A New Edit for Identifying Potential Gender Misclassification in Central Cancer Registry Databases.

Soloway LE, Boscoe FP, Kahn AR

New York State Cancer Registry, New York Department of Health, Albany, NY 12204

Introduction

- Many registries rely only on sex-specific cancers to identify errors in the sex field
  - accounts for about 15% of the cases in which sex may be incorrectly coded.

Introduction

- Categories that incorrect coding of sex may be identified with:
  - Names that are highly sex-specific
  - Names that have sex-specific spellings and may be spelled wrong
  - Names that have changed sex assignment over the decade
  - Names that may have different sex assignments according to a culture or language
  - Names that have no specific sex assignment (gender-neutral)

Introduction

- We focused on the first three of these points
  - Names that are highly sex-specific
  - Names that have sex-specific spellings and may be spelled wrong
  - Names that have changed sex assignment over the decade
- Our program flags suspicious name/sex combinations for manual review

Methods

- For the period 1890-2008, the most popular birth names per decade were downloaded from the SSA database

- If the ratio of males to females with that name in that decade was greater than 49:1 (98%), it was considered to be a male name and vice versa

Methods

- This was then merged in with the New York State Cancer Registry database
- Sex-specific cancers were excluded
- Those with conflicting data for name and sex fields were flagged

Results

- 21,784 name/birth decade combinations were identified as questionable
- 8,285 cases flagged
- 0.3% of the registry
- 1,495 different names flagged

Results

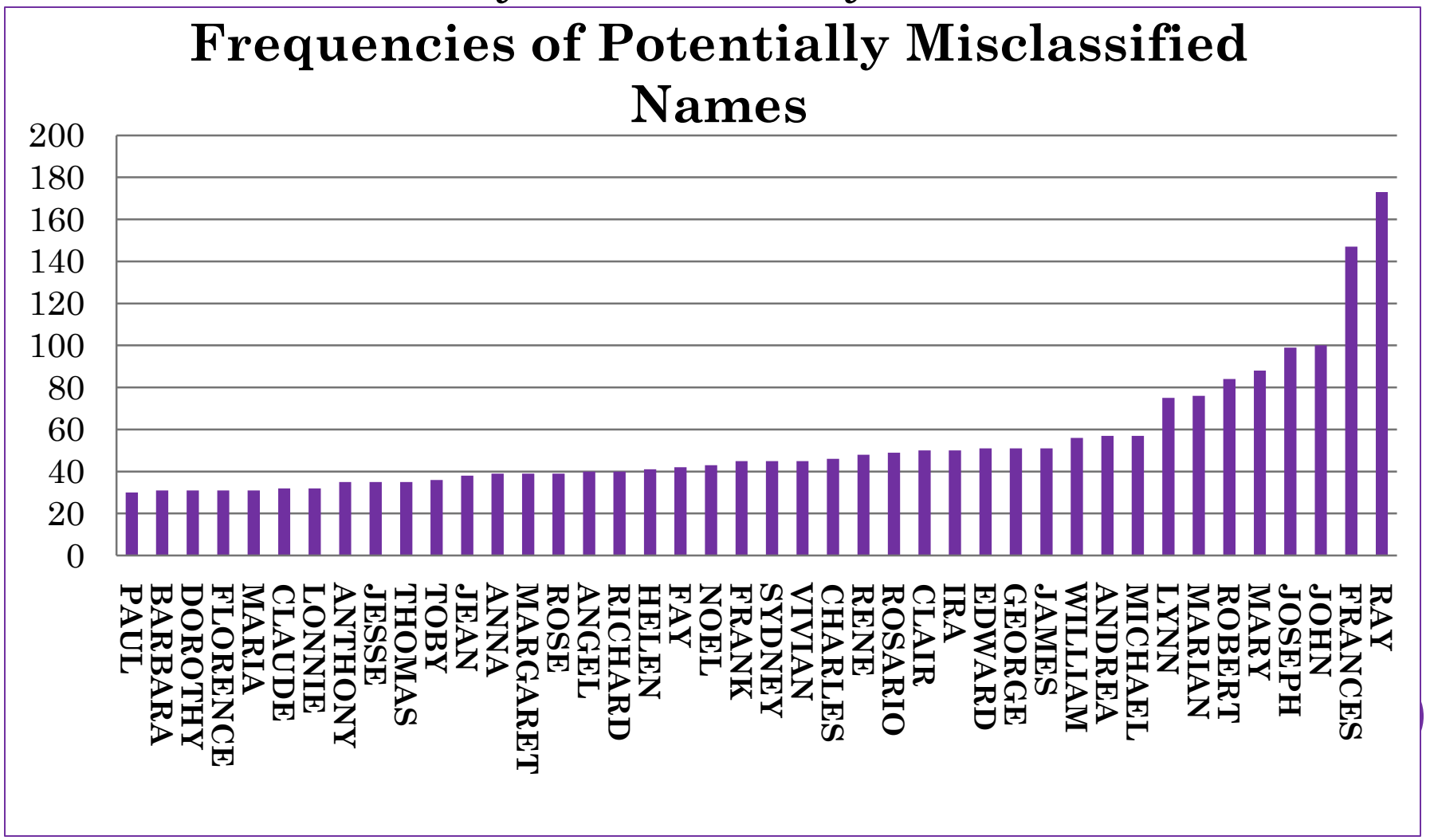
- Many names changed gender affiliation over the decades

Decade	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Allison	M	M	M	M	---	F	F	F	F	F	F	F
Artie	pF	pF	pF	pF	F	M	M	M	---	---	---	---
Ashley	M	M	M	---	---	M	M	pF	pF	pF	pF	F
Beverly	M	pF	pF	pF	pF	pF	F	F	F	F	F	---
Elisha	M	M	M	M	M	pF	M	---	---	F	pF	M
Lauren	---	---	M	M	M	pF	pF	F	F	pF	F	F
Lindsay	M	M	M	---	---	M	M	M	F	F	F	F
Lindsey	M	M	M	M	M	M	M	M	pF	pF	F	F
Lindy	---	---	---	M	---	---	F	---	F	F	---	---
Meredith	---	M	pM	pF	pF	pF	F	F	F	F	F	F
Robbie	F	F	F	F	pF	pF	pF	pM	pM	M	---	---
Rosario	---	M	M	M	M	F	F	F	F	---	---	---
Sandy	M	M	M	M	pM	pF	pF	pF	pF	F	F	F
Sydney	M	M	M	M	pM	pF	pF	F	---	F	F	F

M=Male F=Female pM=Probable Male pF=Probable female

Results

- The most common potentially misclassified name was “Ray” followed by “Frances”

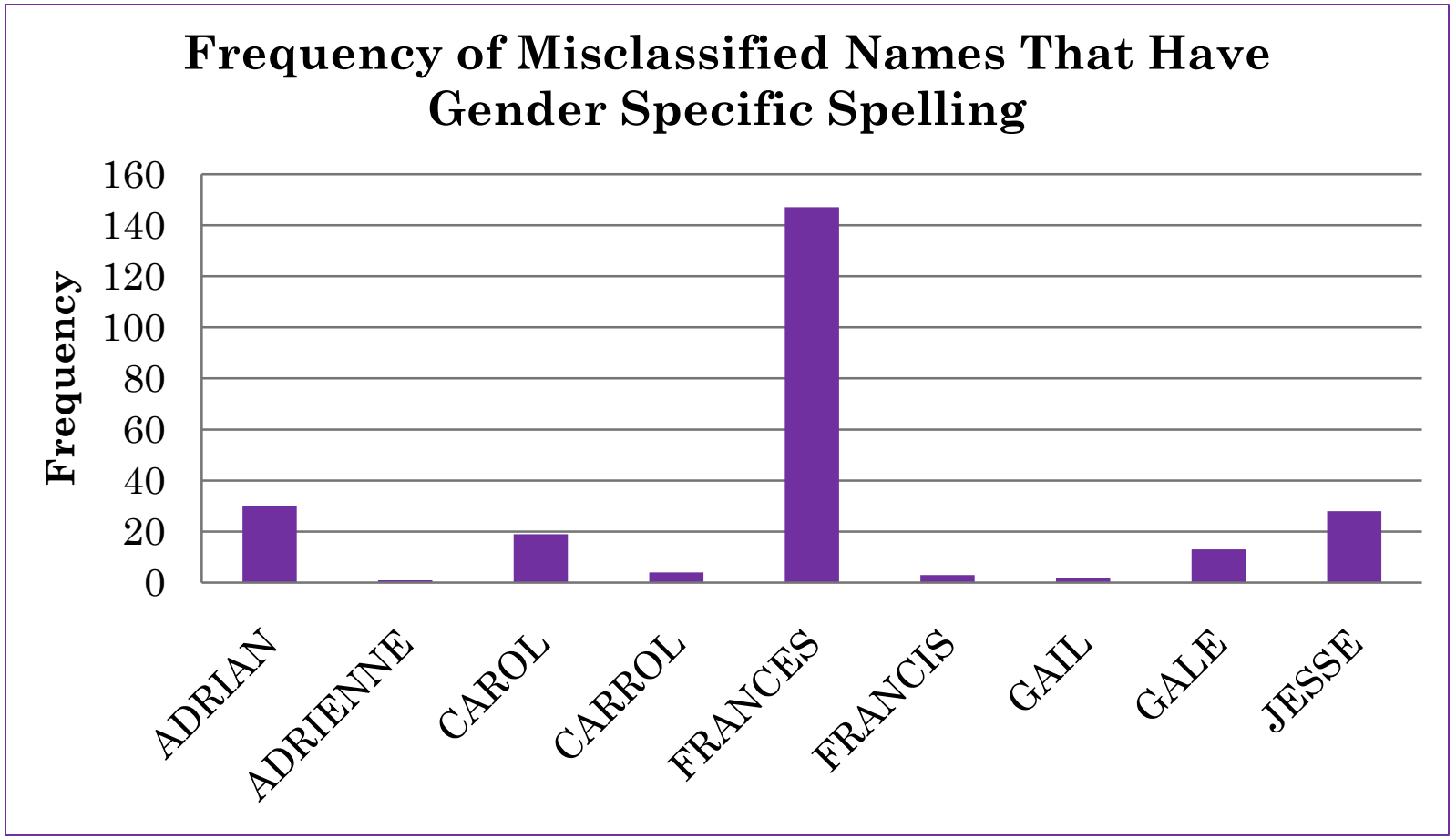


Results

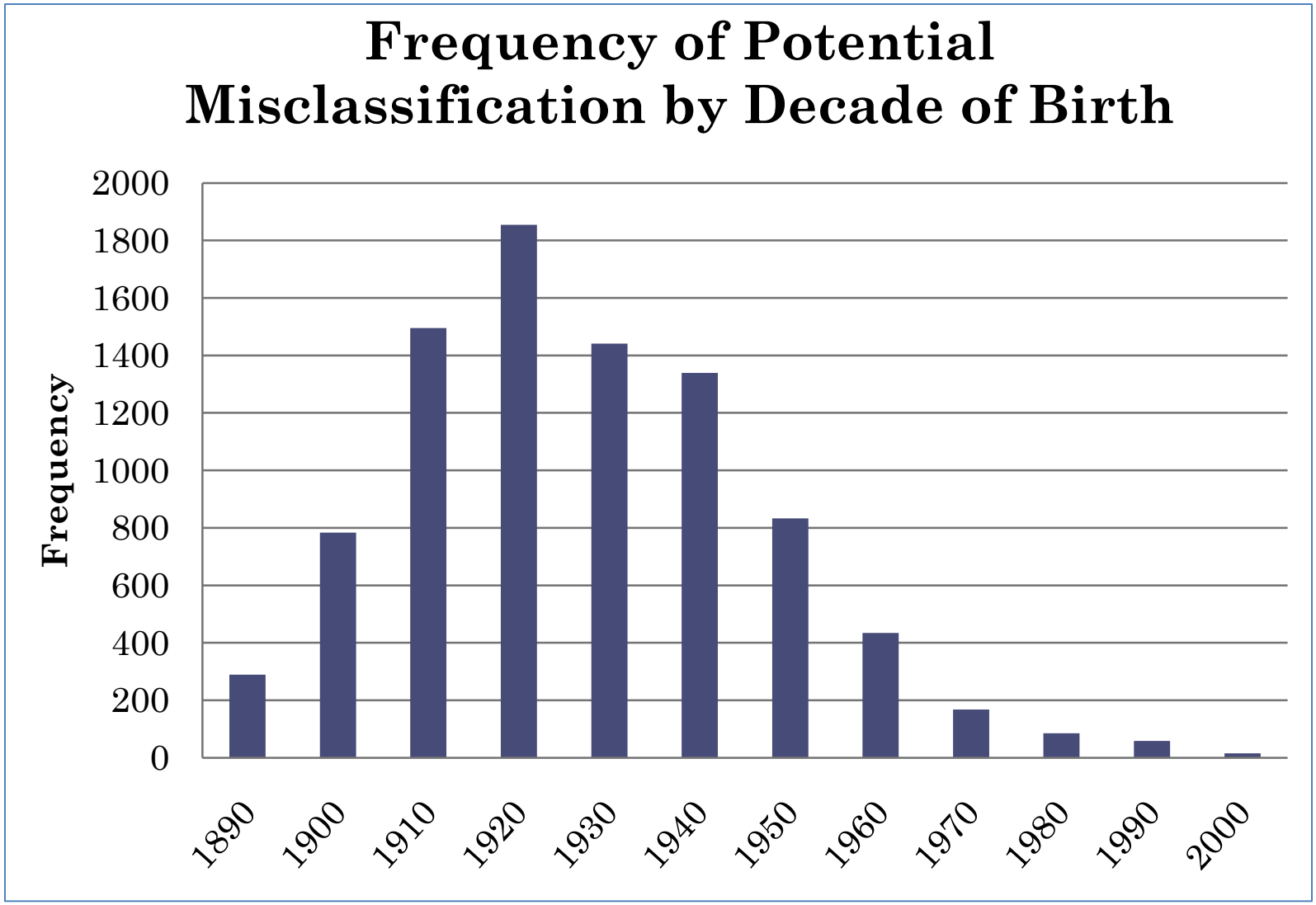
- In 1920's:
  - Frances was the 10th most popular female name
  - Frances was the 596th most popular male name
  - Francis was the 35th most popular male name

Results

- Several names are frequently misspelled leading to a false flag



RESULTS


Frequency of Potential Misclassification by Decade of Birth
Frequency
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

Results

- Those born in 1990 had the highest percentage of potential misclassification

Decade	Number Misclassified	Number of Cases	Percent Misclassified
1890	289	99783	0.289628
1900	783	289896	0.270097
1910	1495	517754	0.288747
1920	1855	632208	0.293416
1930	1441	452180	0.318678
1940	1339	344261	0.388949
1950	833	218656	0.380964
1960	434	100535	0.43169
1970	168	34860	0.481928
1980	85	16940	0.501771
1990	58	8486	0.683479
2000	15	3331	0.450315

Results

- Because this new edit caused a lot of new records to be flagged, we decided to cut down the number of records flagged by doing the following:
  - Limiting the names on the list to those that were present in 1000 or more people born in that decade
  - Eliminating names with sex-specific spellings
  - Eliminating ambiguous nicknames.
    - Pat
    - Lou
    - Chris
    - Merle

RESULTS

- Eliminating names from certain decades where there was an unusual pattern:
  - Dana 1920
  - Dale 1910,1920
  - Lee 1890
  - Leslie 1900
  - Randy 1950

- Eliminating certain cases that often were born outside the US with language/culture specific names:
  - Jean
  - Carmen
  - Andrea
  - Angel

RESULTS

- Nearly all of the flagged cases did indeed have incorrect gender, when verified using external sources or source text information
- This project had a strong impact on male breast cancer
- Nearly 10% of our male breast cancer cases had miscoded gender
- Male breast cancer is even rarer than we think it is

Final Database

- 15788 name/birth decade combinations were identified as questionable
- 7500 cases flagged
- 1203 different names

Conclusions

- Those born between 1910 and 1940 are the most likely to be misclassified, but they are also the most likely to have a cancer diagnosis.

- The decade with the highest percent of potential misclassification is 1990 and this may continue to be more prevalent in the future because of ambiguous naming conventions

Conclusions

- Ambiguous naming conventions
  - names that change over decades increasing in frequency
  - the use of the mother's maiden name or family last names as a child's first name
  - increasing use of gender-neutral names.

Conclusions

- Need to take into account:
  - names that change sex affiliation over decades
  - names that may be misspelled frequently
  - names that are likely to be male or female based on the decade.

Strengths

- More sensitive and more specific than other algorithms of which we are aware

Limitations

- Social security cards not required before 1937
- Over 7000 names already on the registry flagged and many more coming in to be flagged

Future Aspects to Think About

- Names of people born outside the US
- Names that have no specific sex assignment (gender-neutral)
- When next to update the tables

Acknowledgements

- CA Cancer registry
- Walter Fuller
- Funding Sources:
  - This work was supported in part by the Centers for Disease Control and Prevention's Cooperative Agreement U58/DP000783, awarded to the New York State Department of Health through the National Program of Cancer Registries.

Questions?