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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Allison</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>---</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Artie</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Ashley</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>---</td>
<td>---</td>
<td>M</td>
<td>M</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>F</td>
</tr>
<tr>
<td>Beverly</td>
<td>M</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>---</td>
</tr>
<tr>
<td>Elisha</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>pF</td>
<td>M</td>
<td>---</td>
<td>---</td>
<td>F</td>
<td>pF</td>
<td>M</td>
</tr>
<tr>
<td>Lauren</td>
<td>---</td>
<td>---</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>pF</td>
<td>pF</td>
<td>F</td>
<td>F</td>
<td>pF</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Lindsay</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>---</td>
<td>---</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Lindsey</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>pF</td>
<td>pF</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Lindy</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>M</td>
<td>---</td>
<td>---</td>
<td>F</td>
<td>---</td>
<td>F</td>
<td>F</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Meredith</td>
<td>---</td>
<td>M</td>
<td>pM</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Robbie</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>pM</td>
<td>pM</td>
<td>M</td>
<td>---</td>
</tr>
<tr>
<td>Rosario</td>
<td>---</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Sandy</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>pM</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>pF</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Sydney</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
<td>pM</td>
<td>pF</td>
<td>pF</td>
<td>F</td>
<td>---</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

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.

![Graph showing frequency of misclassified names with gender-specific spelling.](attachment:graph.png)
RESULTS

Frequency of Potential Misclassification by Decade of Birth
## RESULTS

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

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