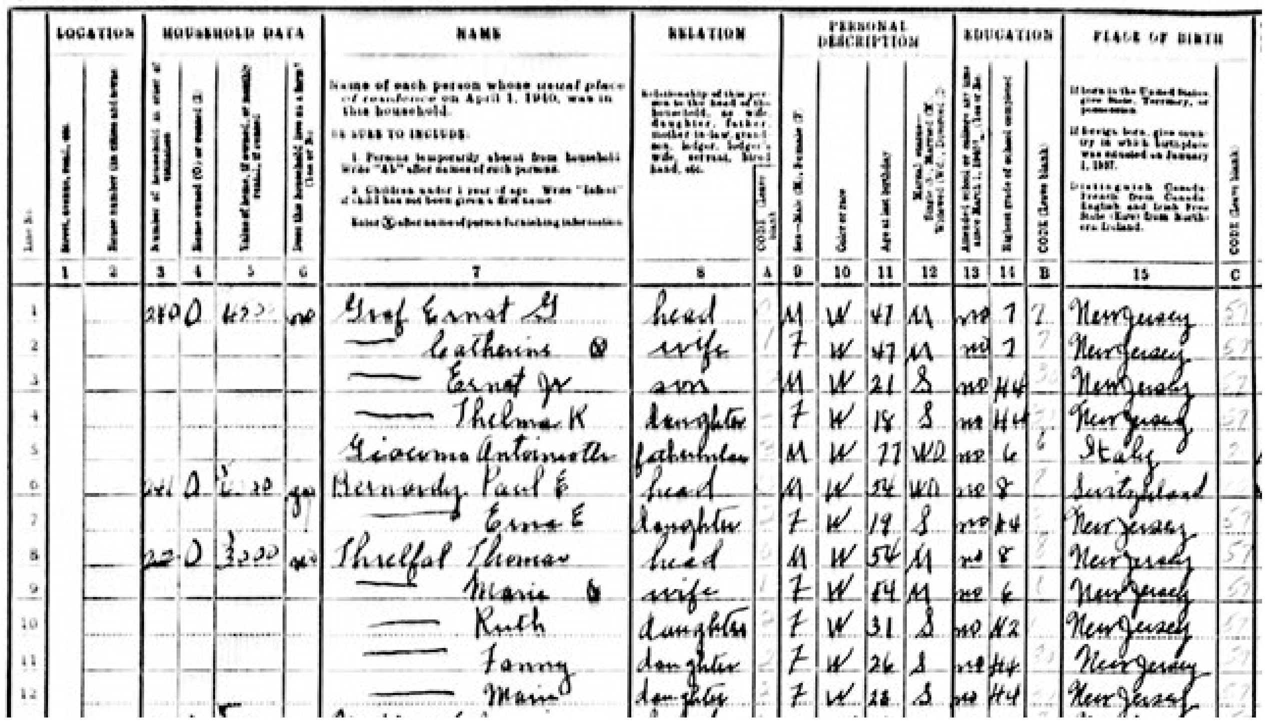




US Record Linkage (Early 20th Century)

Problems

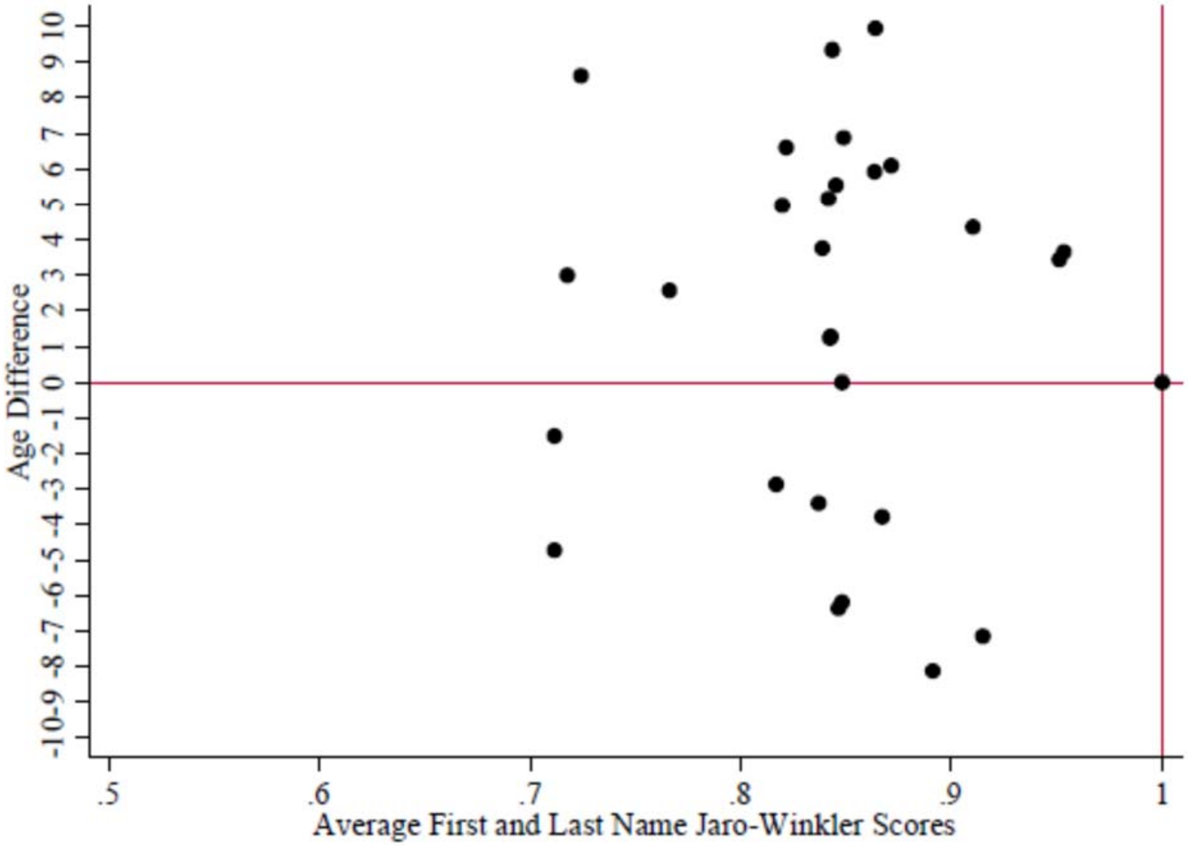
1. Misreports by individual
2. Errors in enumeration
3. Errors in transcription



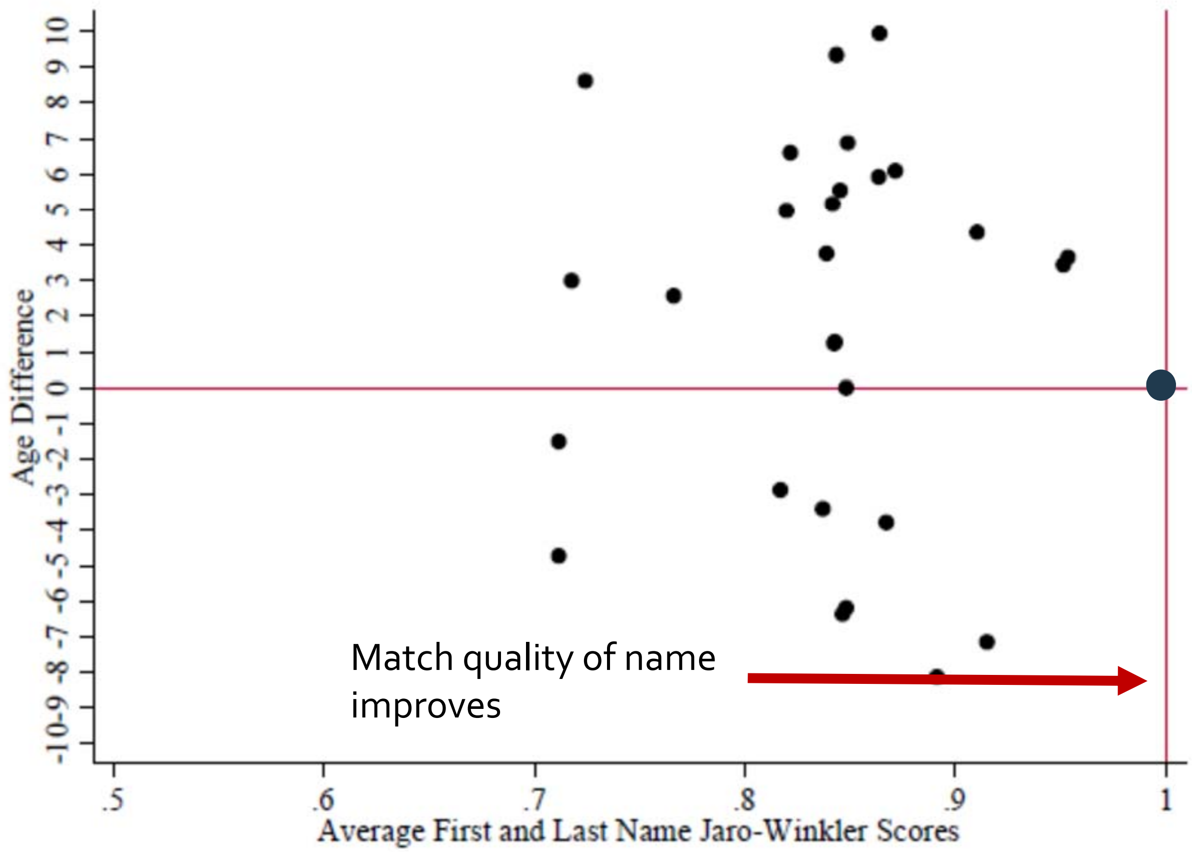
What Record Linking Algorithms Do

LINKING TO US CENSUS DATA

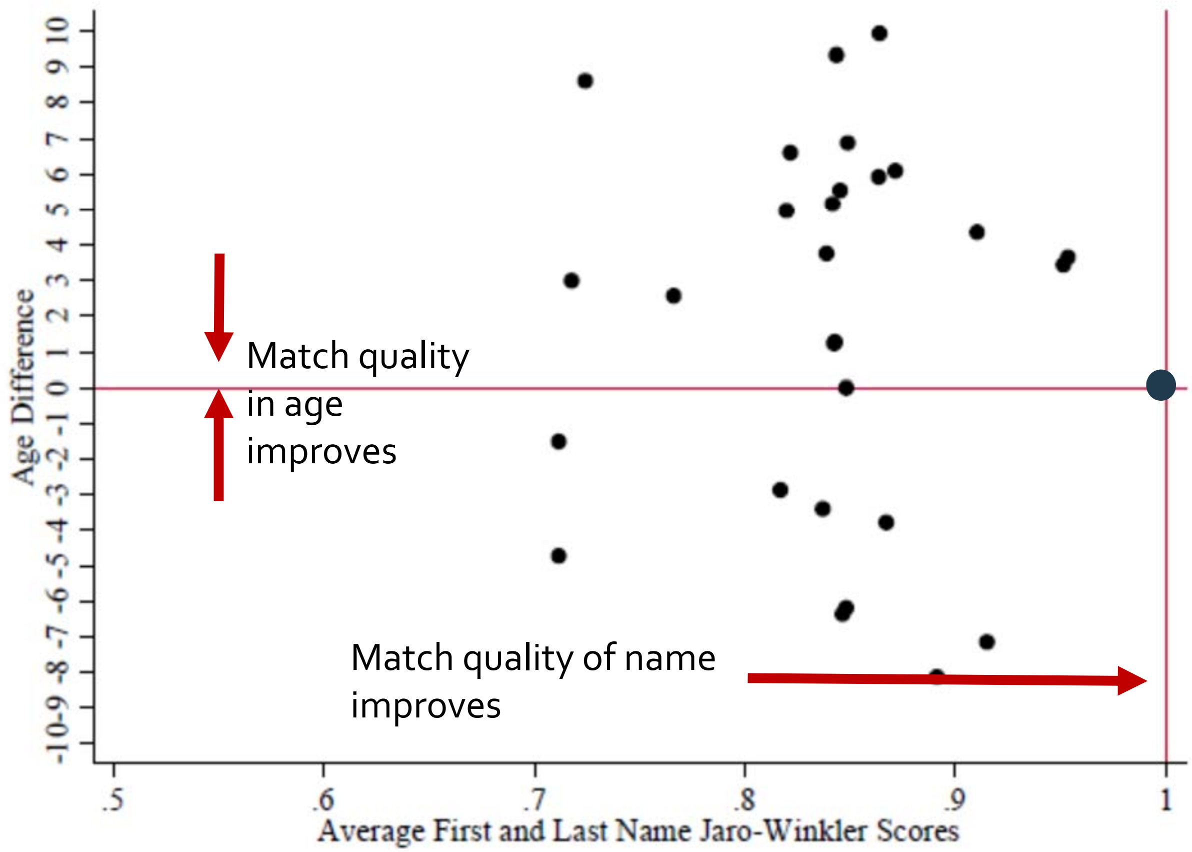
Example Matching Problem



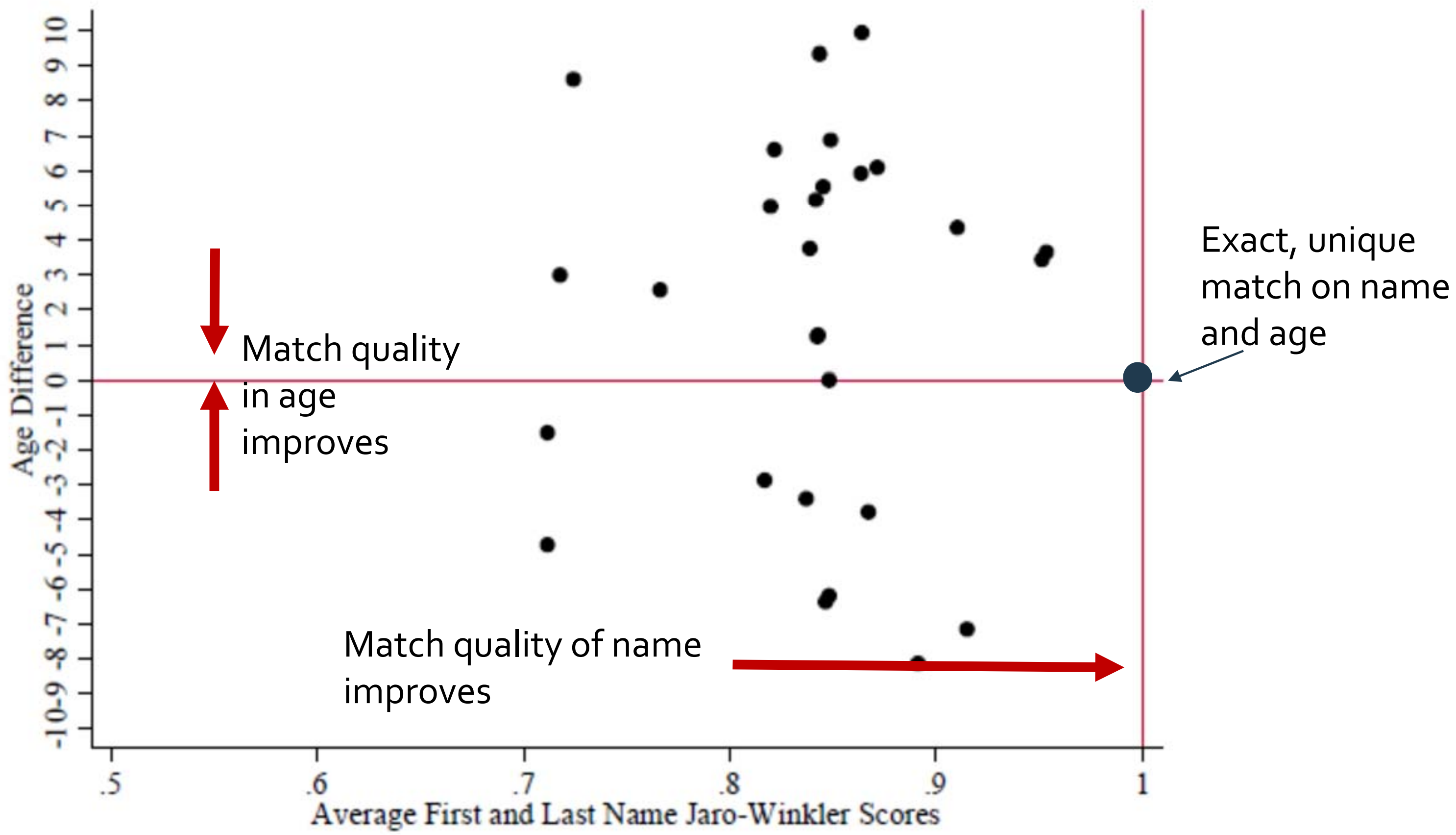
Example Matching Problem



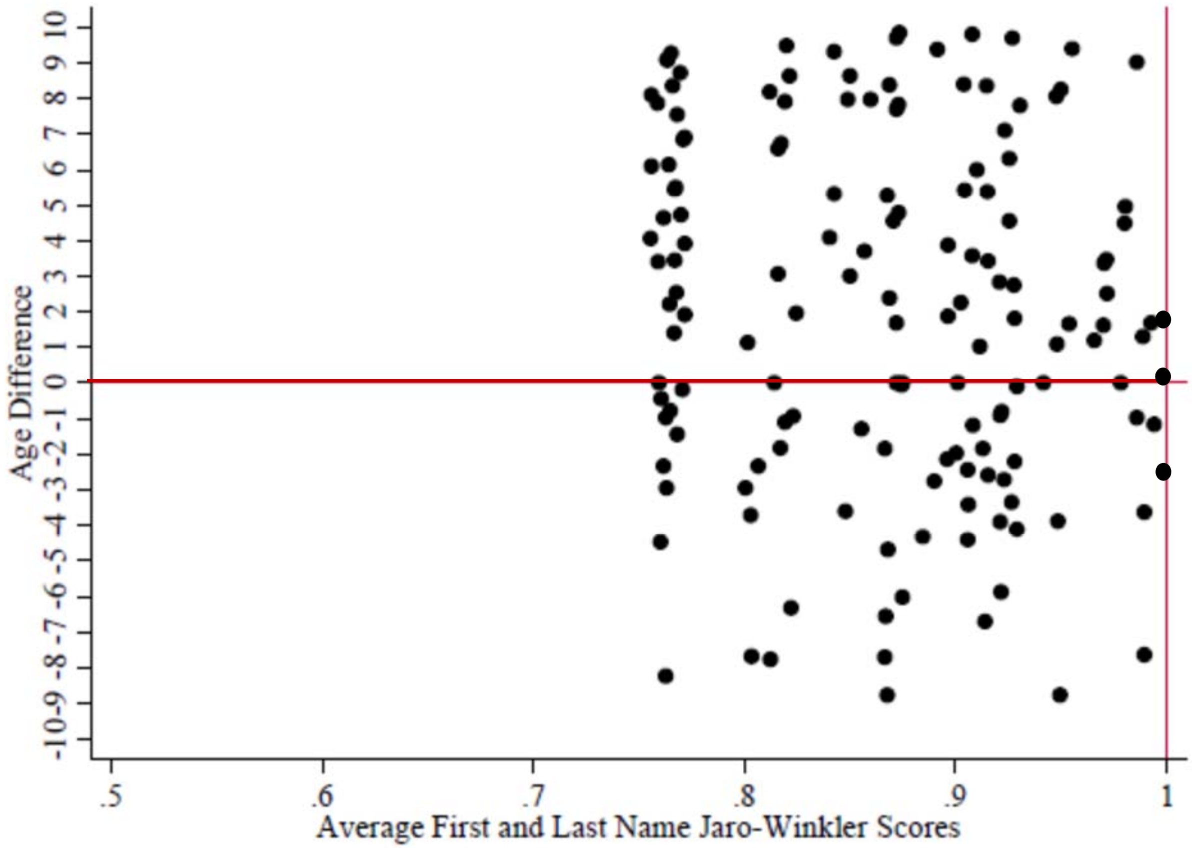
Example Matching Problem



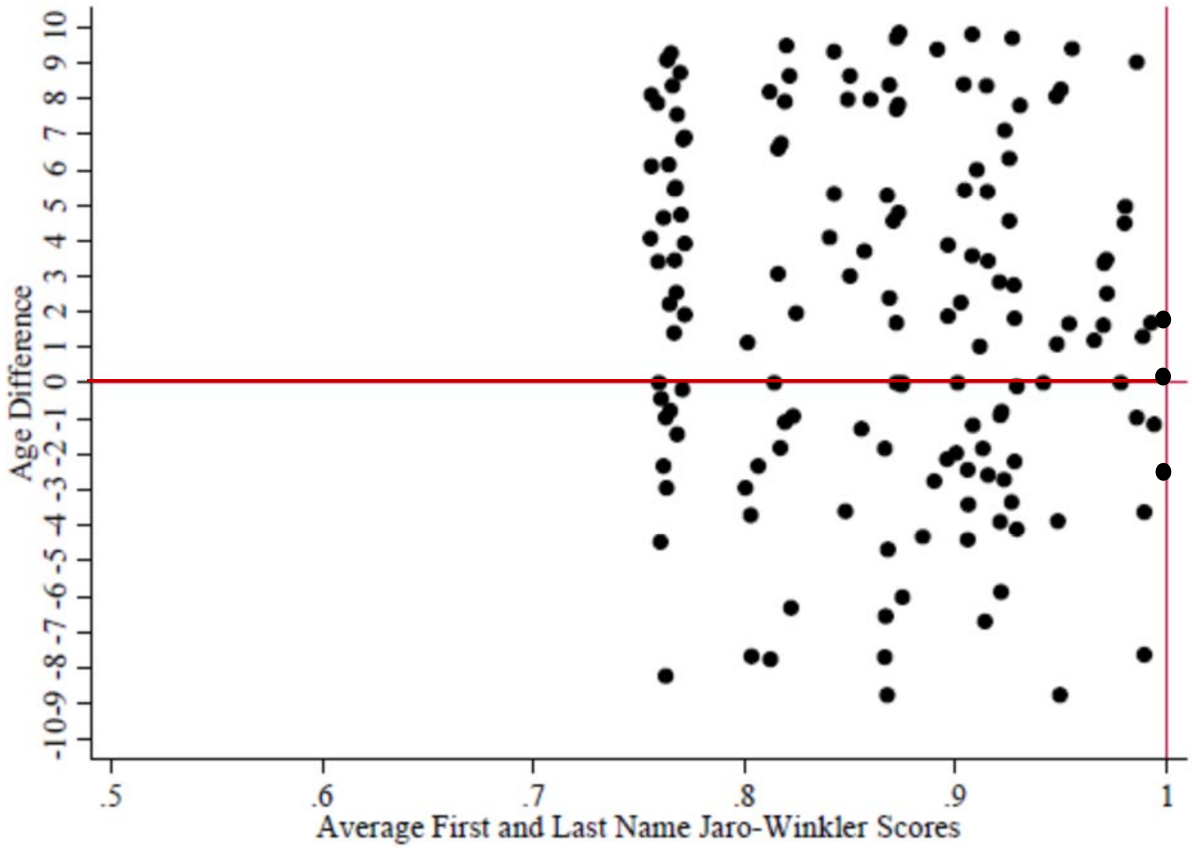
Example Matching Problem



Ferrie (1996)

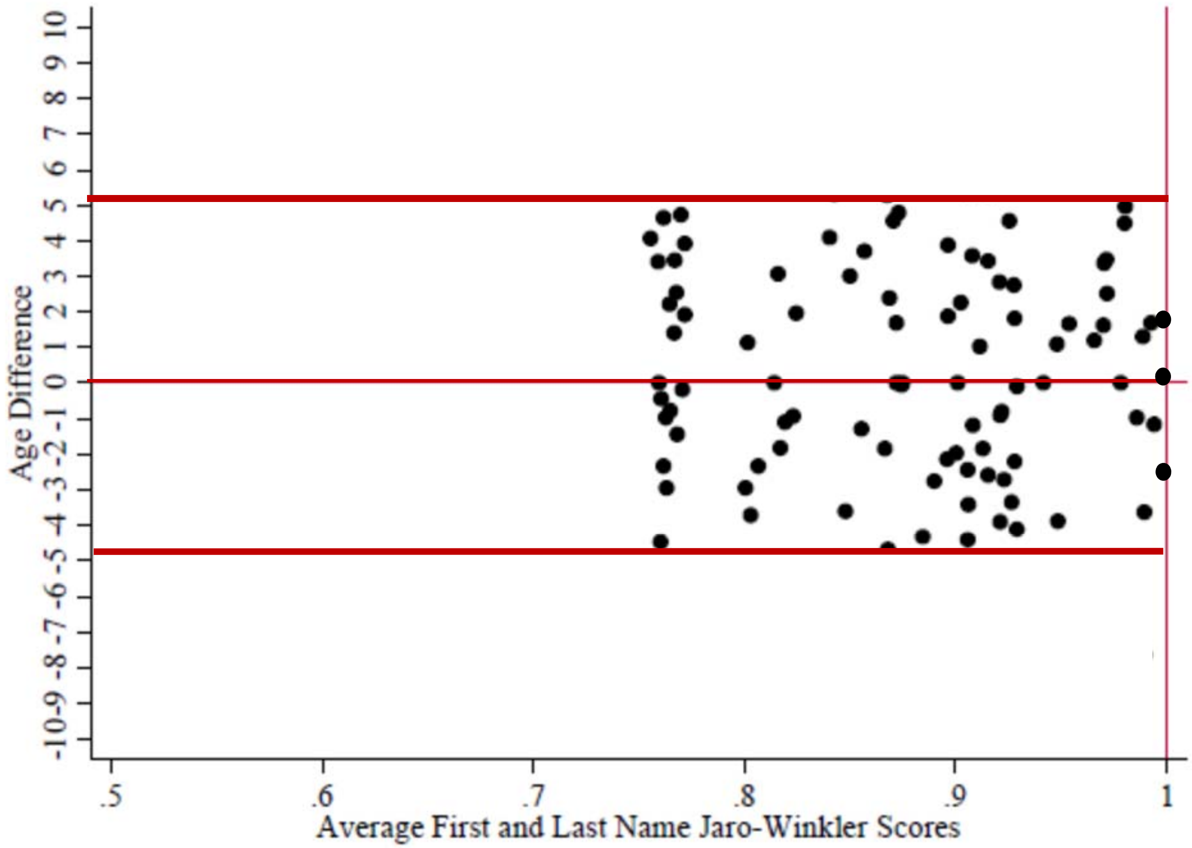

Age Difference
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
.5
.6
.7
.8
.9
1
Average First and Last Name Jaro-Winkler Scores

Ferrie (1996)



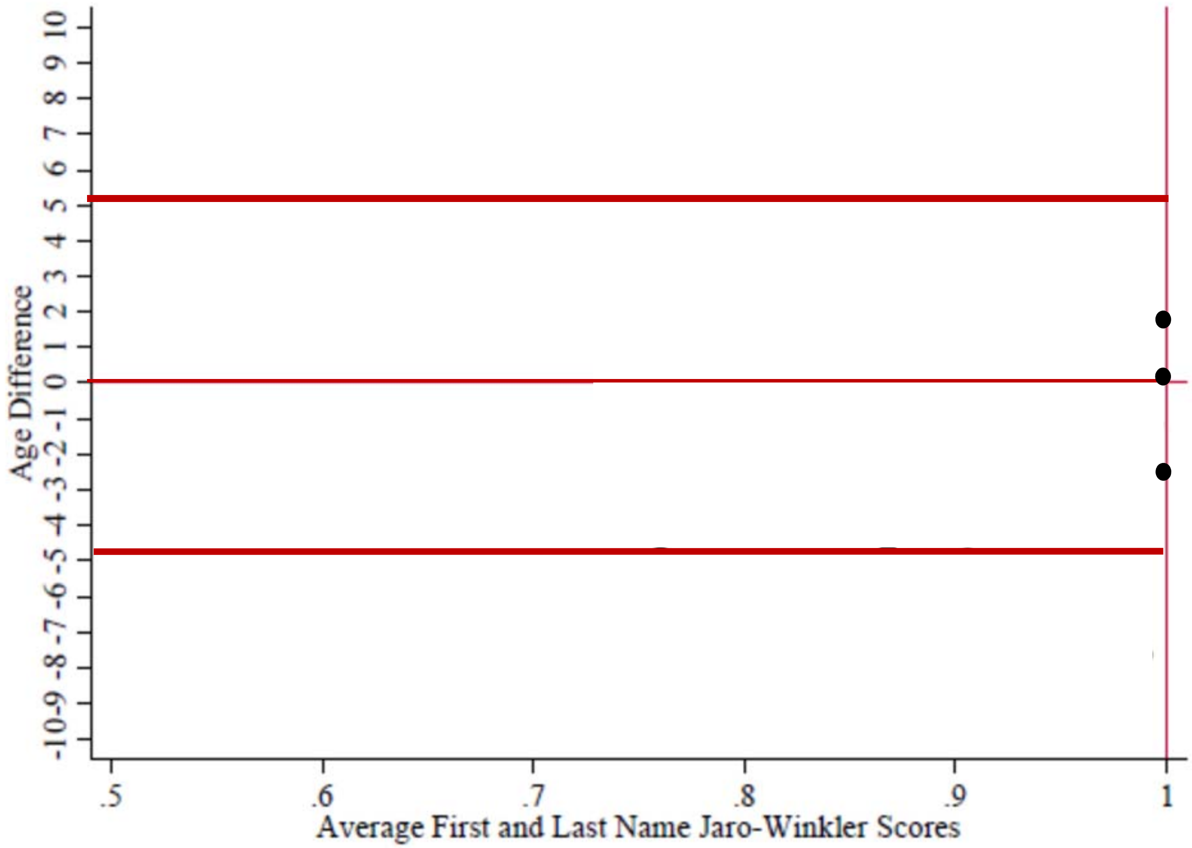
1. Uses uncommon name sample

Ferrie (1996)



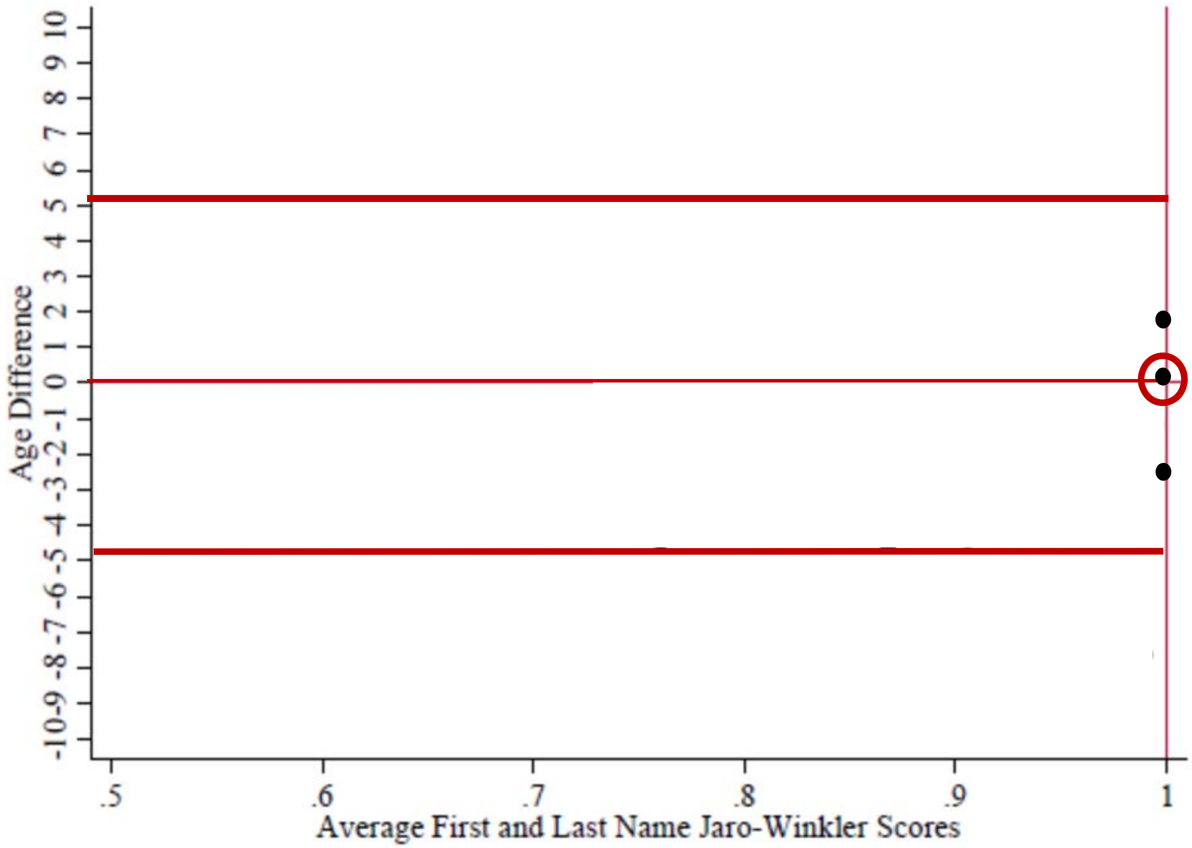
1. Uses uncommon name sample
2. Restricts age difference to be +/-5

Ferrie (1996)



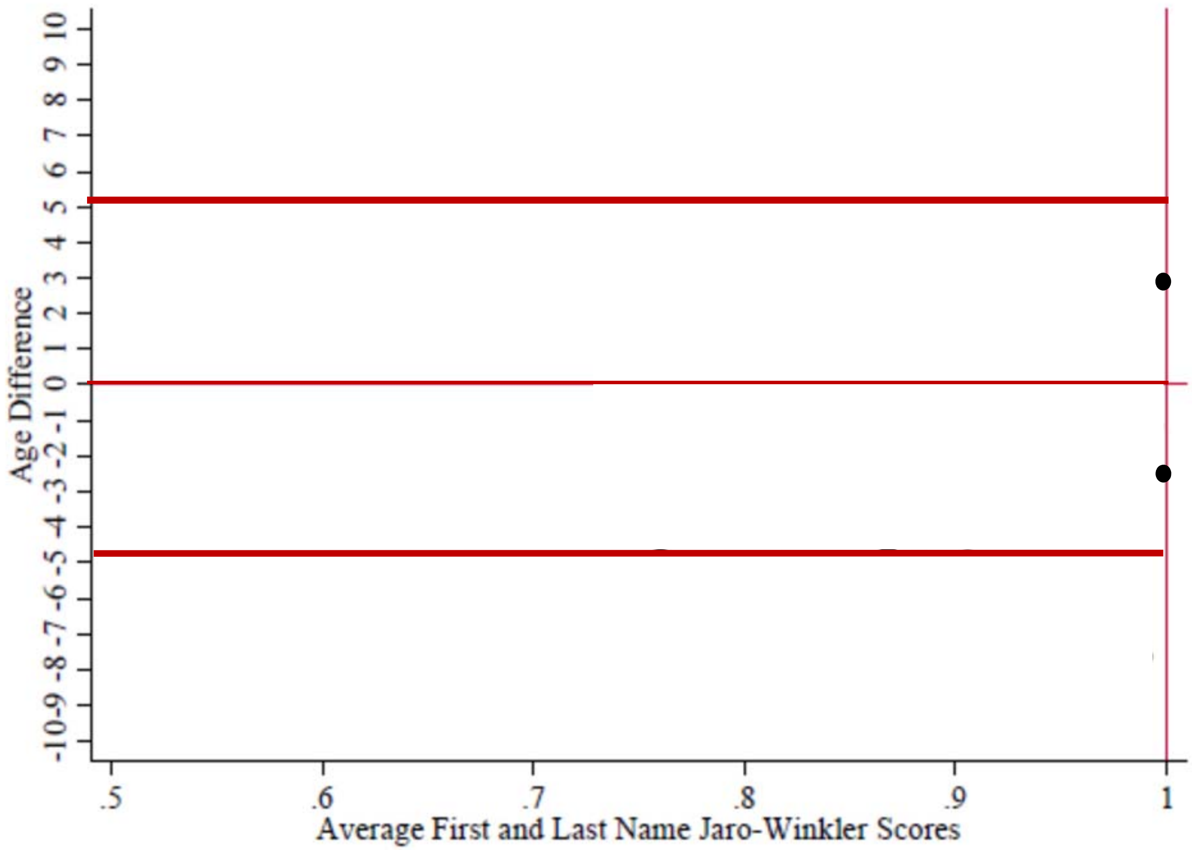
1. Uses uncommon name sample
2. Restricts age difference to be +/-5
3. Finds exact name matches

Ferrie (1996)



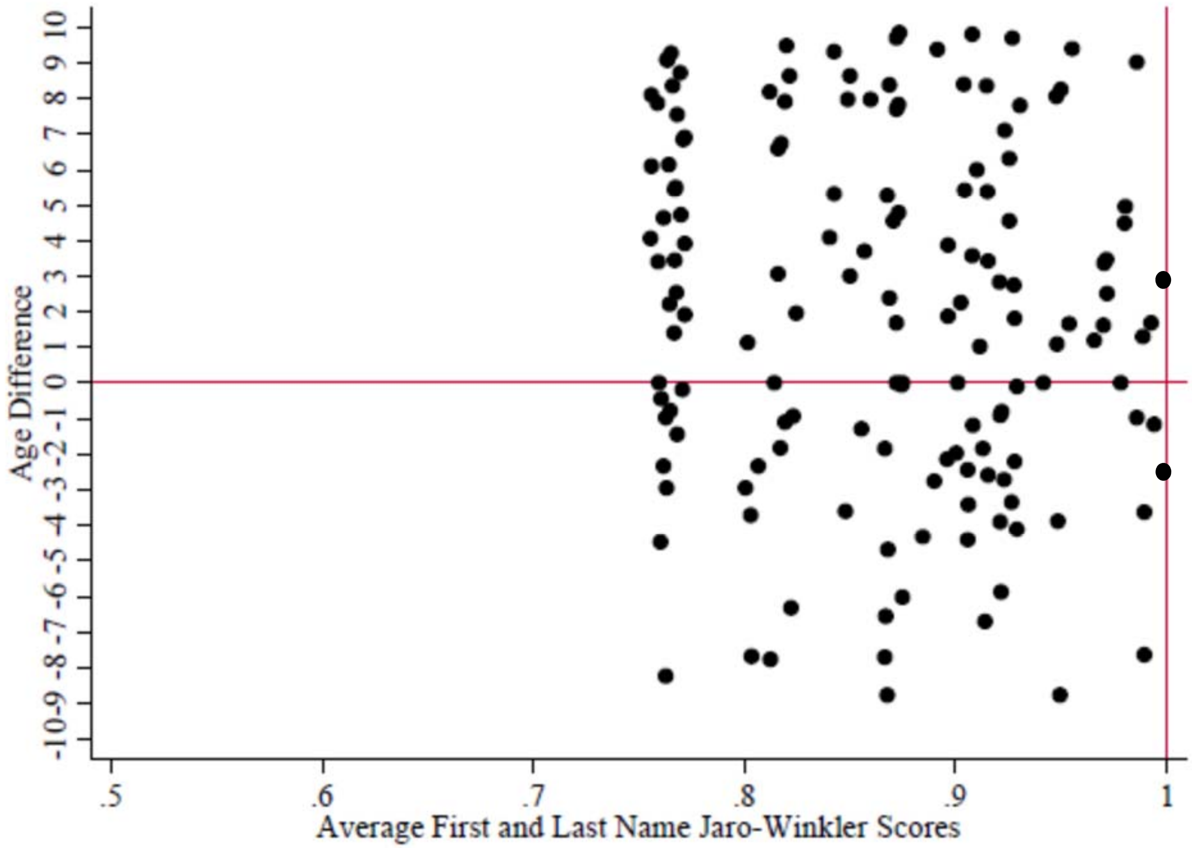
1. Uses uncommon name sample
2. Restricts age difference to be +/-5
3. Finds exact name matches
4. Minimizes age difference

Ferrie (1996)



5. No match chosen when ties

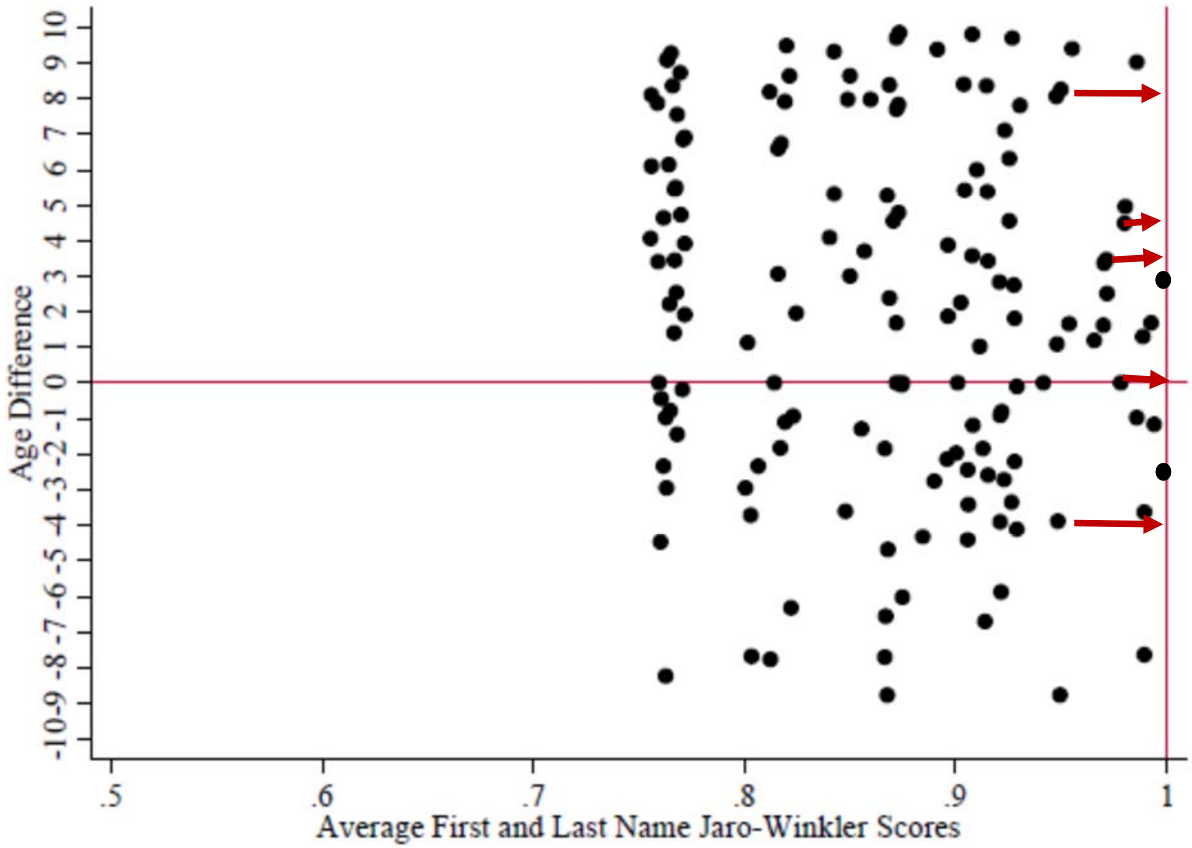
Phonetic Name Cleaning



Soundex and NYSIIS

- Soundex: “Smith,” “Smyth” and “Smythe” to the same code (S530)
- NYSIIS: “Wilhem” and “William” to WALAN

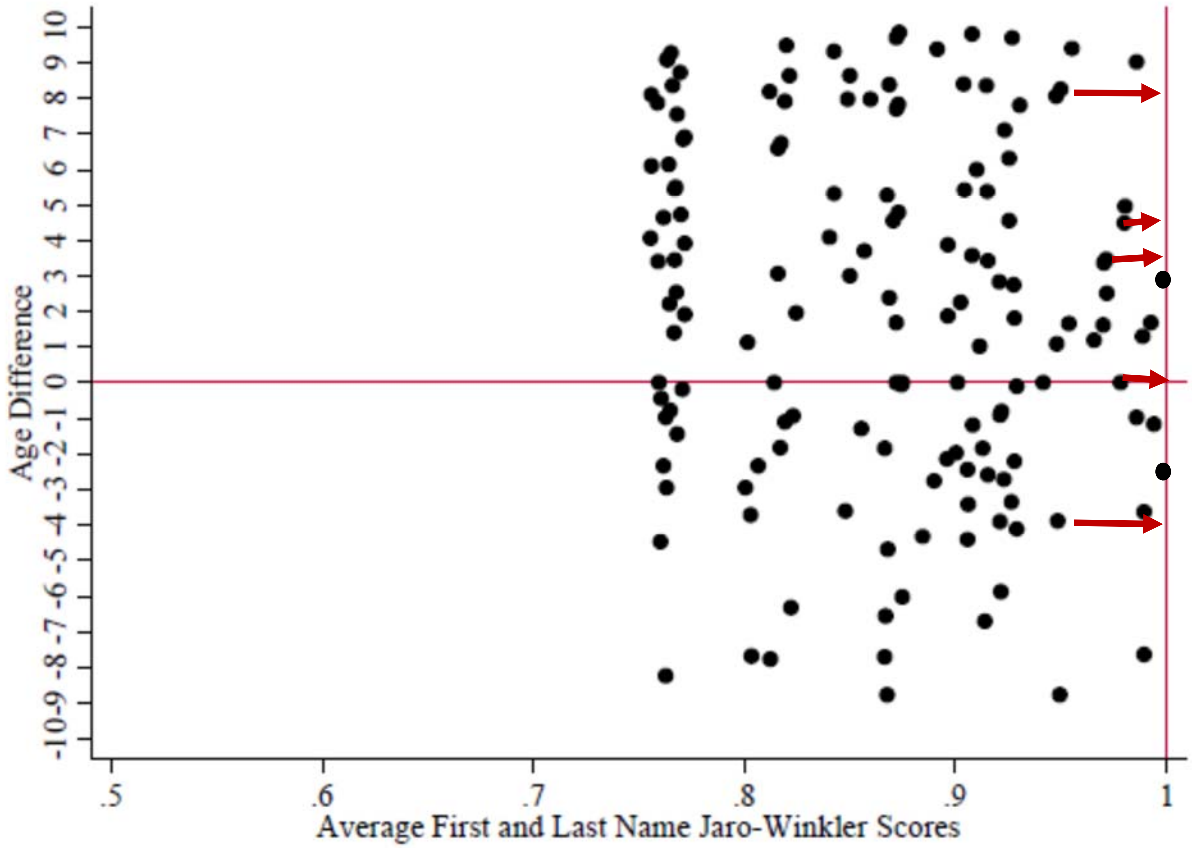
Phonetic Name Cleaning



Soundex and NYSIIS

- Soundex: “Smith,” “Smyth” and “Smythe” to the same code (S530)
- NYSIIS: “Wilhem” and “William” to WALAN

Phonetic Name Cleaning

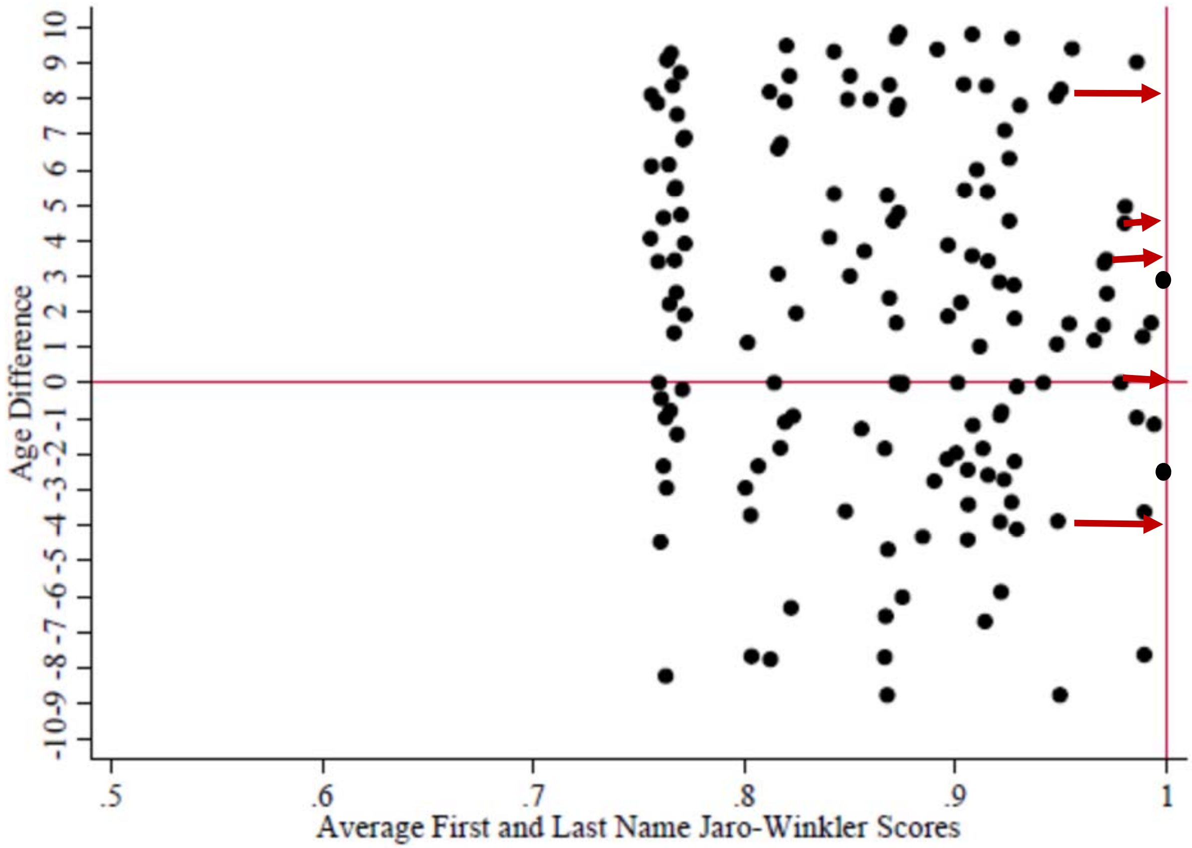


Soundex and NYSIIS

- Soundex: "Smith," "Smyth" and "Smythe" to the same code (S530)
- NYSIIS: "Wilhem" and "William" to WALAN

Ferrie (1996) uses NYSIIS

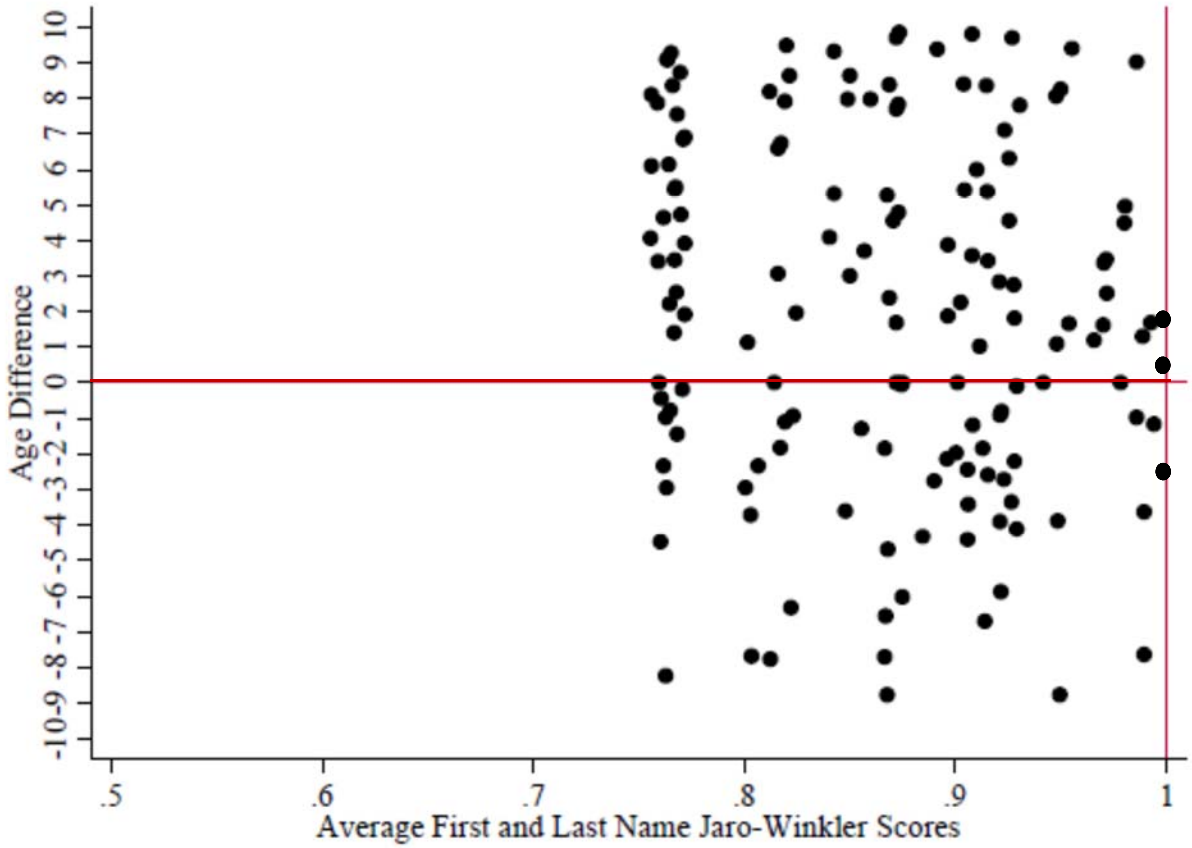
Phonetic Name Cleaning



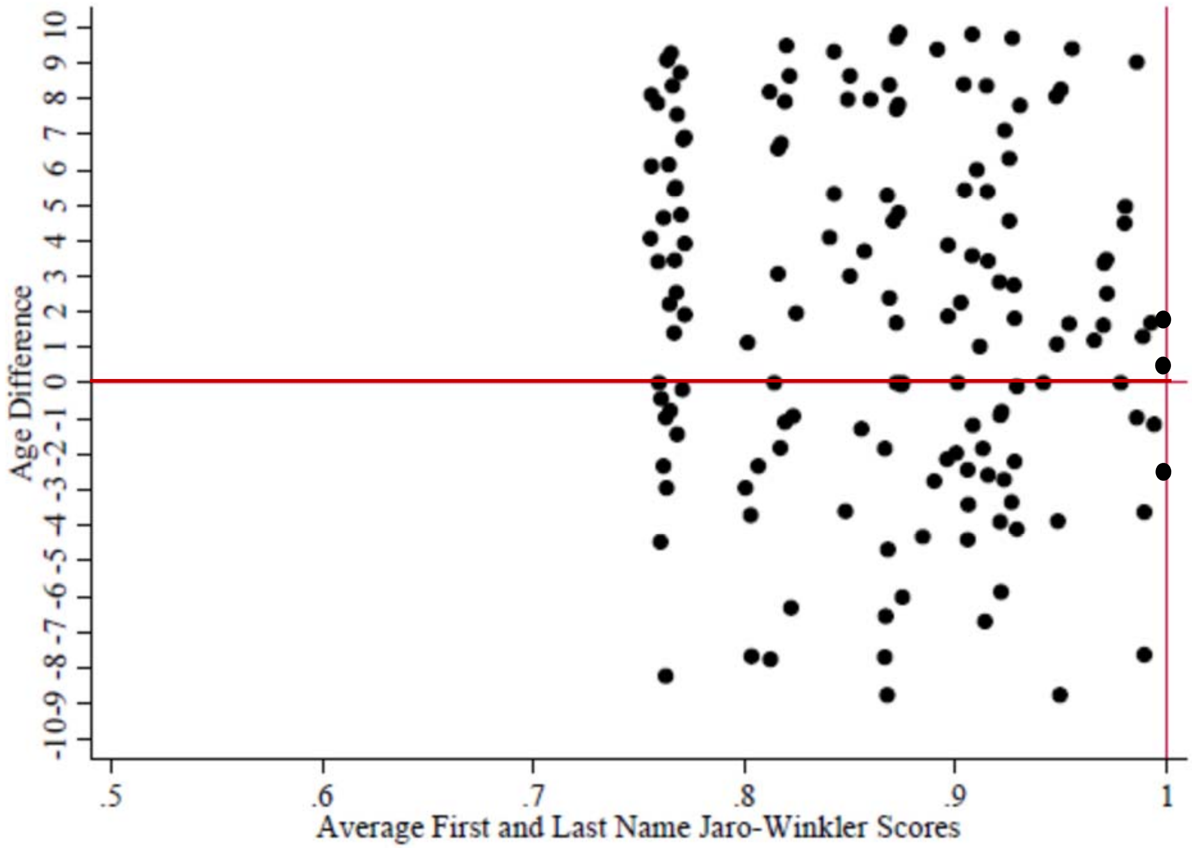
Soundex and NYSIIS

- may increase the number match candidates
- may worsen name matching
- may increase problems with match ties

Abramitzky, Boustan, and Erickson (2012)

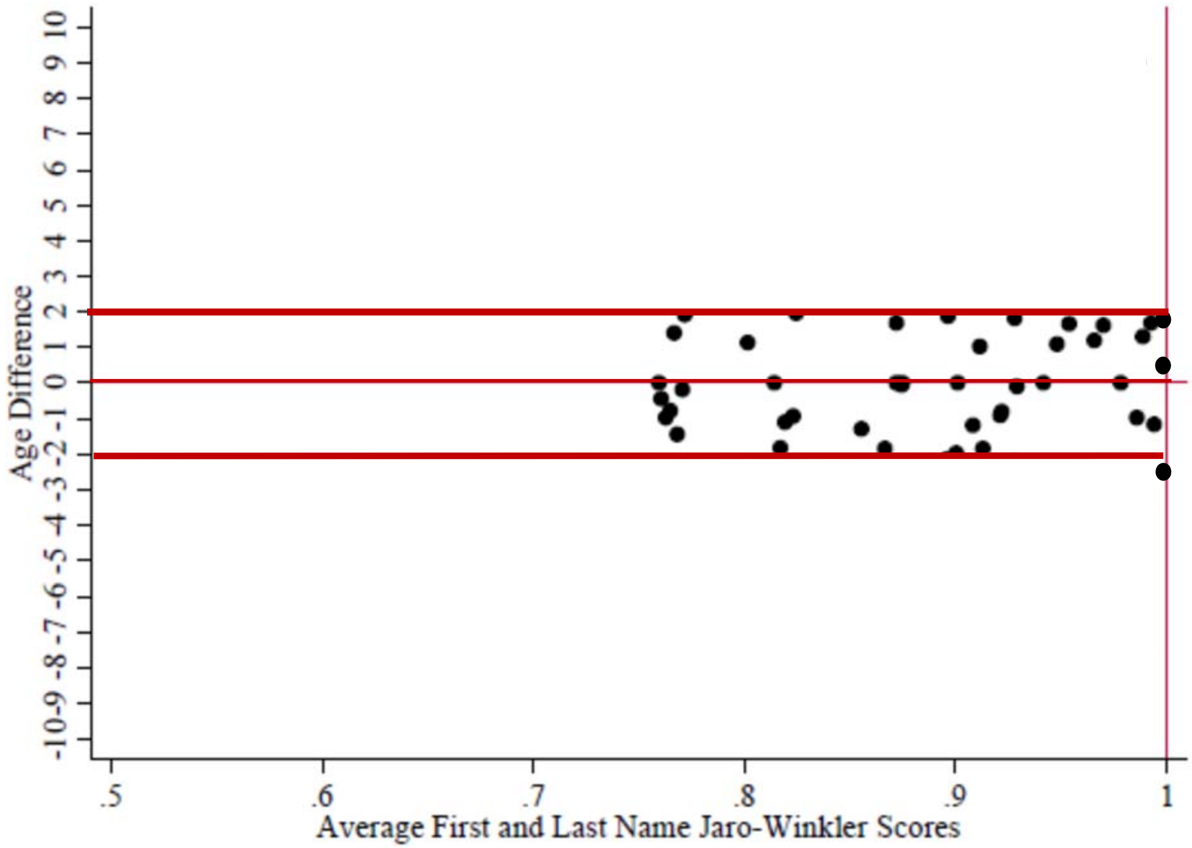


Abramitzky, Boustan, and Erickson (2012)



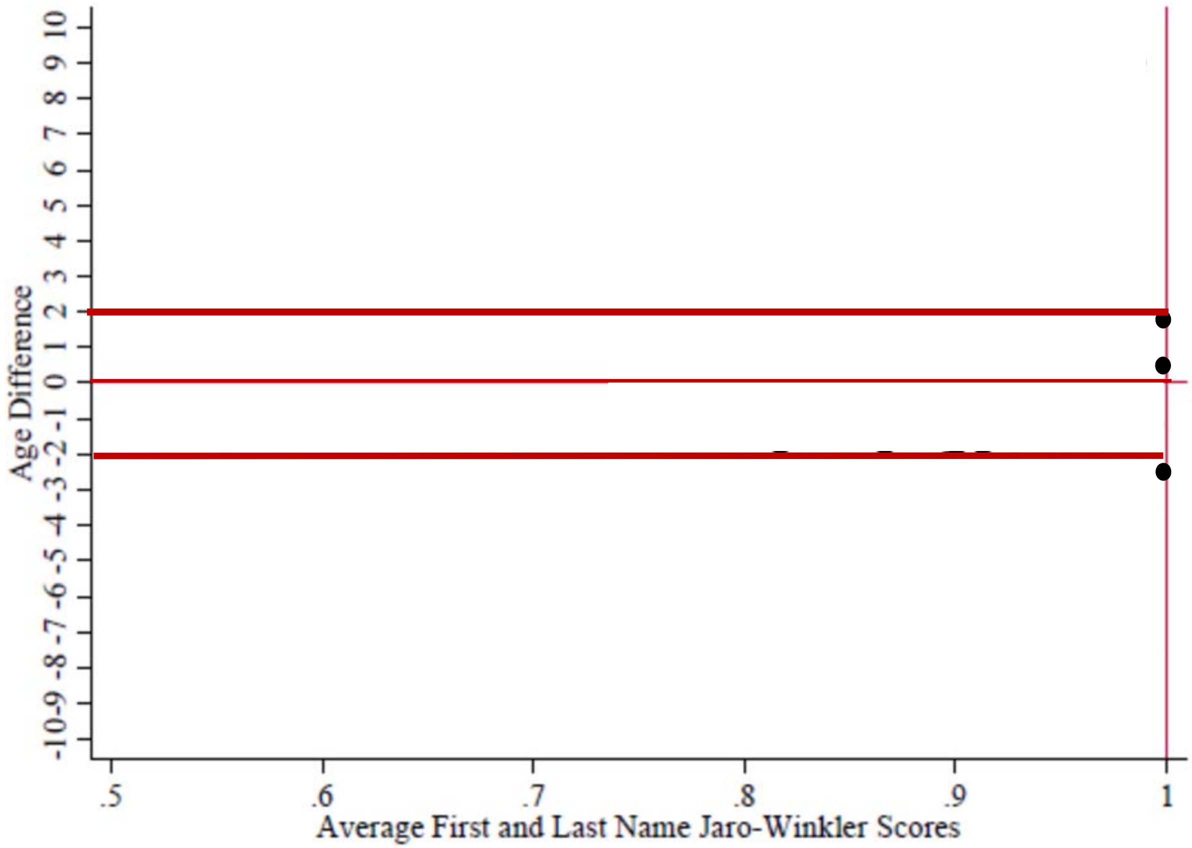
1. Keeps common names

Abramitzky, Boustan, and Erickson (2012)



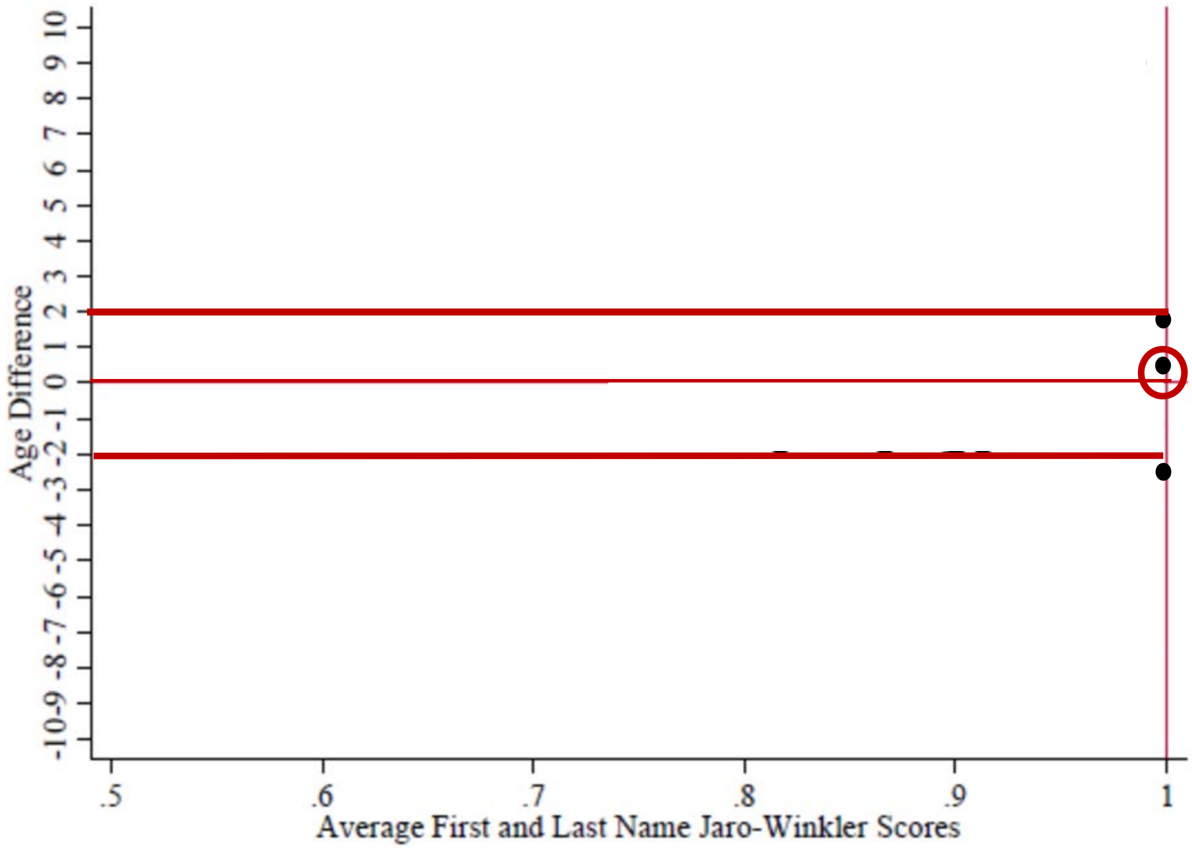
1. Keeps common names
2. Restricts age difference to be +/-2

Abramitzky, Boustan, and Erickson (2012)



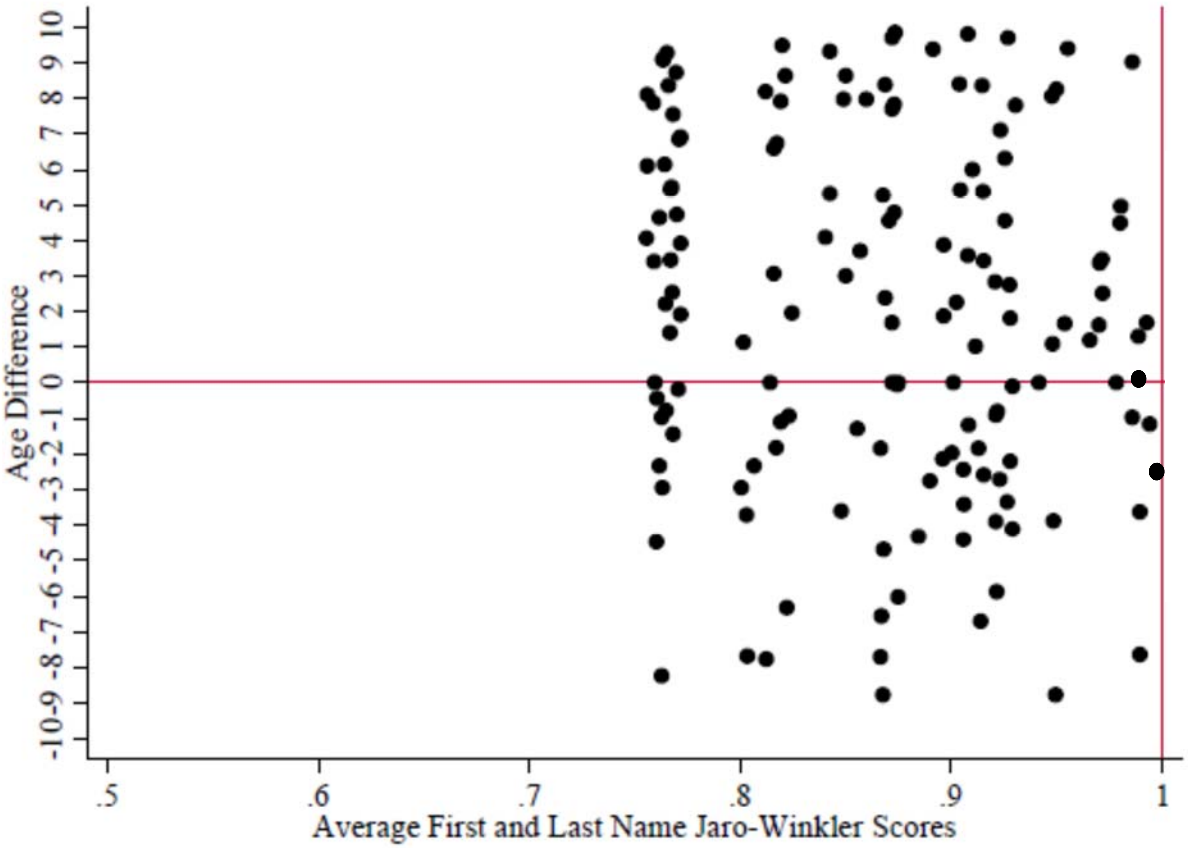
1. Keeps common names
2. Restricts age difference to be +/-2
3. Finds exact name matches (NYSIIS)

Abramitzky, Boustan, and Erickson (2012)



1. Keeps common names
2. Restricts age difference to be +/-2
3. Finds exact name matches (NYSIIS)
4. Searches iteratively +1, -1, +2, -2, etc. over age difference
5. No match chosen with multiples

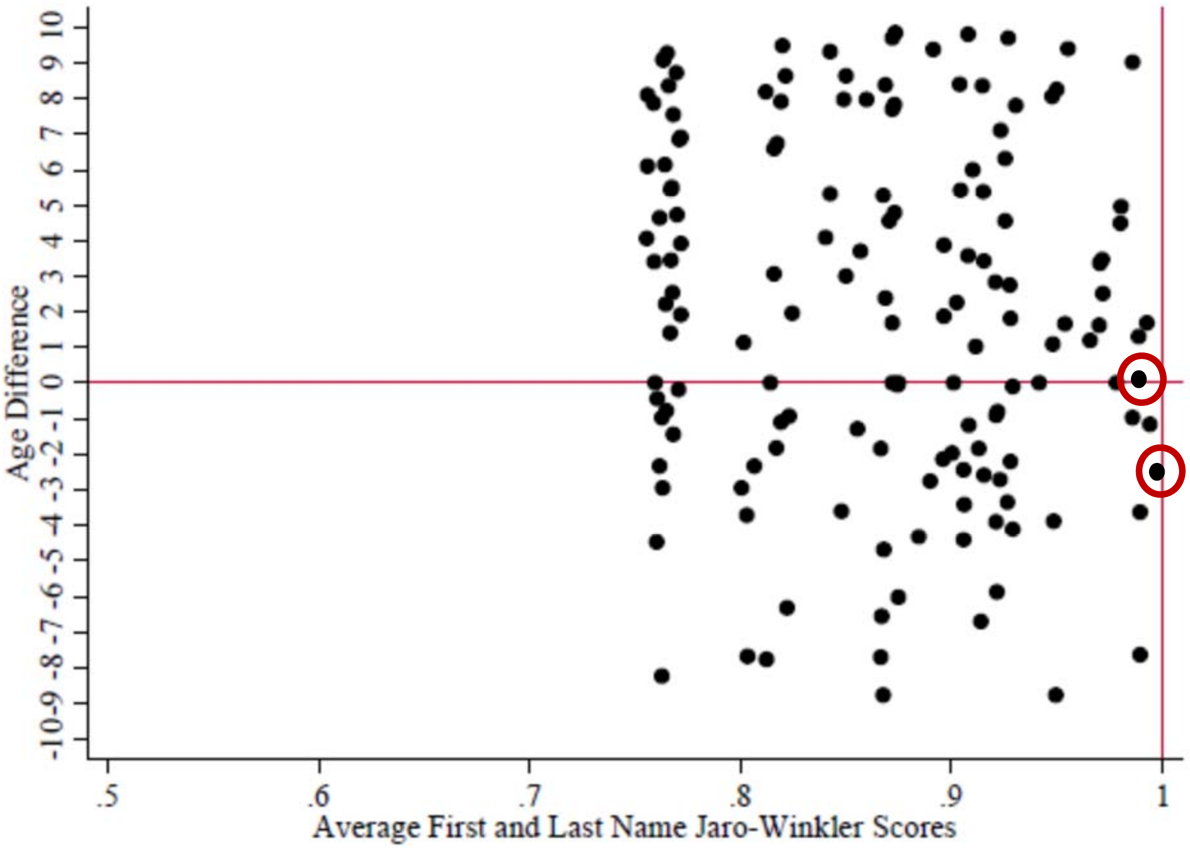
Trade-Offs: Age vs. Name Similarity?



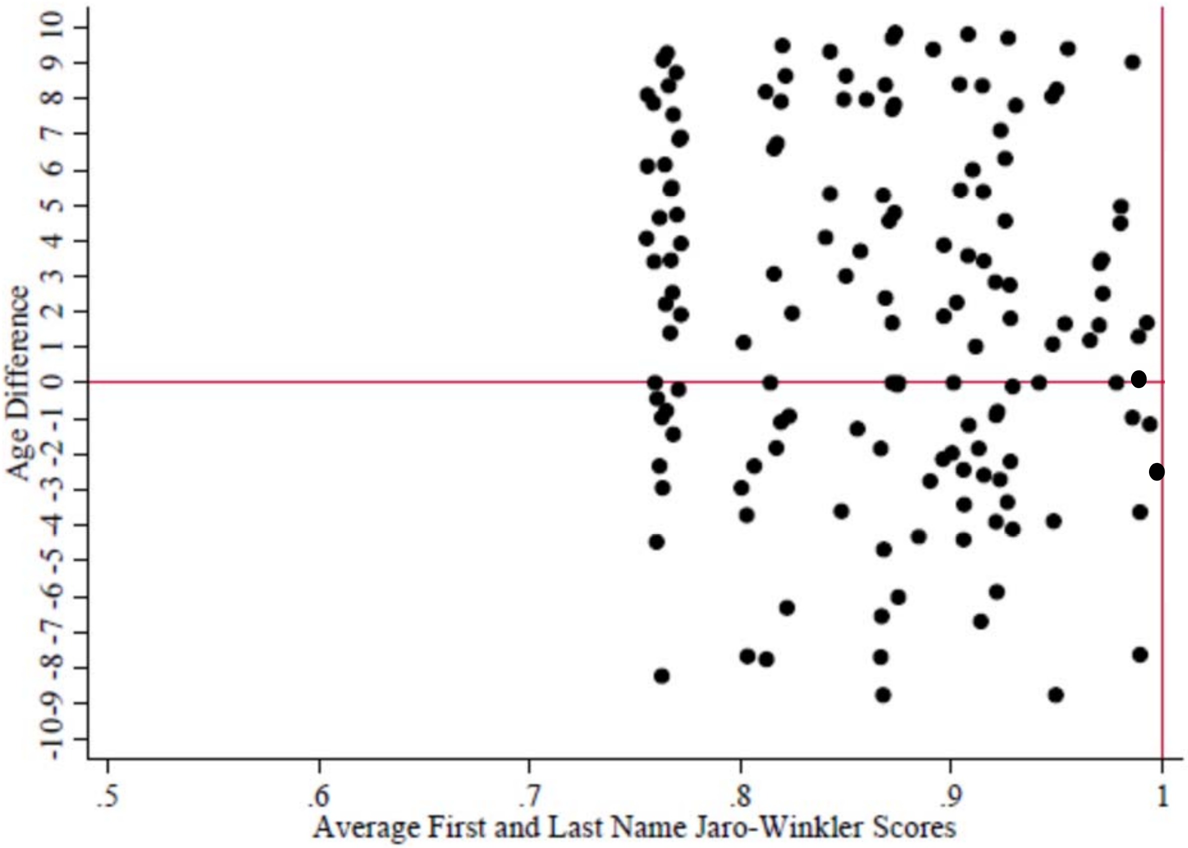
Trade-Offs: Age vs. Name Similarity?

Trade-Offs: Age vs. Name Similarity?

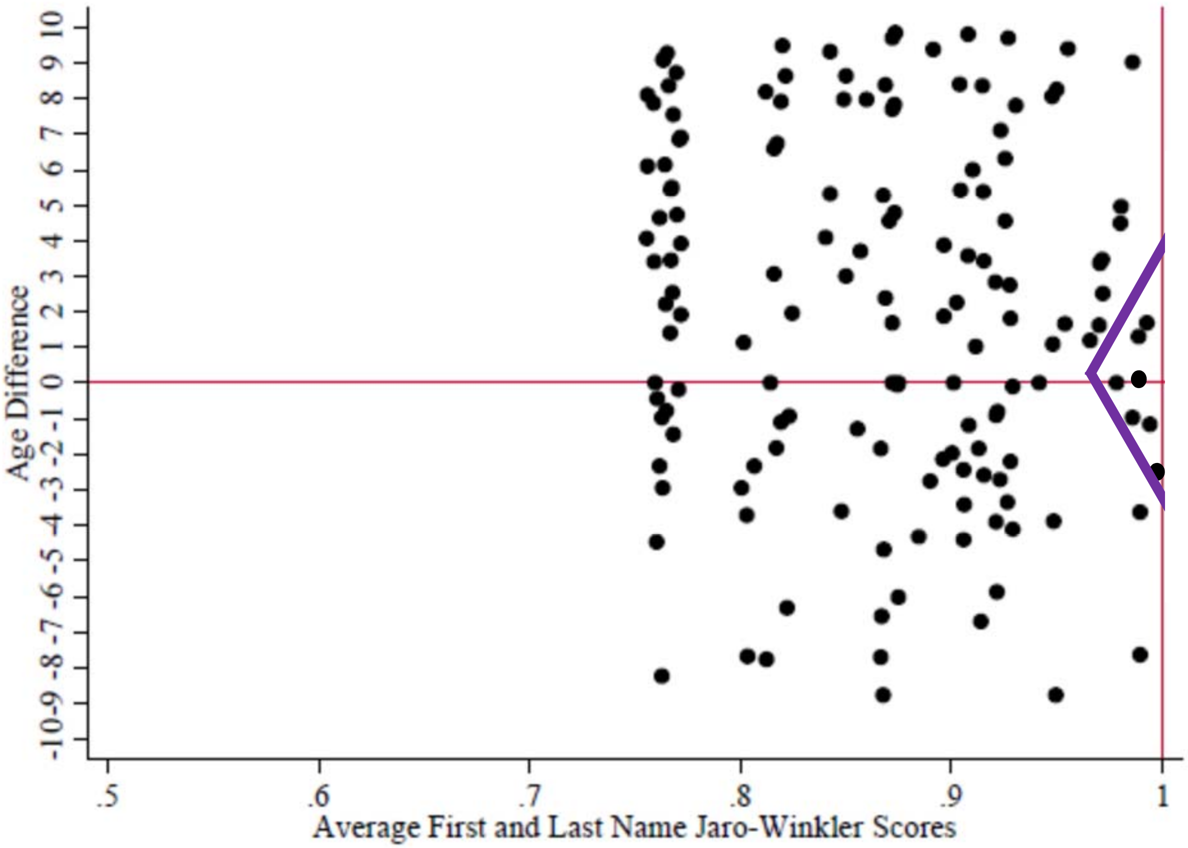
Trade-Offs: Age vs. Name Similarity?



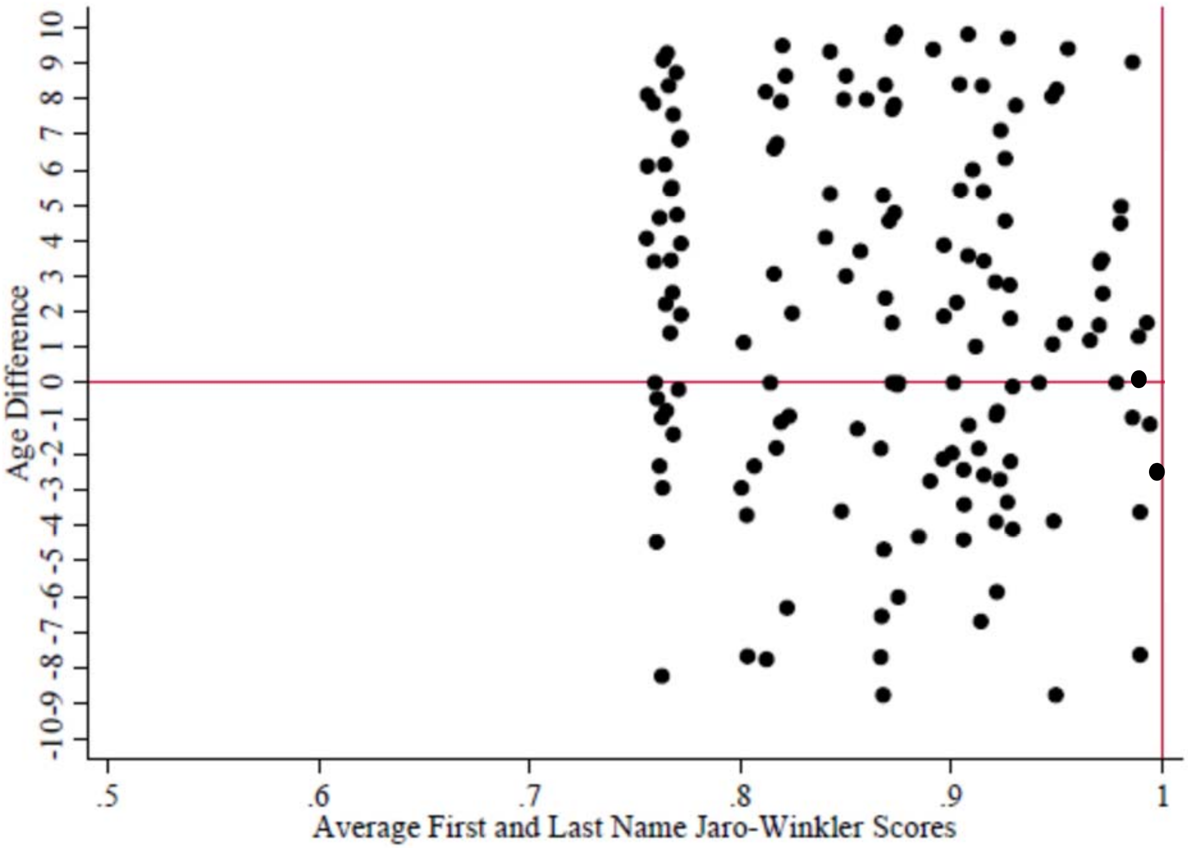
Trade-Offs: Age vs. Name Similarity?



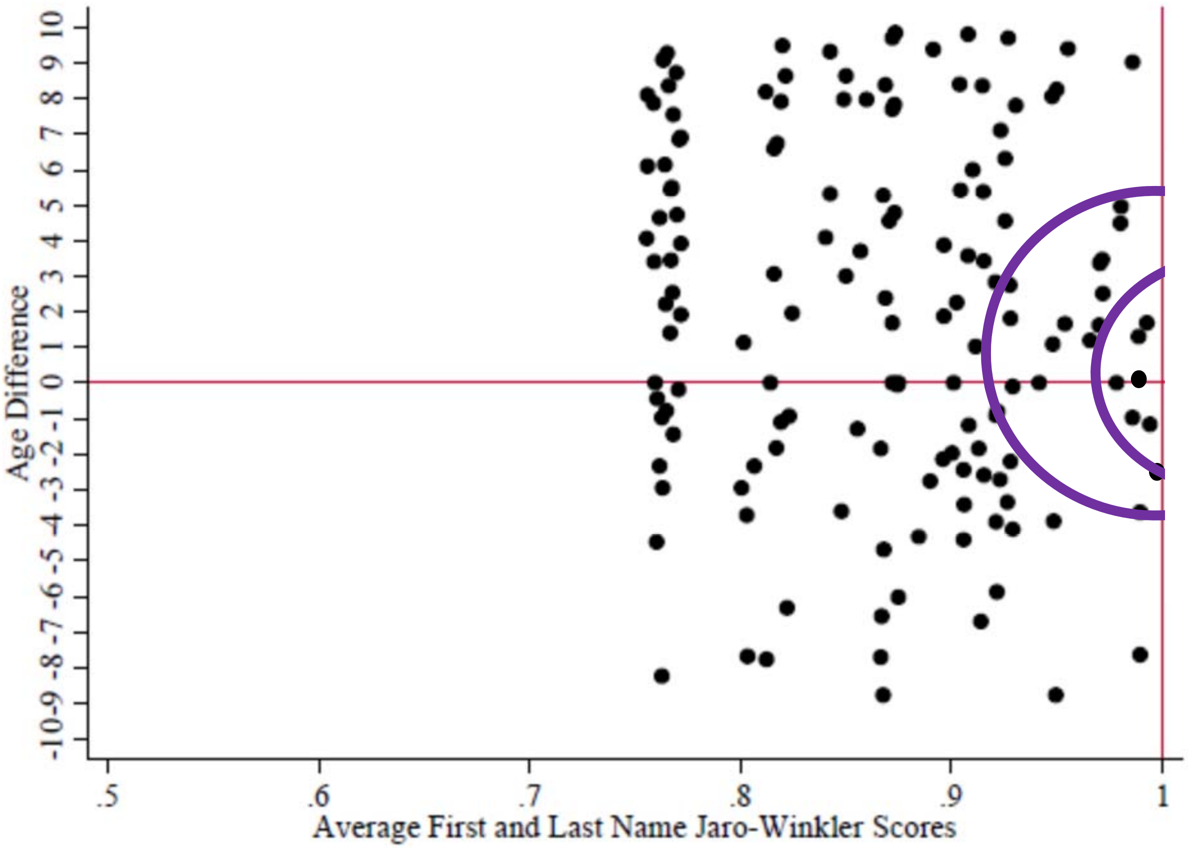
Trade-Offs: Age vs. Name Similarity?



Trade-Offs: Age vs. Name Similarity?



Trade-Offs: Age vs. Name Similarity?



Machine Learning

Key idea: use information in a “truth dataset” to “train” a model to classify links

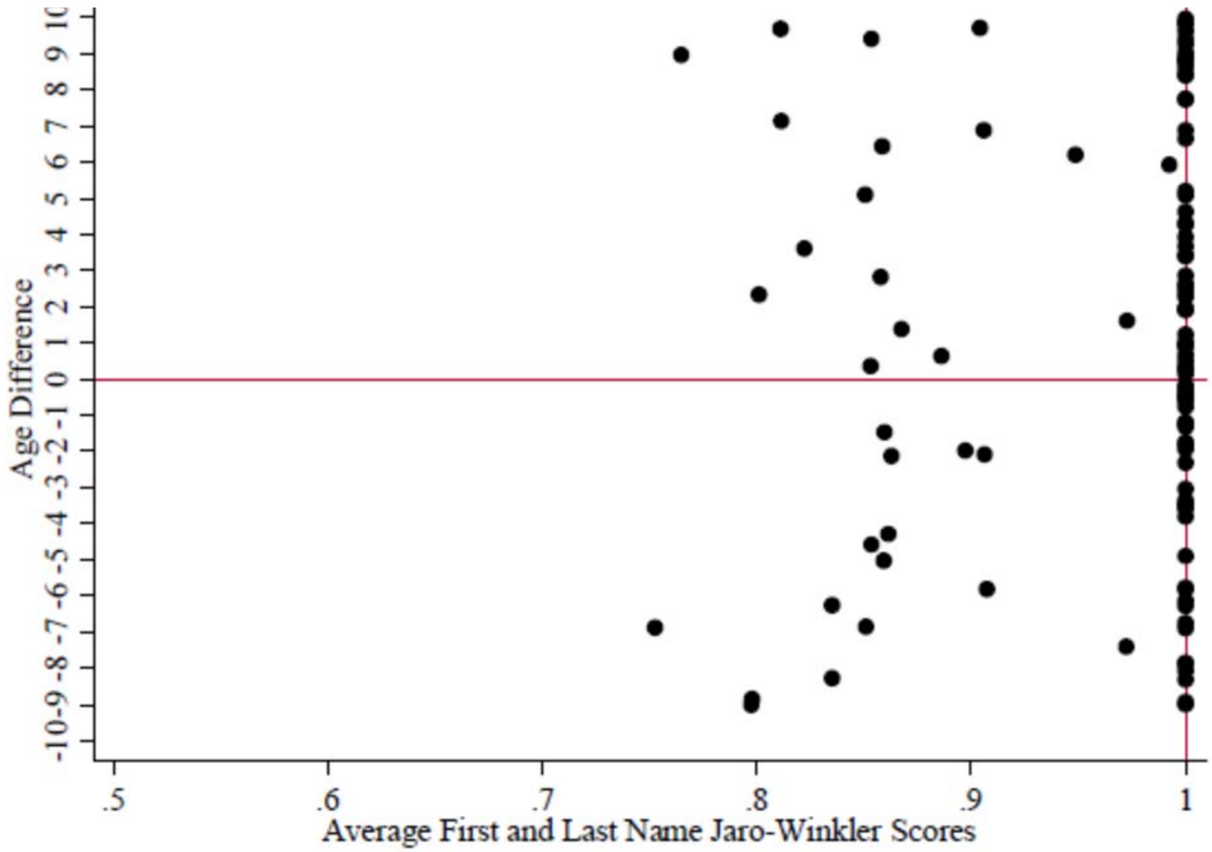
Machine Learning

Key idea: use information in a “truth dataset” to “train” a model to classify links

IPUMS Linked Historical samples uses a SVM to model trade-offs in multiple dimensions

Feigenbaum (2016) “regression-based” method to model trade-offs in multiple dimensions

Final Frontier: How to Choose Among Ties?



Choosing among (Exact) Ties

Exact ties in name-age: ~20 to 35 percent of U.S. samples; higher in some subsamples

Choosing among (Exact) Ties

Exact ties in name-age: ~20 to 35 percent of U.S. samples; higher in some subsamples

Statistics literature suggests probabilistic weighting:

- For exact ties, weight by p=1/m (where m=number of ties)

Choosing among (Exact) Ties

Exact ties in name-age: ~20 to 35 percent of U.S. samples; higher in some subsamples

Statistics literature suggests probabilistic weighting:

- For exact ties, weight by p=1/m (where m=number of ties)

Nix and Qian (2015) suggest random selection among ties

Choosing among (Exact) Ties

Exact ties in name-age: ~20 to 35 percent of U.S. samples; higher in some subsamples

Statistics literature suggests probabilistic weighting:

- For exact ties, weight by p=1/m (where m=number of ties)

Nix and Qian (2015) suggest random selection among ties

➔Assuming one of ties is correct, expected number of "wrong links" is the *same* for both methods

Method Performance

MATCH RATES AND REPRESENTATIVENESS

Method Performance

MATCH RATES AND REPRESENTATIVENESS

INCIDENCE OF TYPE I ERRORS

Method Performance

MATCH RATES AND REPRESENTATIVENESS

INCIDENCE OF TYPE I ERRORS

INCIDENCE OF TYPE II ERRORS

Ground Truth Samples

1. LIFE-M data
2. Synthetic data
3. Early Indicators data
4. IPUMS Historical Linked Censuses, 1850-1900

LIFE-M :
NC & Ohio Boys linked to 1940 Census

Births Records

1. Random samples of NC and Ohio birth certificates from 1909-20
2. Add in all siblings

N=45,442

1940 Census

birth place*, age*,
name*,
education, wages

LIFE-M Linking Process

1. Every link reviewed by two independent “data trainers”
2. Agreements assumed to be correct

LIFE-M Linking Process

1. Every link reviewed by two independent “data trainers”
2. Agreements assumed to be correct
3. Disagreements send records to *re*-review by an additional three individuals to resolve these discrepancies

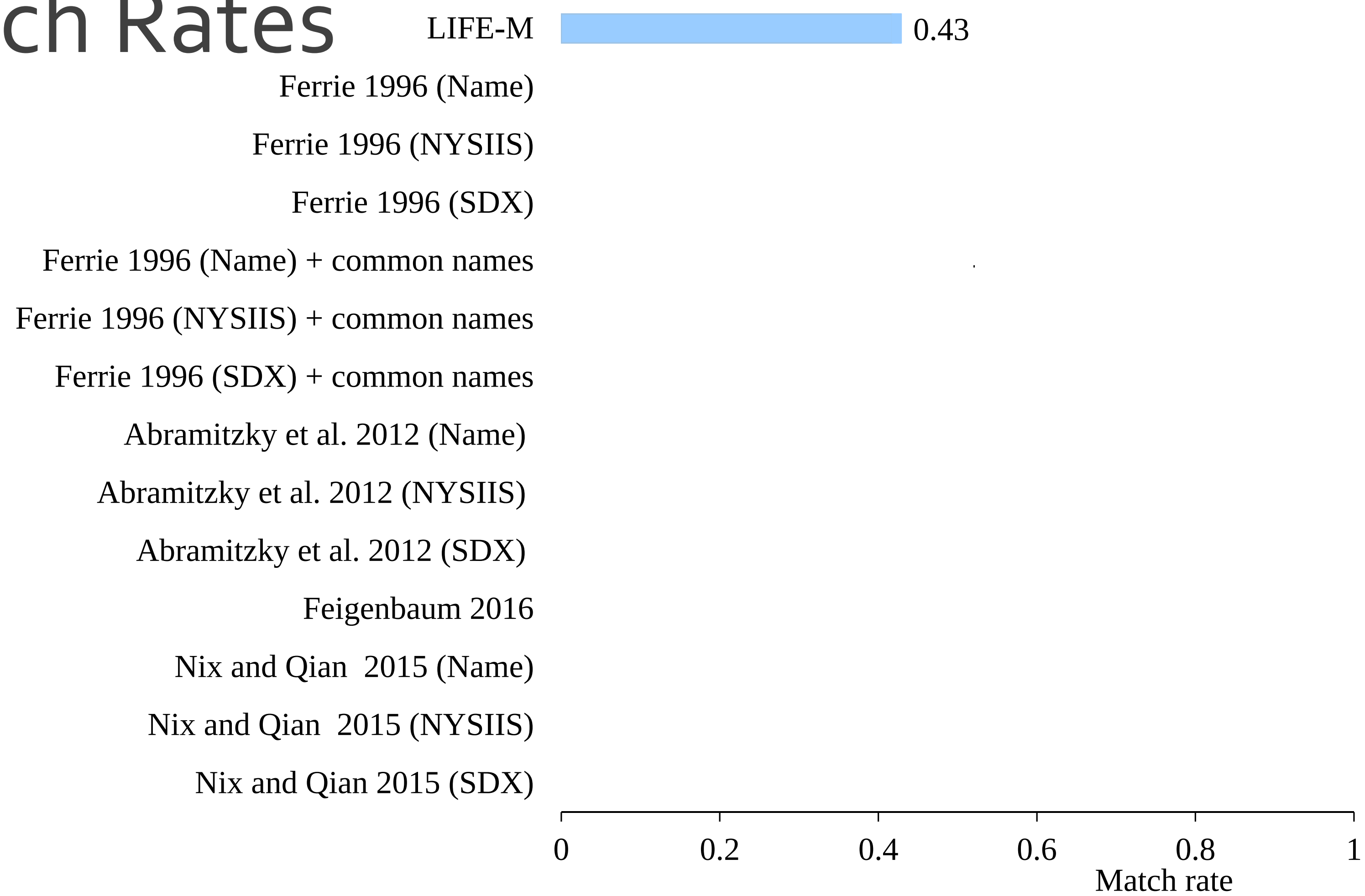
LIFE-M Linking Process

1. Every link reviewed by two independent “data trainers”
2. Agreements assumed to be correct
3. Disagreements send records to *re*-review by an additional three individuals to resolve these discrepancies
4. “Audit batches” and weekly meetings help maintain quality

LIFE-M Linking Process

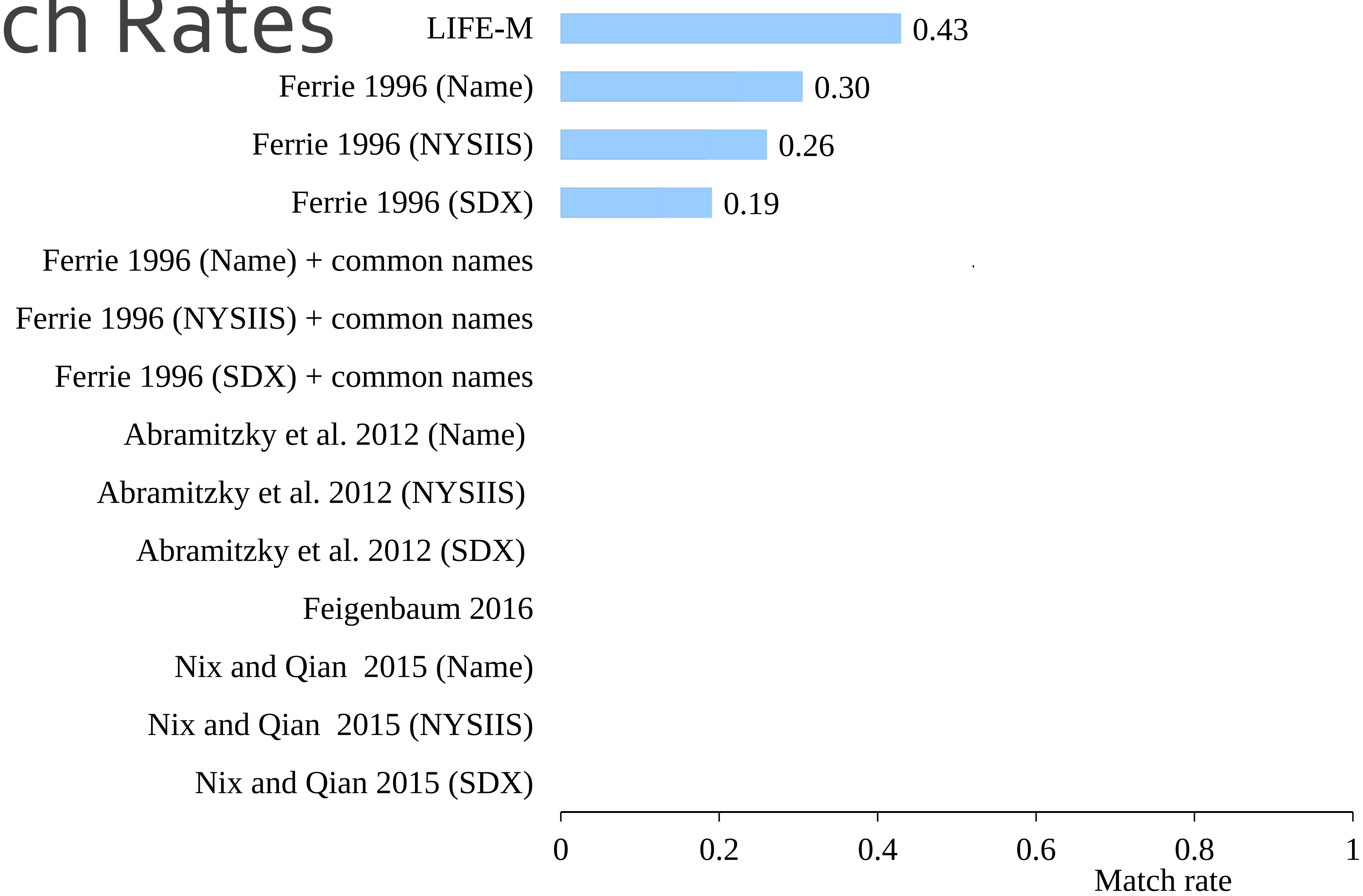
1. Every link reviewed by two independent "data trainers"
2. Agreements assumed to be correct
3. Disagreements send records to *re*-review by an additional three individuals to resolve these discrepancies
4. "Audit batches" and weekly meetings help maintain quality
5. Independent validation of links by BYU agrees 96% of the time

Match Rates

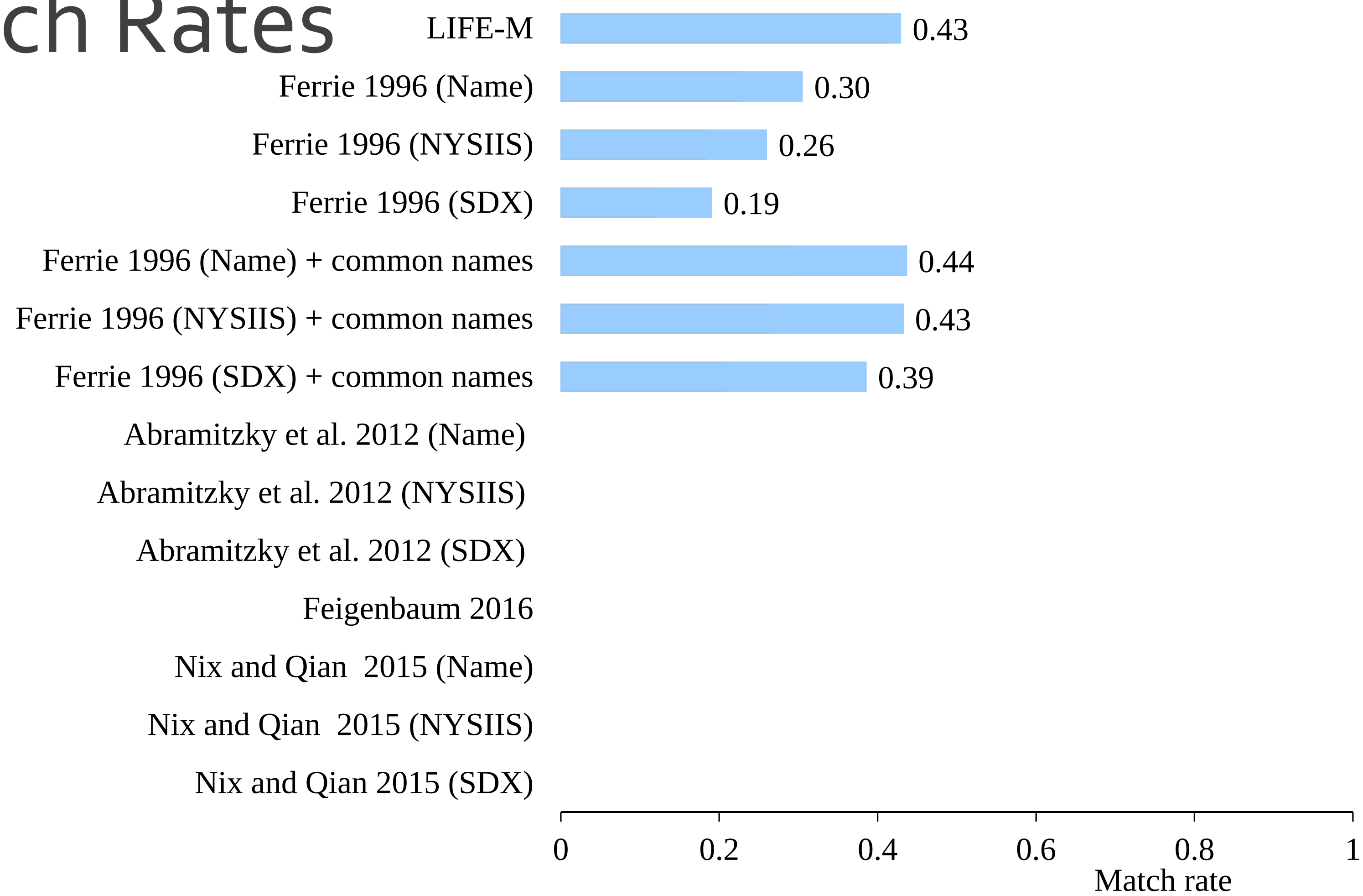


LIFE-M
0.43
Ferrie 1996 (Name)
Ferrie 1996 (NYSIIS)
Ferrie 1996 (SDX)
Ferrie 1996 (Name) + common names
Ferrie 1996 (NYSIIS) + common names
Ferrie 1996 (SDX) + common names
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0
0.2
0.4
0.6
0.8
1
Match rate

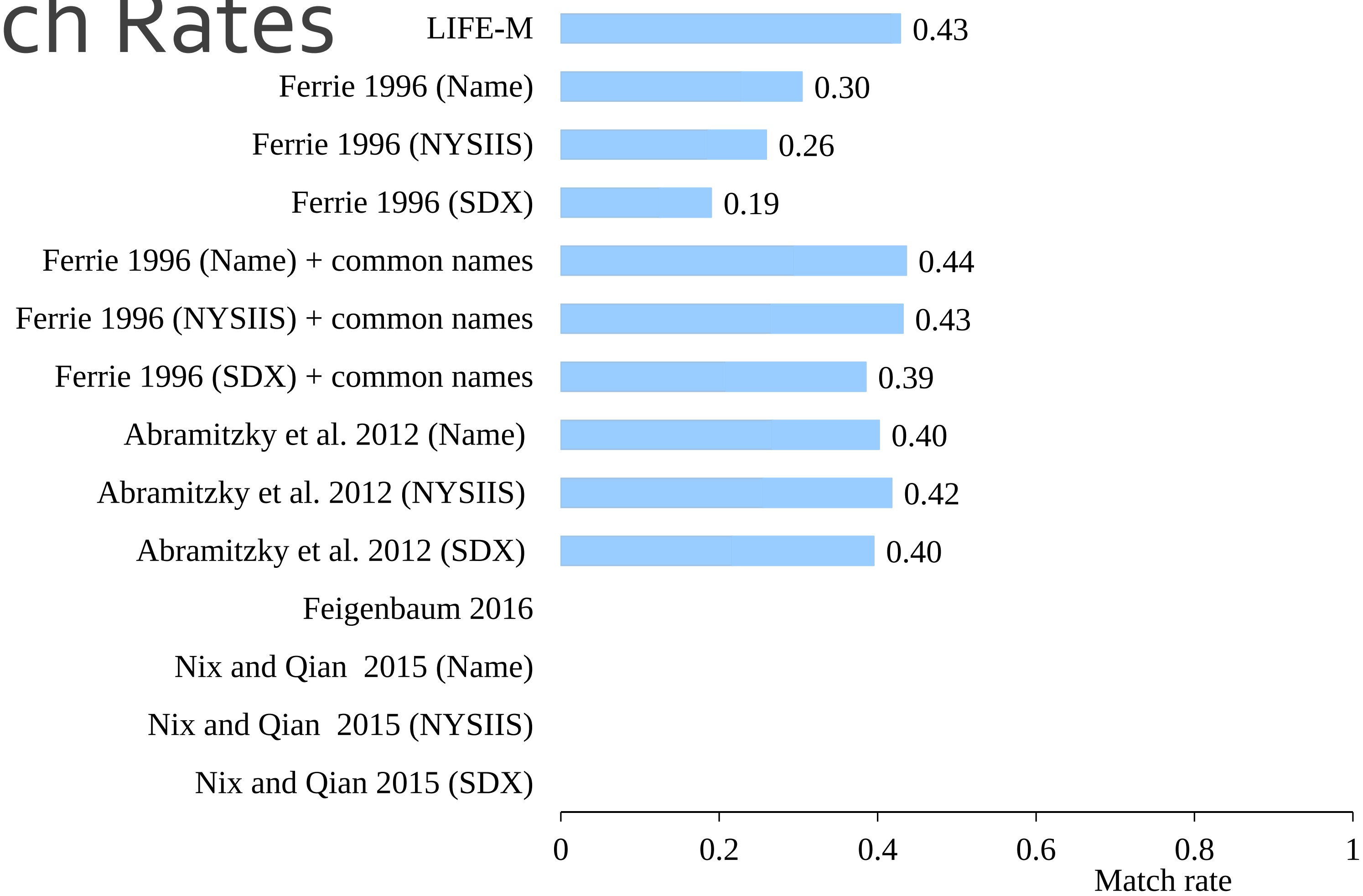
Match Rates


LIFE-M
0.43
Ferrie 1996 (Name)
0.30
Ferrie 1996 (NYSIIS)
0.26
Ferrie 1996 (SDX)
0.19
Ferrie 1996 (Name) + common names
Ferrie 1996 (NYSIIS) + common names
Ferrie 1996 (SDX) + common names
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0
0.2
0.4
0.6
0.8
1
Match rate

Match Rates

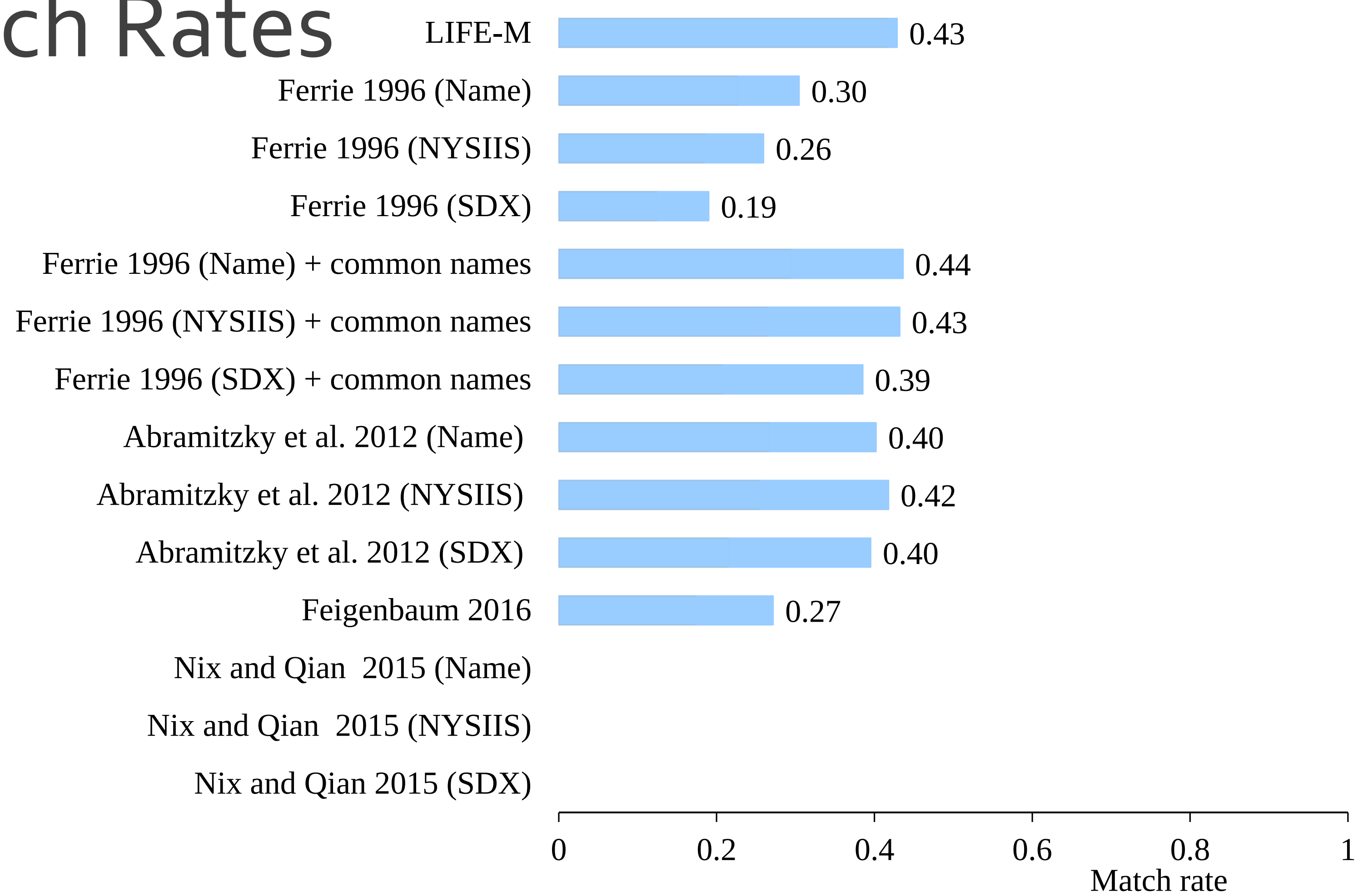

LIFE-M
0.43
Ferrie 1996 (Name)
0.30
Ferrie 1996 (NYSIIS)
0.26
Ferrie 1996 (SDX)
0.19
Ferrie 1996 (Name) + common names
0.44
Ferrie 1996 (NYSIIS) + common names
0.43
Ferrie 1996 (SDX) + common names
0.39
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0
0.2
0.4
0.6
0.8
1
Match rate

Match Rates



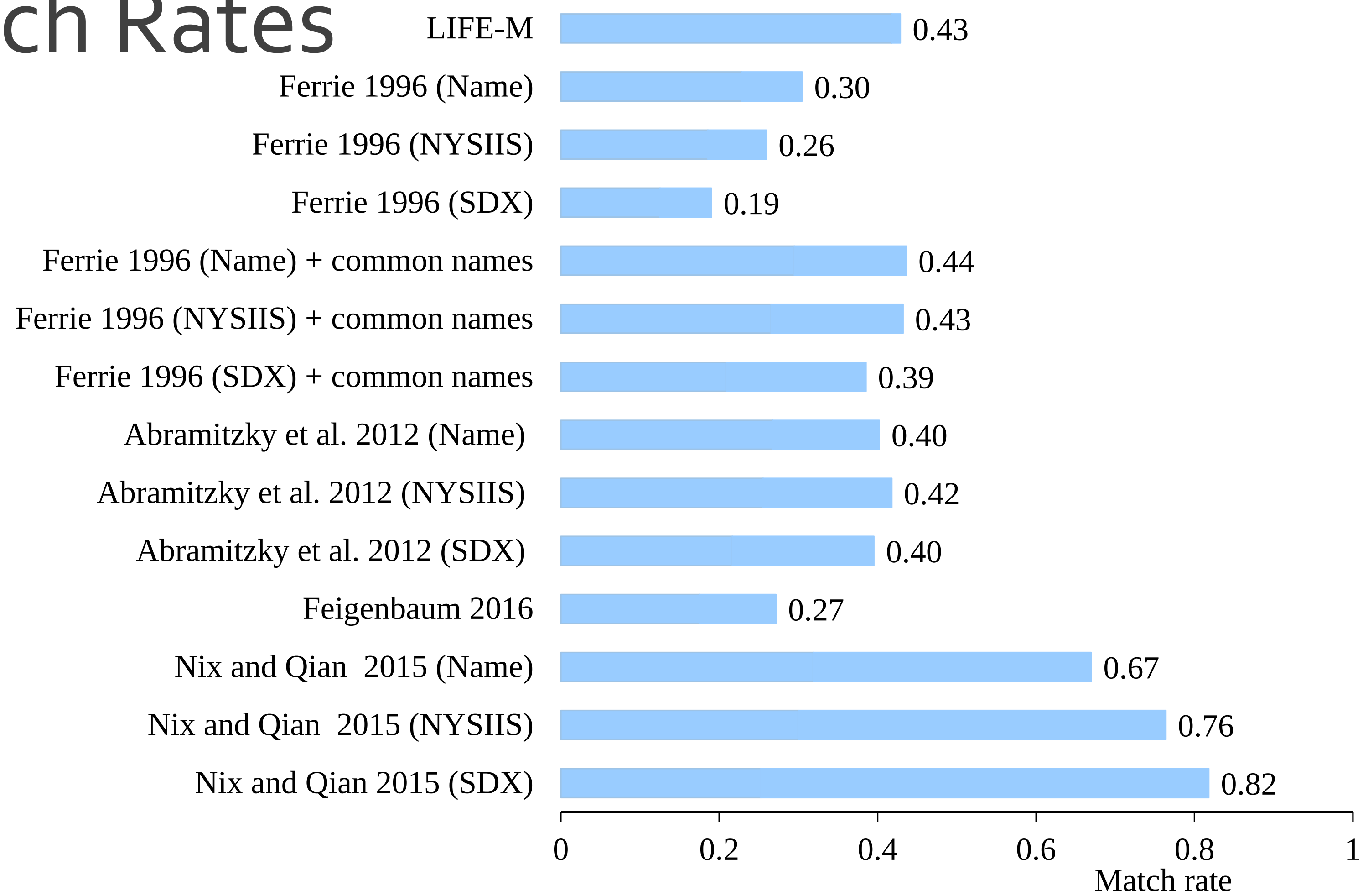
LIFE-M
0.43
Ferrie 1996 (Name)
0.30
Ferrie 1996 (NYSIIS)
0.26
Ferrie 1996 (SDX)
0.19
Ferrie 1996 (Name) + common names
0.44
Ferrie 1996 (NYSIIS) + common names
0.43
Ferrie 1996 (SDX) + common names
0.39
Abramitzky et al. 2012 (Name)
0.40
Abramitzky et al. 2012 (NYSIIS)
0.42
Abramitzky et al. 2012 (SDX)
0.40
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0
0.2
0.4
0.6
0.8
1
Match rate

Match Rates



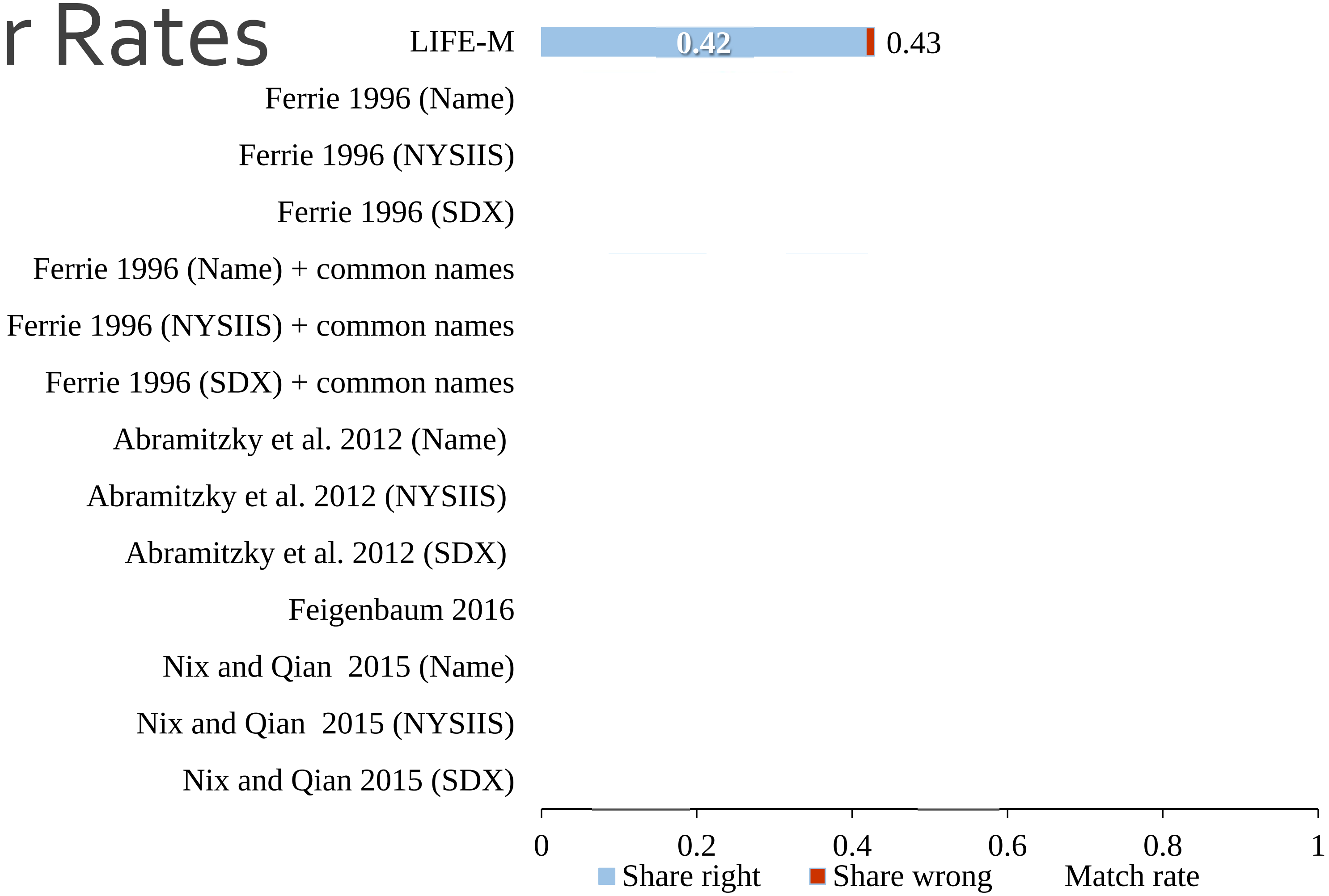
LIFE-M
Ferrie 1996 (Name)
Ferrie 1996 (NYSIIS)
Ferrie 1996 (SDX)
Ferrie 1996 (Name) + common names
Ferrie 1996 (NYSIIS) + common names
Ferrie 1996 (SDX) + common names
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0.43
0.30
0.26
0.19
0.44
0.43
0.39
0.40
0.42
0.40
0.27
0
0.2
0.4
0.6
0.8
1
Match rate

Match Rates

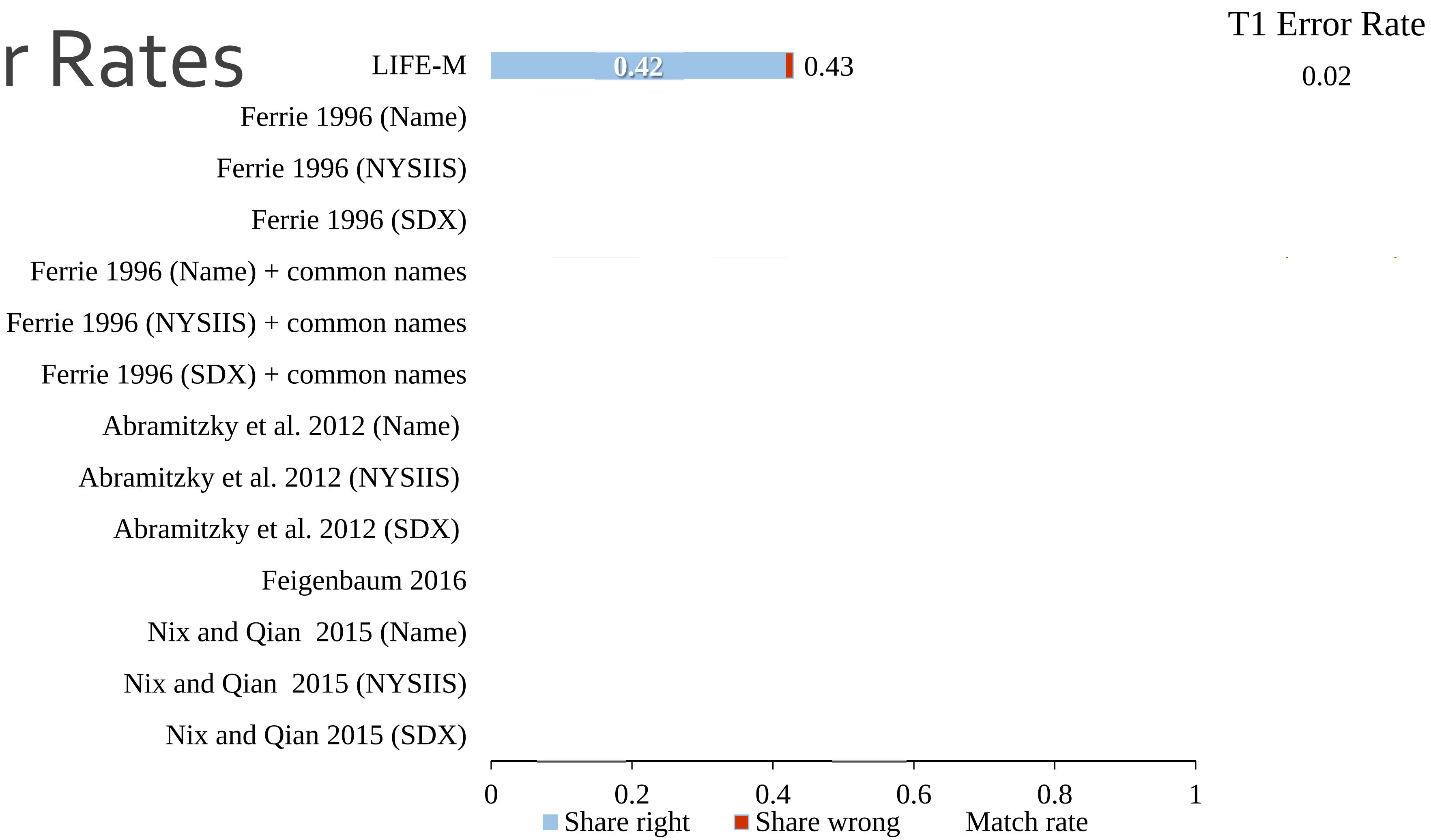


LIFE-M 0.43
Ferrie 1996 (Name) 0.30
Ferrie 1996 (NYSIIS) 0.26
Ferrie 1996 (SDX) 0.19
Ferrie 1996 (Name) + common names 0.44
Ferrie 1996 (NYSIIS) + common names 0.43
Ferrie 1996 (SDX) + common names 0.39
Abramitzky et al. 2012 (Name) 0.40
Abramitzky et al. 2012 (NYSIIS) 0.42
Abramitzky et al. 2012 (SDX) 0.40
Feigenbaum 2016 0.27
Nix and Qian 2015 (Name) 0.67
Nix and Qian 2015 (NYSIIS) 0.76
Nix and Qian 2015 (SDX) 0.82
0
0.2
0.4
0.6
0.8
1
Match rate

Error Rates

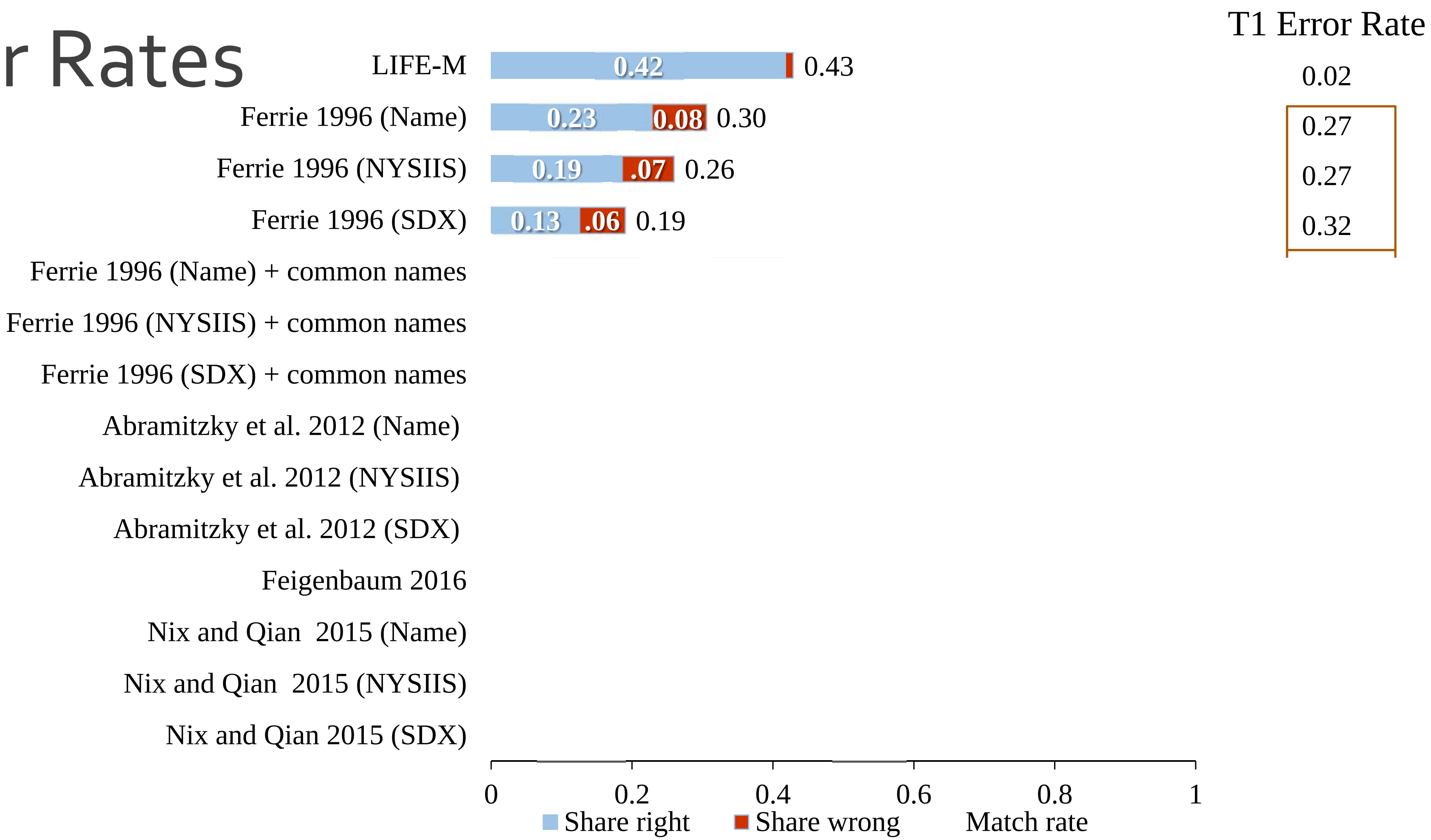

LIFE-M
0.42
0.43
Ferrie 1996 (Name)
Ferrie 1996 (NYSIIS)
Ferrie 1996 (SDX)
Ferrie 1996 (Name) + common names
Ferrie 1996 (NYSIIS) + common names
Ferrie 1996 (SDX) + common names
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0
0.2
0.4
0.6
0.8
1
Share right
Share wrong
Match rate

Error Rates

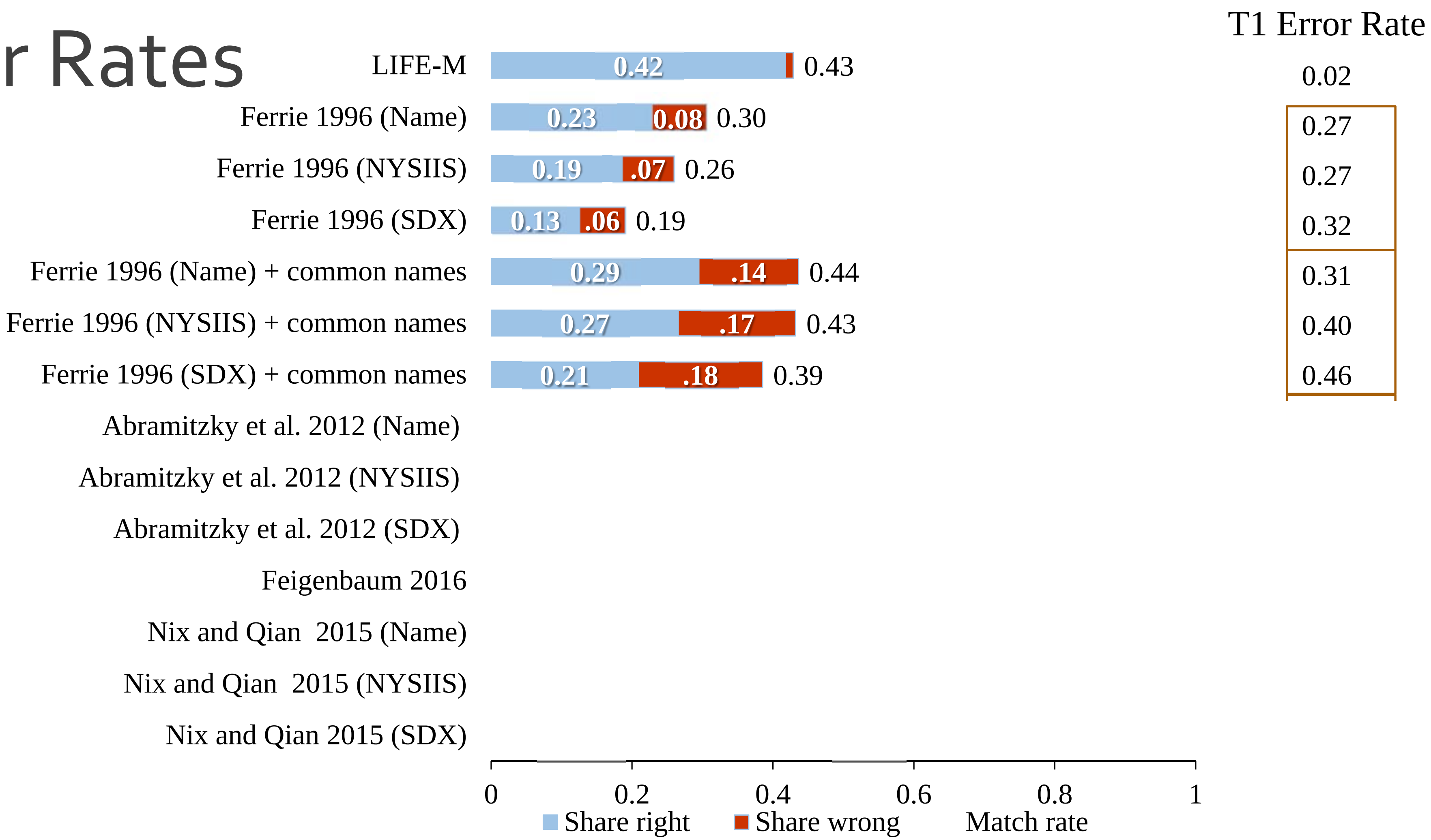


T1 Error Rate
LIFE-M
0.42
0.43
0.02
Ferrie 1996 (Name)
Ferrie 1996 (NYSIIS)
Ferrie 1996 (SDX)
Ferrie 1996 (Name) + common names
Ferrie 1996 (NYSIIS) + common names
Ferrie 1996 (SDX) + common names
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0
0.2
0.4
0.6
0.8
1
Share right
Share wrong
Match rate

Error Rates

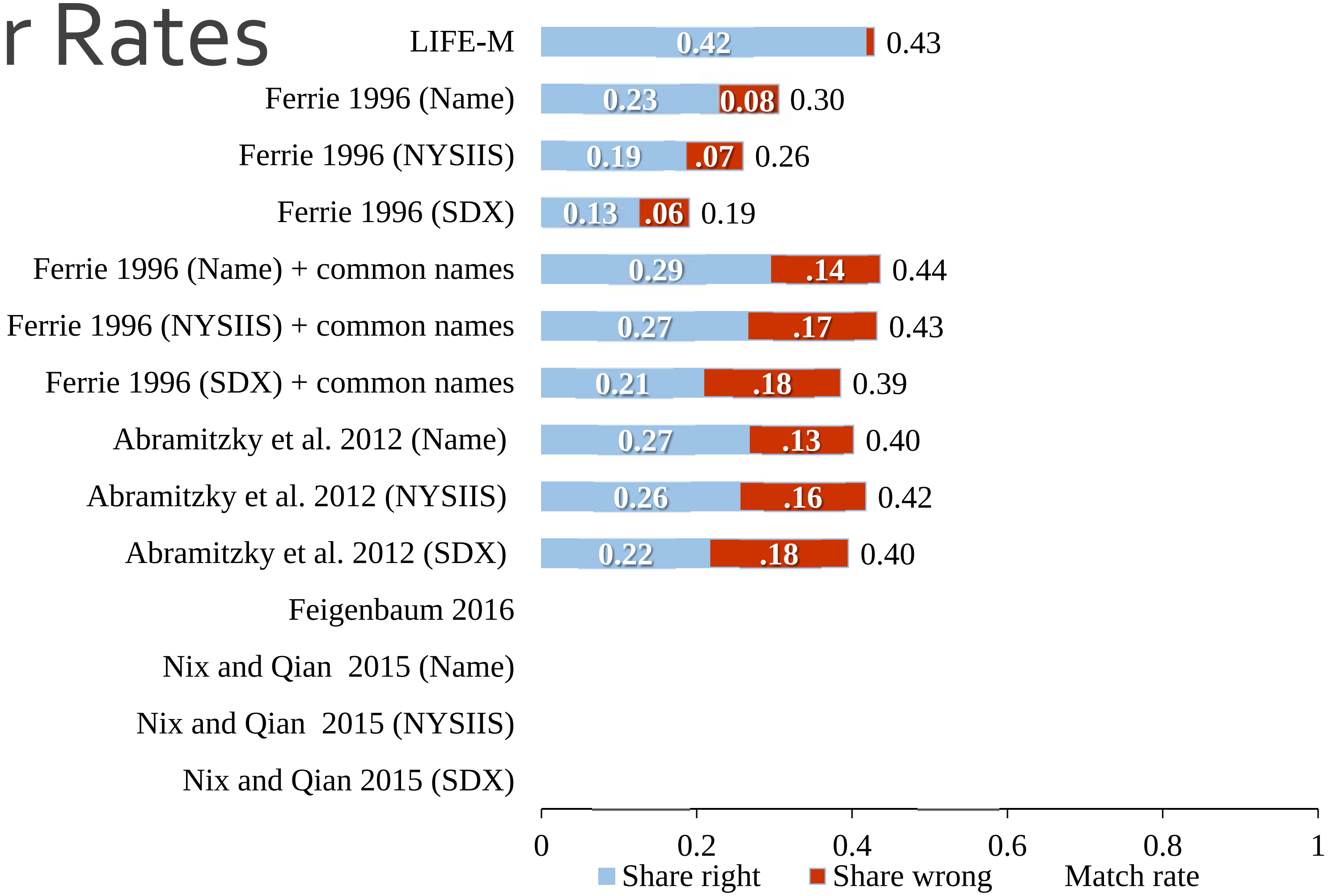

LIFE-M
Ferrie 1996 (Name)
Ferrie 1996 (NYSIIS)
Ferrie 1996 (SDX)
Ferrie 1996 (Name) + common names
Ferrie 1996 (NYSIIS) + common names
Ferrie 1996 (SDX) + common names
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0.42
0.43
0.23
0.08
0.30
0.19
.07
0.26
0.13
.06
0.19
0
0.2
0.4
0.6
0.8
1
Share right
Share wrong
Match rate
T1 Error Rate
0.02
0.27
0.27
0.32

Error Rates



T1 Error Rate
LIFE-M
Ferrie 1996 (Name)
Ferrie 1996 (NYSIIS)
Ferrie 1996 (SDX)
Ferrie 1996 (Name) + common names
Ferrie 1996 (NYSIIS) + common names
Ferrie 1996 (SDX) + common names
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0.42
0.43
0.23
0.08
0.30
0.19
.07
0.26
0.13
.06
0.19
0.29
.14
0.44
0.27
.17
0.43
0.21
.18
0.39
0.02
0.27
0.27
0.32
0.31
0.40
0.46
0
0.2
0.4
0.6
0.8
1
Share right
Share wrong
Match rate

Error Rates



T1 Error Rate

0.02

0.27

0.27

0.32

0.31

0.40

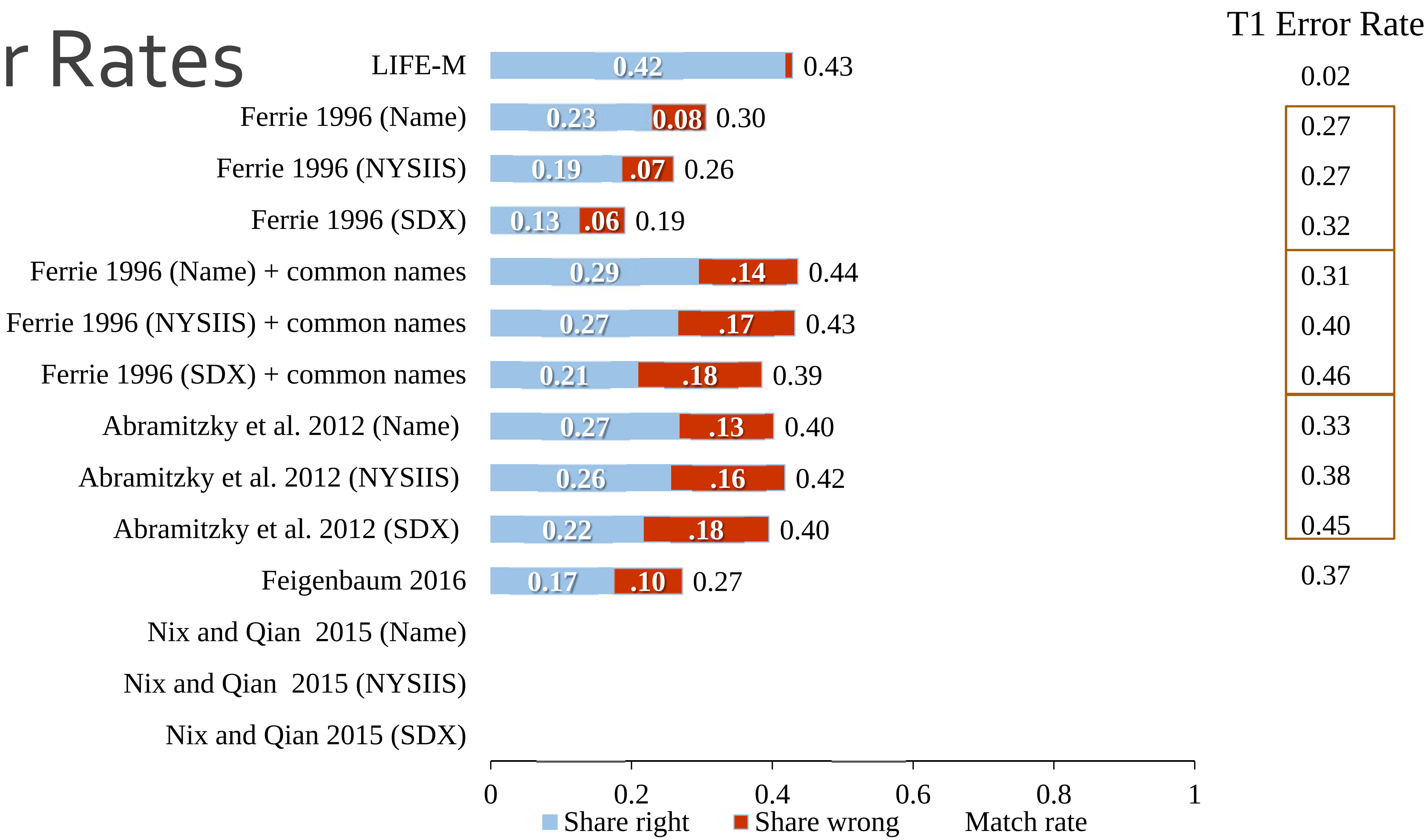
0.46

0.33

0.38

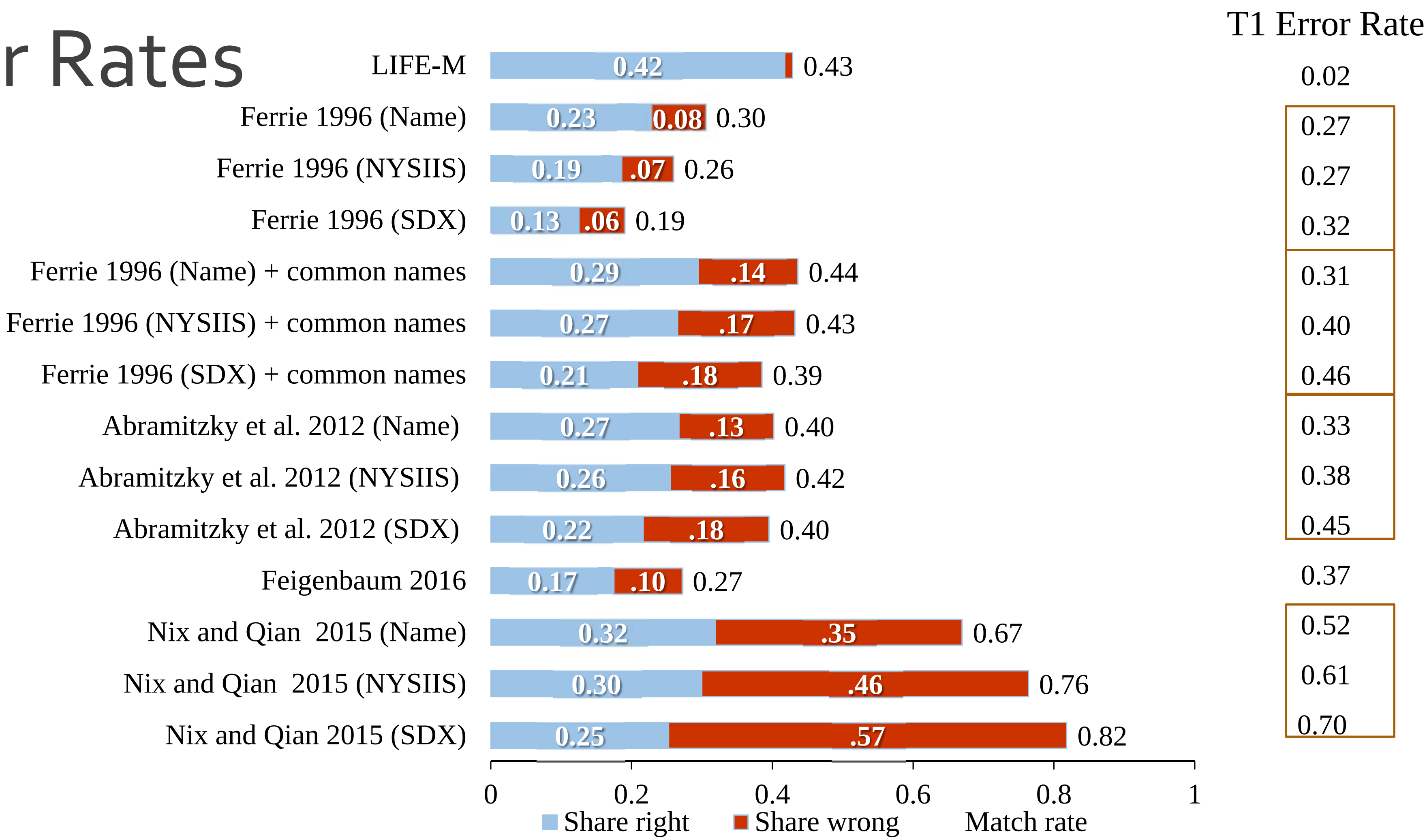
0.45

Error Rates

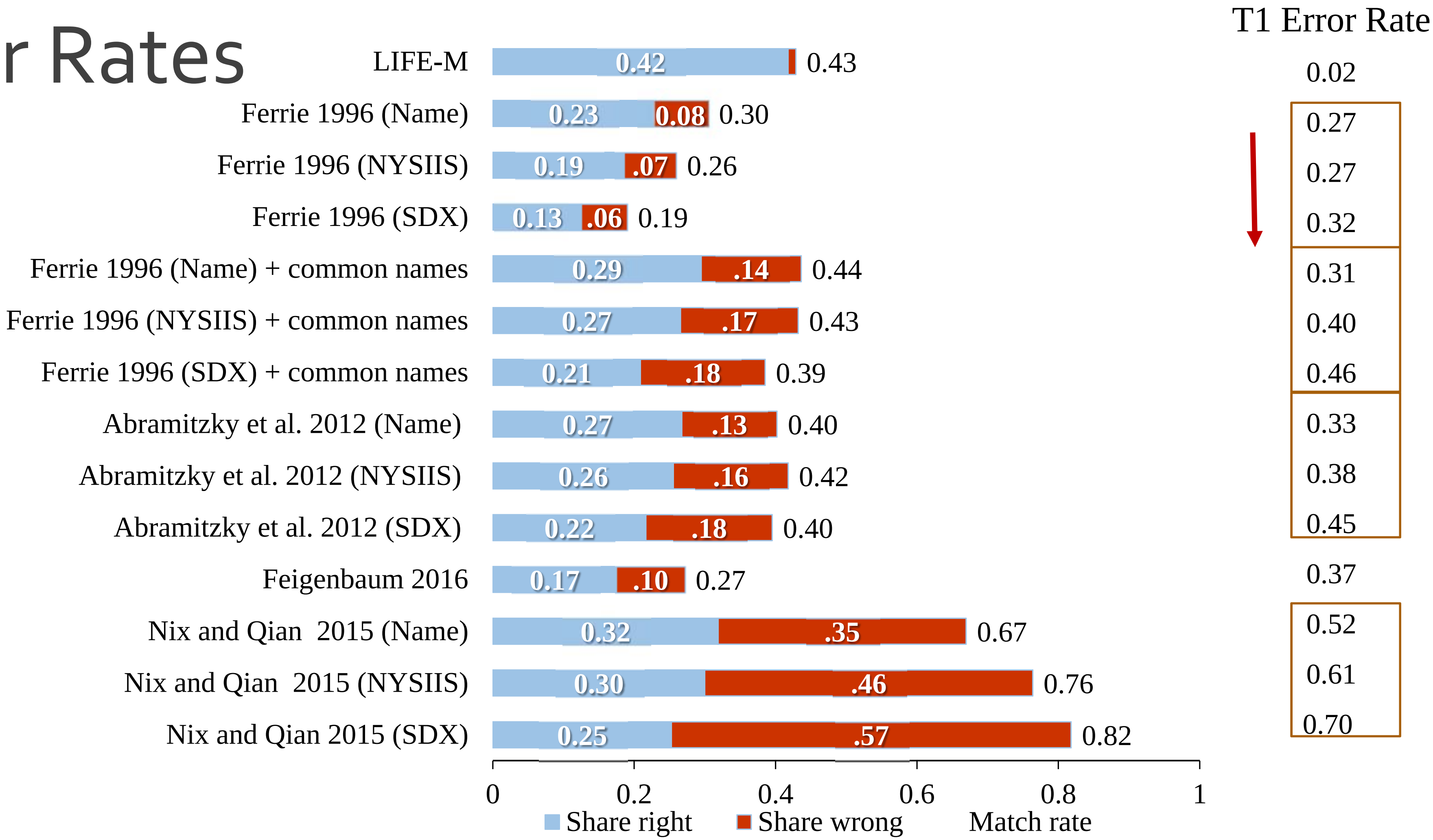


T1 Error Rate
LIFE-M 0.42 0.43 0.02
Ferrie 1996 (Name) 0.23 0.08 0.30 0.27
Ferrie 1996 (NYSIIS) 0.19 .07 0.26 0.27
Ferrie 1996 (SDX) 0.13 .06 0.19 0.32
Ferrie 1996 (Name) + common names 0.29 .14 0.44 0.31
Ferrie 1996 (NYSIIS) + common names 0.27 .17 0.43 0.40
Ferrie 1996 (SDX) + common names 0.21 .18 0.39 0.46
Abramitzky et al. 2012 (Name) 0.27 .13 0.40 0.33
Abramitzky et al. 2012 (NYSIIS) 0.26 .16 0.42 0.38
Abramitzky et al. 2012 (SDX) 0.22 .18 0.40 0.45
Feigenbaum 2016 0.17 .10 0.27 0.37
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0
0.2
0.4
0.6
0.8
1
Share right
Share wrong
Match rate

Error Rates

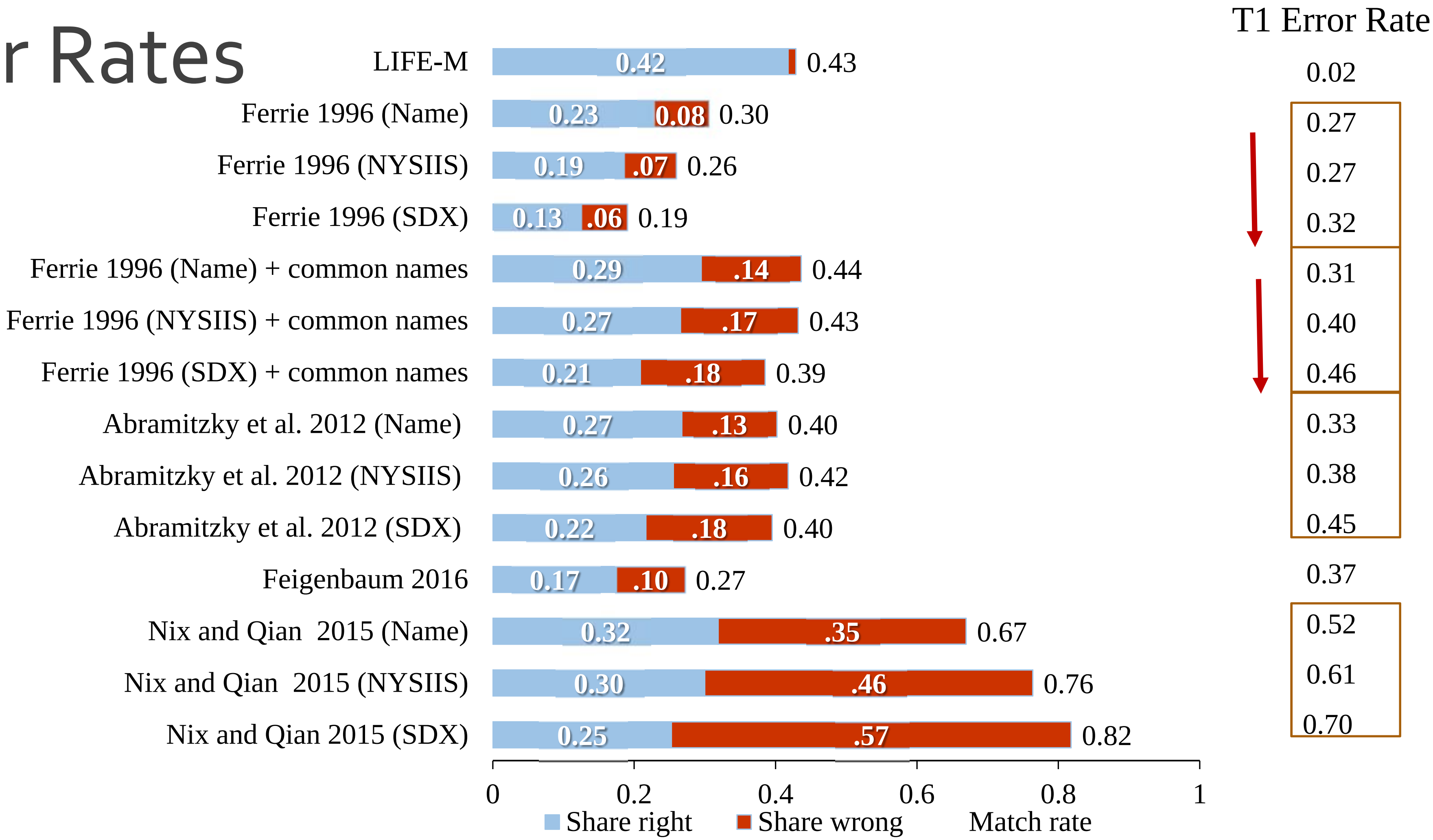

T1 Error Rate
LIFE-M 0.42 0.43 0.02
Ferrie 1996 (Name) 0.23 0.08 0.30 0.27
Ferrie 1996 (NYSIIS) 0.19 .07 0.26 0.27
Ferrie 1996 (SDX) 0.13 .06 0.19 0.32
Ferrie 1996 (Name) + common names 0.29 .14 0.44 0.31
Ferrie 1996 (NYSIIS) + common names 0.27 .17 0.43 0.40
Ferrie 1996 (SDX) + common names 0.21 .18 0.39 0.46
Abramitzky et al. 2012 (Name) 0.27 .13 0.40 0.33
Abramitzky et al. 2012 (NYSIIS) 0.26 .16 0.42 0.38
Abramitzky et al. 2012 (SDX) 0.22 .18 0.40 0.45
Feigenbaum 2016 0.17 .10 0.27 0.37
Nix and Qian 2015 (Name) 0.32 .35 0.67 0.52
Nix and Qian 2015 (NYSIIS) 0.30 .46 0.76 0.61
Nix and Qian 2015 (SDX) 0.25 .57 0.82 0.70
0
0.2
0.4
0.6
0.8
1
Share right
Share wrong
Match rate

Error Rates



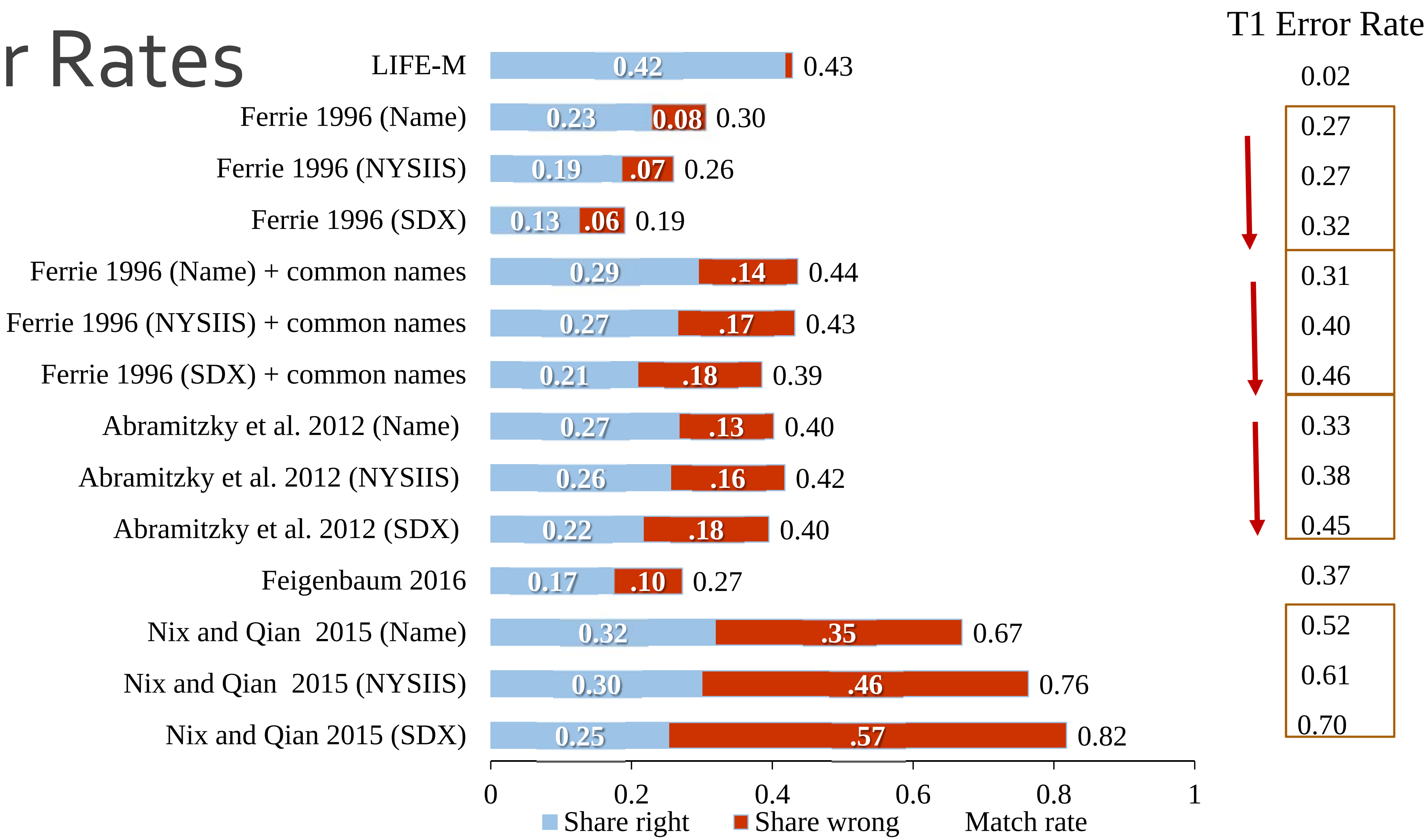
T1 Error Rate
LIFE-M 0.42 0.43 0.02
Ferrie 1996 (Name) 0.23 0.08 0.30 0.27
Ferrie 1996 (NYSIIS) 0.19 .07 0.26 0.27
Ferrie 1996 (SDX) 0.13 .06 0.19 0.32
Ferrie 1996 (Name) + common names 0.29 .14 0.44 0.31
Ferrie 1996 (NYSIIS) + common names 0.27 .17 0.43 0.40
Ferrie 1996 (SDX) + common names 0.21 .18 0.39 0.46
Abramitzky et al. 2012 (Name) 0.27 .13 0.40 0.33
Abramitzky et al. 2012 (NYSIIS) 0.26 .16 0.42 0.38
Abramitzky et al. 2012 (SDX) 0.22 .18 0.40 0.45
Feigenbaum 2016 0.17 .10 0.27 0.37
Nix and Qian 2015 (Name) 0.32 .35 0.67 0.52
Nix and Qian 2015 (NYSIIS) 0.30 .46 0.76 0.61
Nix and Qian 2015 (SDX) 0.25 .57 0.82 0.70
0
0.2
0.4
0.6
0.8
1
Share right
Share wrong
Match rate

Error Rates

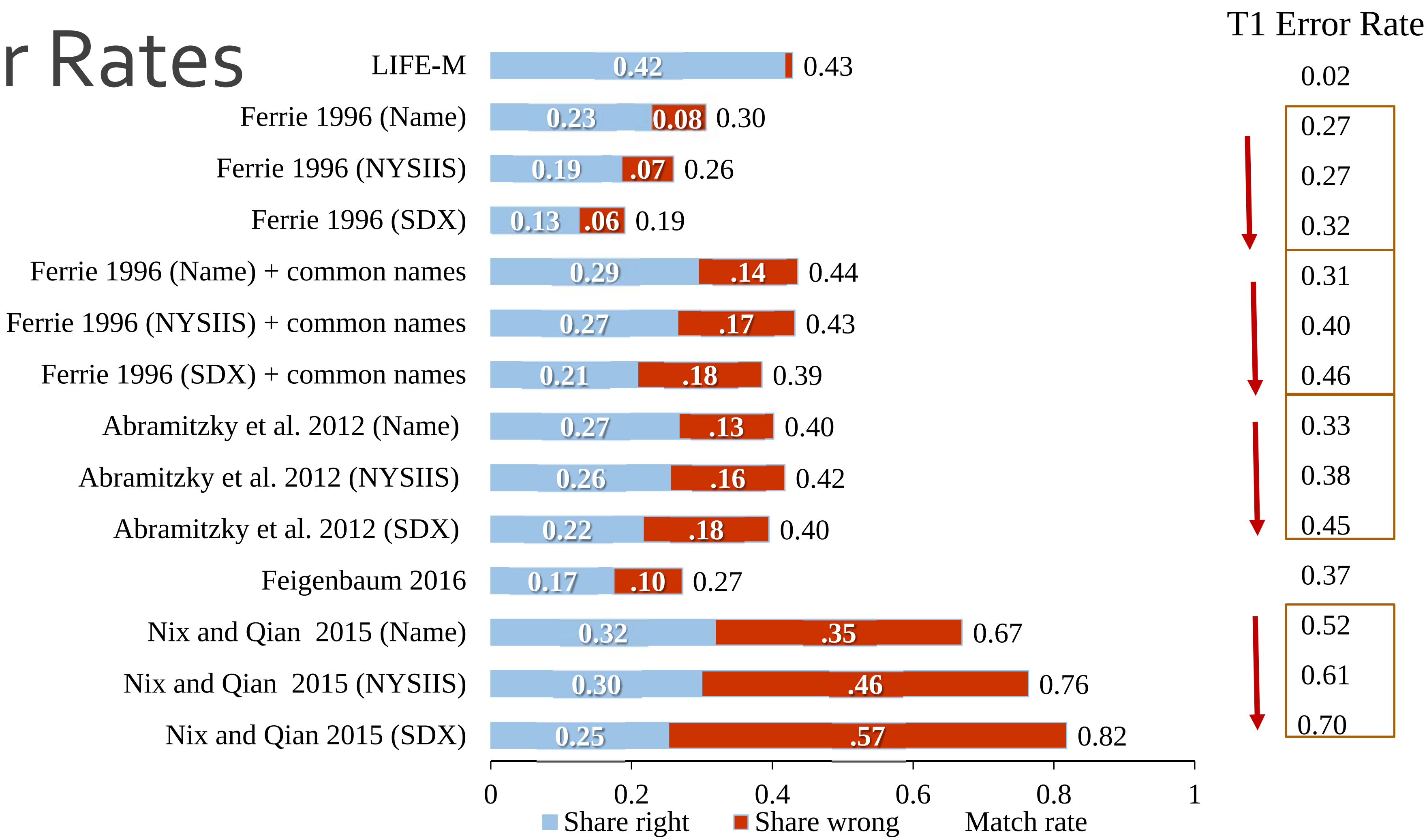


T1 Error Rate
LIFE-M 0.42 0.43 0.02
Ferrie 1996 (Name) 0.23 0.08 0.30 0.27
Ferrie 1996 (NYSIIS) 0.19 .07 0.26 0.27
Ferrie 1996 (SDX) 0.13 .06 0.19 0.32
Ferrie 1996 (Name) + common names 0.29 .14 0.44 0.31
Ferrie 1996 (NYSIIS) + common names 0.27 .17 0.43 0.40
Ferrie 1996 (SDX) + common names 0.21 .18 0.39 0.46
Abramitzky et al. 2012 (Name) 0.27 .13 0.40 0.33
Abramitzky et al. 2012 (NYSIIS) 0.26 .16 0.42 0.38
Abramitzky et al. 2012 (SDX) 0.22 .18 0.40 0.45
Feigenbaum 2016 0.17 .10 0.27 0.37
Nix and Qian 2015 (Name) 0.32 .35 0.67 0.52
Nix and Qian 2015 (NYSIIS) 0.30 .46 0.76 0.61
Nix and Qian 2015 (SDX) 0.25 .57 0.82 0.70
0 0.2 0.4 0.6 0.8 1
Share right
Share wrong
Match rate

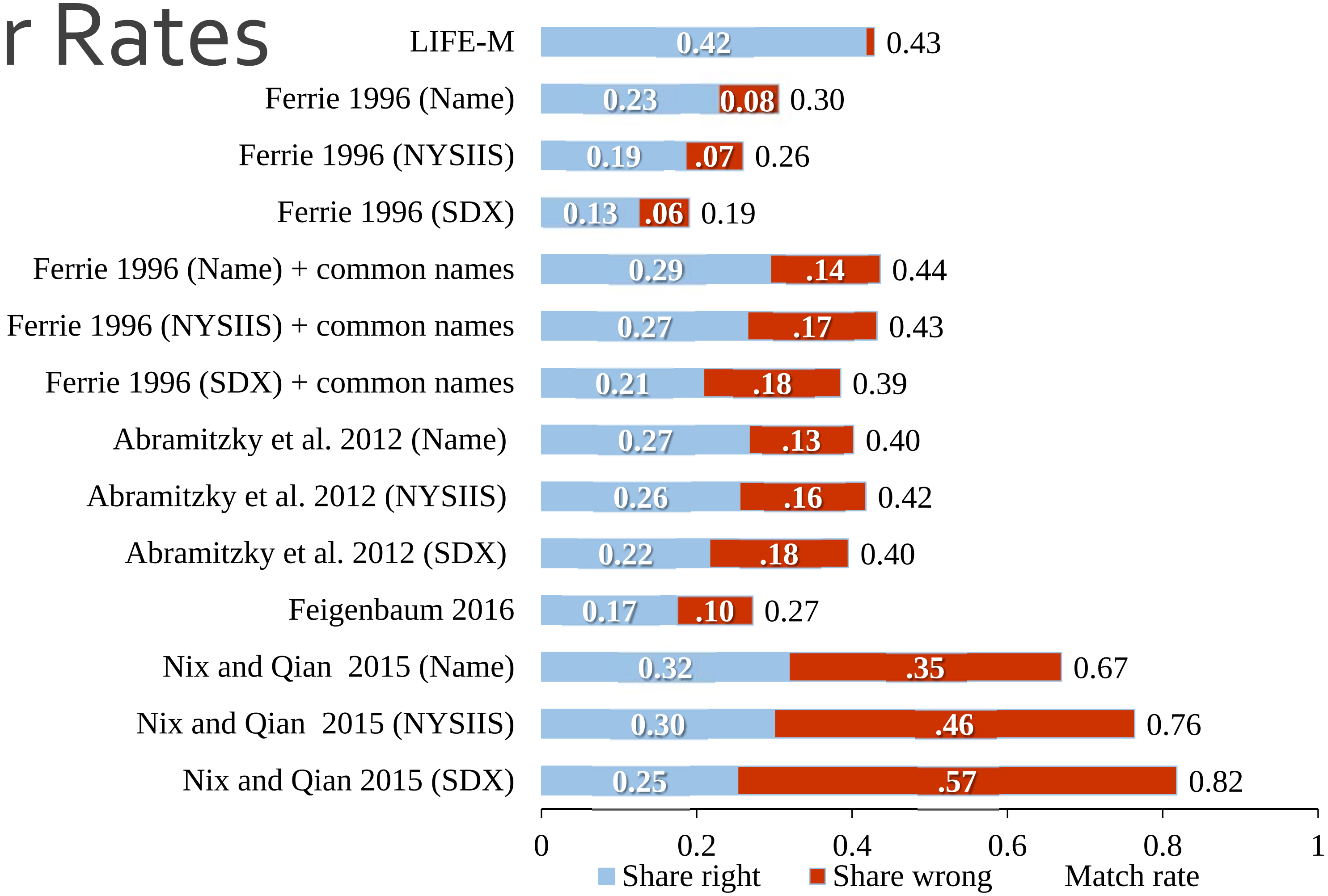
Error Rates


T1 Error Rate
LIFE-M 0.42 0.43 0.02
Ferrie 1996 (Name) 0.23 0.08 0.30 0.27
Ferrie 1996 (NYSIIS) 0.19 .07 0.26 0.27
Ferrie 1996 (SDX) 0.13 .06 0.19 0.32
Ferrie 1996 (Name) + common names 0.29 .14 0.44 0.31
Ferrie 1996 (NYSIIS) + common names 0.27 .17 0.43 0.40
Ferrie 1996 (SDX) + common names 0.21 .18 0.39 0.46
Abramitzky et al. 2012 (Name) 0.27 .13 0.40 0.33
Abramitzky et al. 2012 (NYSIIS) 0.26 .16 0.42 0.38
Abramitzky et al. 2012 (SDX) 0.22 .18 0.40 0.45
Feigenbaum 2016 0.17 .10 0.27 0.37
Nix and Qian 2015 (Name) 0.32 .35 0.67 0.52
Nix and Qian 2015 (NYSIIS) 0.30 .46 0.76 0.61
Nix and Qian 2015 (SDX) 0.25 .57 0.82 0.70
0 0.2 0.4 0.6 0.8 1
Share right
Share wrong
Match rate

Error Rates

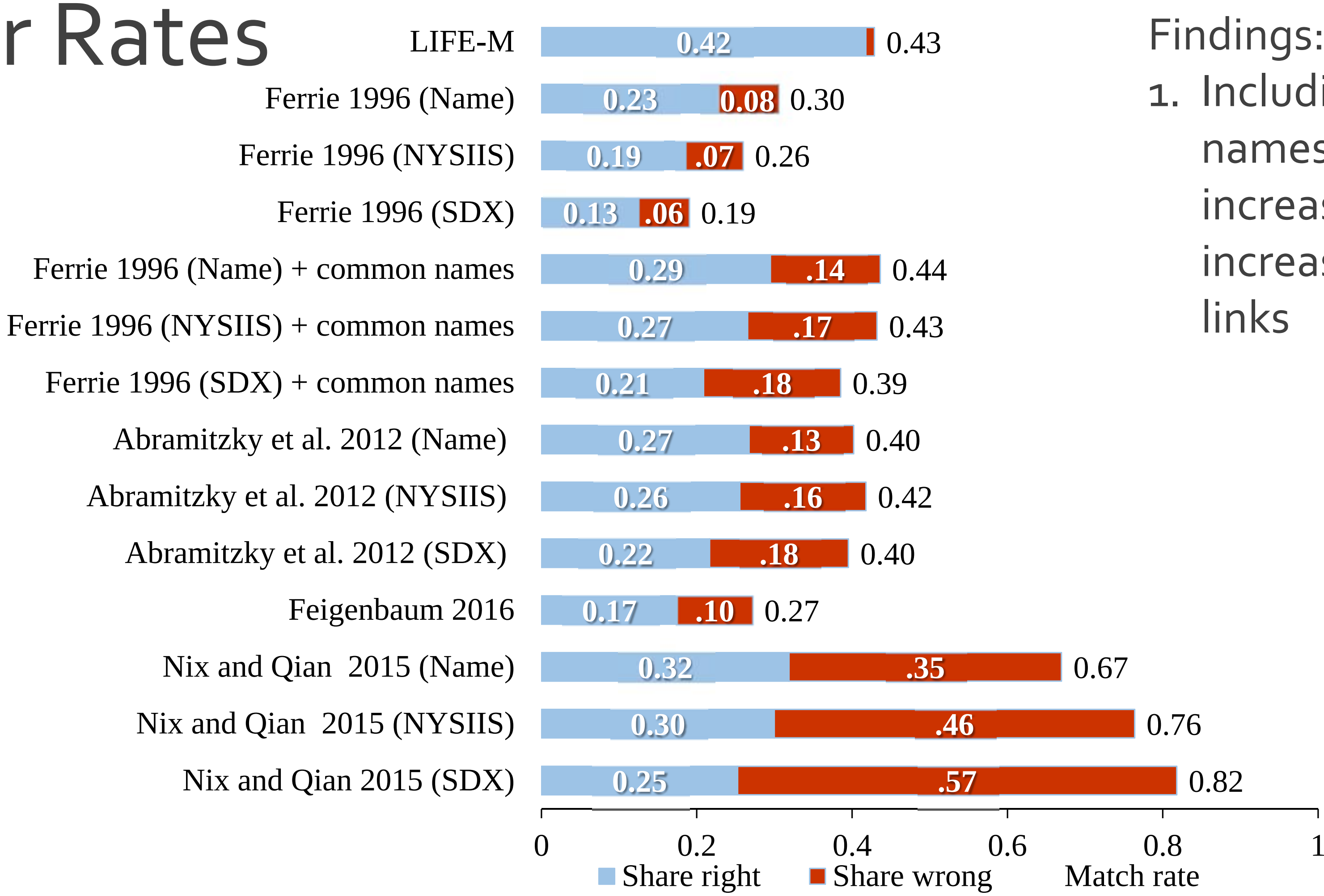

T1 Error Rate
LIFE-M 0.42 0.43 0.02
Ferrie 1996 (Name) 0.23 0.08 0.30 0.27
Ferrie 1996 (NYSIIS) 0.19 .07 0.26 0.27
Ferrie 1996 (SDX) 0.13 .06 0.19 0.32
Ferrie 1996 (Name) + common names 0.29 .14 0.44 0.31
Ferrie 1996 (NYSIIS) + common names 0.27 .17 0.43 0.40
Ferrie 1996 (SDX) + common names 0.21 .18 0.39 0.46
Abramitzky et al. 2012 (Name) 0.27 .13 0.40 0.33
Abramitzky et al. 2012 (NYSIIS) 0.26 .16 0.42 0.38
Abramitzky et al. 2012 (SDX) 0.22 .18 0.40 0.45
Feigenbaum 2016 0.17 .10 0.27 0.37
Nix and Qian 2015 (Name) 0.32 .35 0.67 0.52
Nix and Qian 2015 (NYSIIS) 0.30 .46 0.76 0.61
Nix and Qian 2015 (SDX) 0.25 .57 0.82 0.70
0 0.2 0.4 0.6 0.8 1
Share right
Share wrong
Match rate

Error Rates



LIFE-M
0.42
0.43
Ferrie 1996 (Name)
0.23
0.08
0.30
Ferrie 1996 (NYSIIS)
0.19
.07
0.26
Ferrie 1996 (SDX)
0.13
.06
0.19
Ferrie 1996 (Name) + common names
0.29
.14
0.44
Ferrie 1996 (NYSIIS) + common names
0.27
.17
0.43
Ferrie 1996 (SDX) + common names
0.21
.18
0.39
Abramitzky et al. 2012 (Name)
0.27
.13
0.40
Abramitzky et al. 2012 (NYSIIS)
0.26
.16
0.42
Abramitzky et al. 2012 (SDX)
0.22
.18
0.40
Feigenbaum 2016
0.17
.10
0.27
Nix and Qian 2015 (Name)
0.32
.35
0.67
Nix and Qian 2015 (NYSIIS)
0.30
.46
0.76
Nix and Qian 2015 (SDX)
0.25
.57
0.82
0
0.2
0.4
0.6
0.8
1
Share right
Share wrong
Match rate

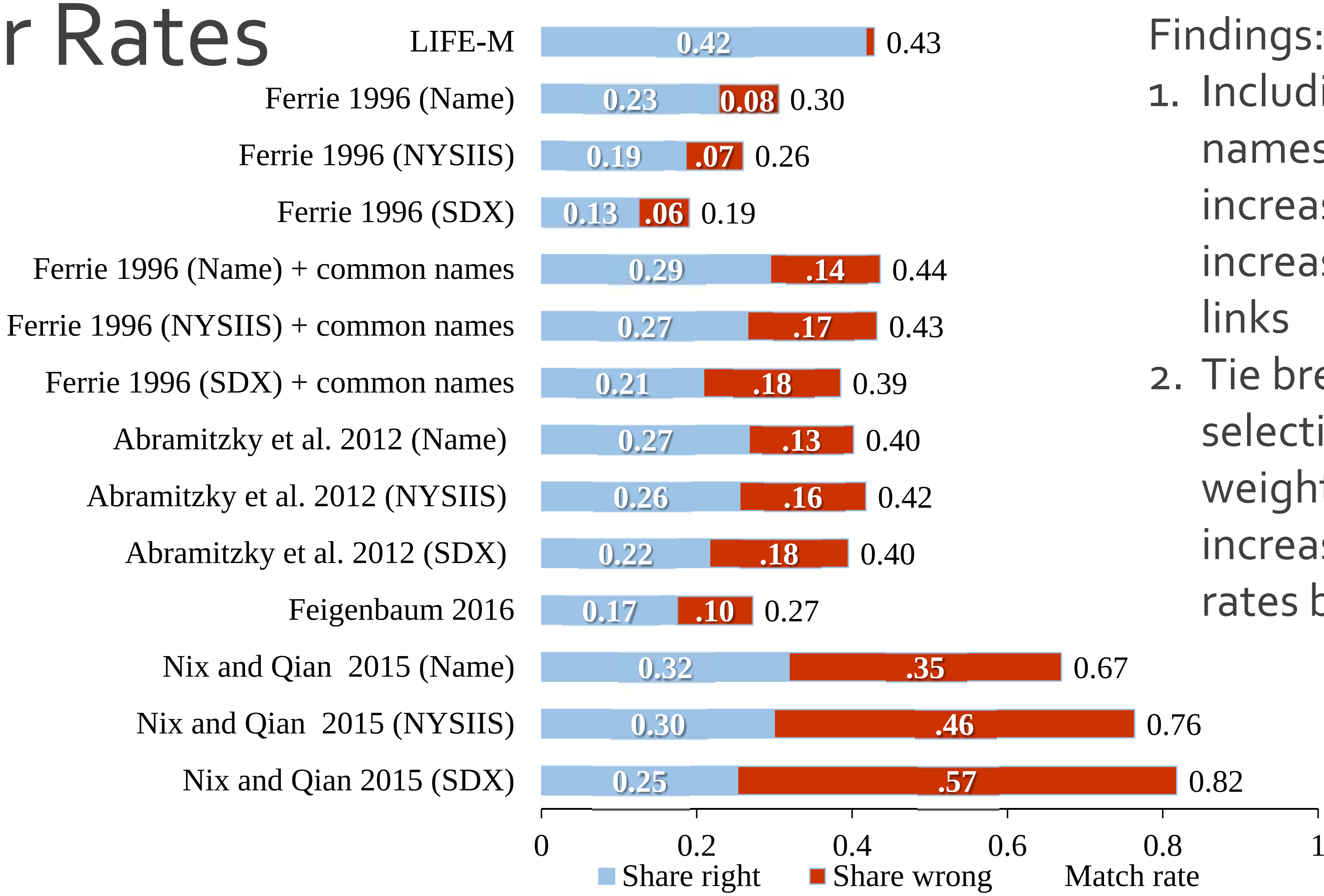
Error Rates



Findings:

1. Including common names sample increases errors but increases correct links

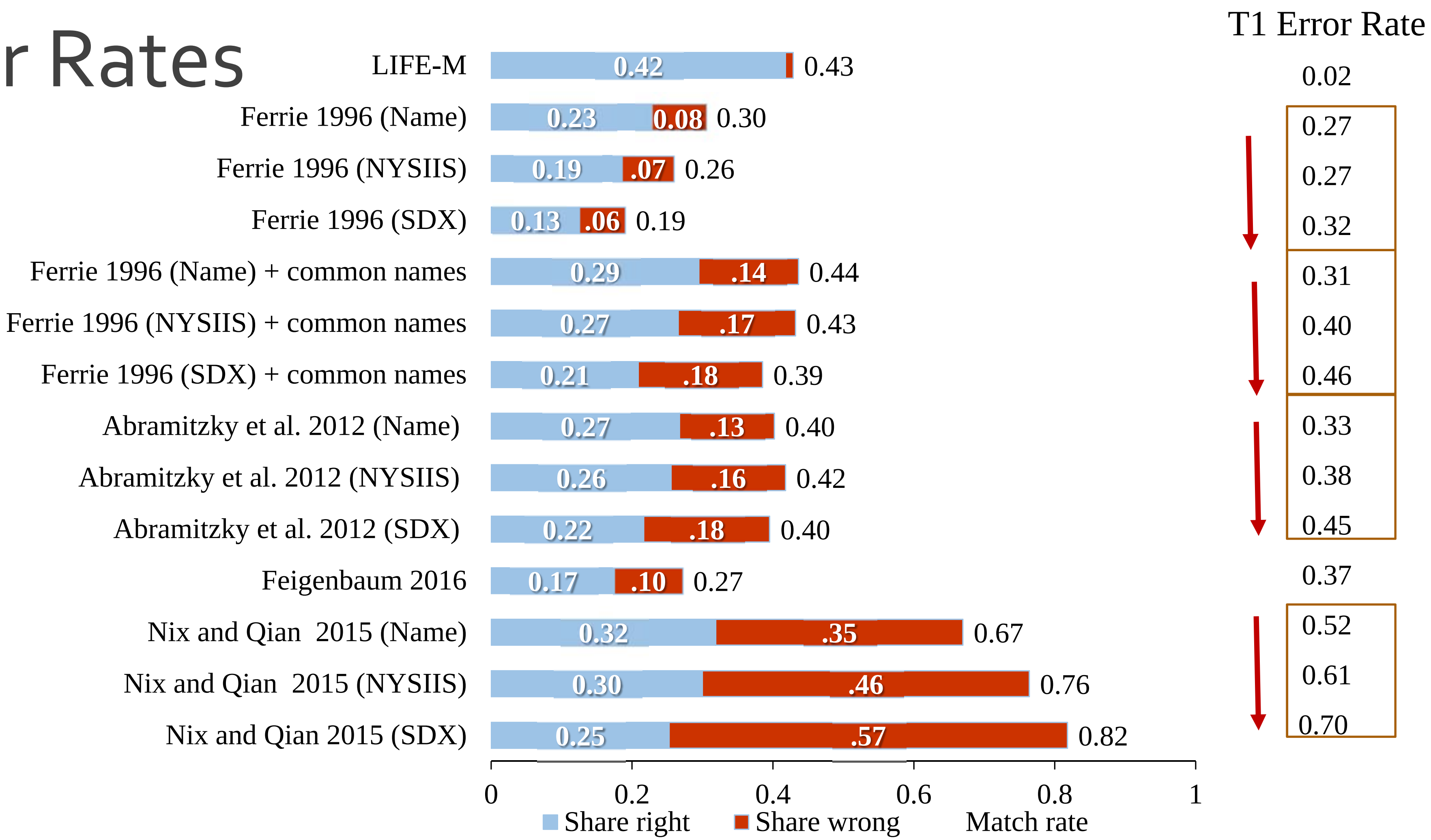
Error Rates



Findings:

1. Including common names sample increases errors but increases correct links
2. Tie breaks (random selection or 1/m weighting) increases match rates but also errors

Error Rates



T1 Error Rate
LIFE-M
Ferrie 1996 (Name)
Ferrie 1996 (NYSIIS)
Ferrie 1996 (SDX)
Ferrie 1996 (Name) + common names
Ferrie 1996 (NYSIIS) + common names
Ferrie 1996 (SDX) + common names
Abramitzky et al. 2012 (Name)
Abramitzky et al. 2012 (NYSIIS)
Abramitzky et al. 2012 (SDX)
Feigenbaum 2016
Nix and Qian 2015 (Name)
Nix and Qian 2015 (NYSIIS)
Nix and Qian 2015 (SDX)
0.42 0.43
0.23 0.08 0.30
0.19 .07 0.26
0.13 .06 0.19
0.29 .14 0.44
0.27 .17 0.43
0.21 .18 0.39
0.27 .13 0.40
0.26 .16 0.42
0.22 .18 0.40
0.17 .10 0.27
0.32 .35 0.67
0.30 .46 0.76
0.25 .57 0.82
0 0.2 0.4 0.6 0.8 1
Share right
Share wrong
Match rate
0.02
0.27
0.27
0.32
0.31
0.40
0.46
0.33
0.38
0.45
0.37
0.52
0.61
0.70