National Center for Health Statistics
Data Linkage

Use of Alternate Information to Improve Linkage with the National Death Index (NDI)

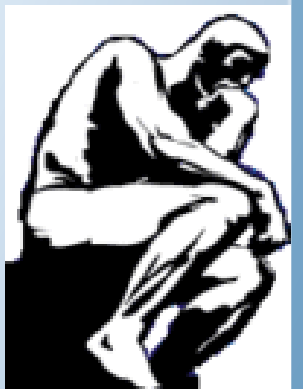
Eric A. Miller, Dean H. Judson, James Brittain, Jennifer D. Parker, Cordell Golden, Patricia Lloyd

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the CDC

Why Am I Here?

- Special Projects Branch (SPB) within the National Center for Health Statistics (NCHS - CDC)
  - Regularly conducts data linkages between population-based surveys and the NDI

- From previous experience in a cancer registry, recognize the importance of mortality linkages and feel these methods could potentially benefit registries

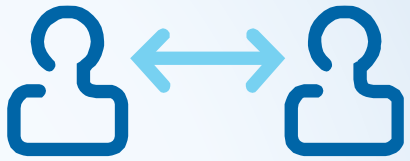


Reasons for Mismatches in Data Linkage

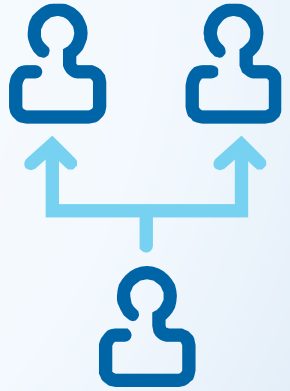
- Data entry errors
- Misreporting
- Variations between reporting sources
  - For example, a nickname versus proper name

Quick Background on Data Linkage

- Deterministic
  - Exact match on linkage variables
    - Frank ≠ Francis
    - 123456789 ≠ 213456789



- Probabilistic
  - Accounts for imperfect data
  - Probability of a match
    - Frank ≈ Francis
    - 123456789 ≈ 213456789



Novel (To Me) SPB Linkage Method

- To increase the likelihood of correctly linking records (and reduce false negatives), SPB creates additional records with alternate information available
- The file of survey participants used to link with administrative data (e.g. NDI) contains multiple records per person with every combination of available information
- Intuitively makes sense but has not been evaluated

Case Study: National Health Interview Survey Linkage with the NDI

- National Death Index (NDI)
  - A national file of identifying death record information (beginning with 1979 deaths)
    - File with records through December 31, 2011
  - Every few years we link survey participants to NDI to identify participant deaths
- NHIS (1986-2009)
  - In-person household survey
  - Conducted continuously by the CDC's NCHS since 1957
  - Extensive data collection
  - Large sample sizes
    - ~35,000 households per year

NDI Matching Fields

- Social Security Number (last 4 digits beginning in 2007)
- First name
- Middle initial
- Last name
- Date of birth (Month, Day, Year)
- Sex
- Father’s surname (for women when available)
- State of birth
- Race
- State of residence
- Marital Status

Alternate Information Includes

- SSNs
  - From other sources or linkages
- Dates of birth
  - From other sources or linkages
- Names
  - From other sources or linkages
  - Name look-up tables
  - Algorithms

Alternate Name Possibilities

- Substitute any alternate names available from other sources
- Substitute nickname with proper name (and *vice versa*)
- Multi-part first name (e.g. Billy Joe)
- Multi-part last name (e.g. Kidd-Gilchrist)
- Switch first name and middle name
- Submit a blank middle name
- Hispanic: Replace last with middle name
- Asian: Switch first and last name

Nickname Lookup Table (Extract)

SEX	NICKNAME	PROPER NAME
M	ABE	ABRAHAM
F	AGGIE	AGNES
M	AL	ALBERT
M	ALEX	ALEXANDER
M	ALF	ALFRED
F	ALLIE	ALBERTA
M	**ANDY**	**ANDREW**

Example: If first name='Andy' then alternate record first name='Andrew'

Alternate Name Records Example

Number	First	Middle	Last
1	David	Américo	Arias Ortiz
2	David	Américo	Ortiz
3	David	Américo	Arias
4	David		Américo
5	Big		Papi

Study Objectives

- Determine the effectiveness of this method
  - **Question 1**: How many links (and %) had an alternate record as the highest scoring record?
  - **Question 2**: Which alternate information was used most often?
  - **Question 3**: How many links were identified with an alternate record that would have been missed otherwise?

Methods

- **Question 1**: How many links (and %) had an alternate record as the highest scoring record?

- **Question 2**: Which alternate information was used most often?

  - File 1
    - Final disposition (vital status) merged with alternate record flags
      - Flags are created for each type of alternate record
        - » Name, SSN, DOB

Methods

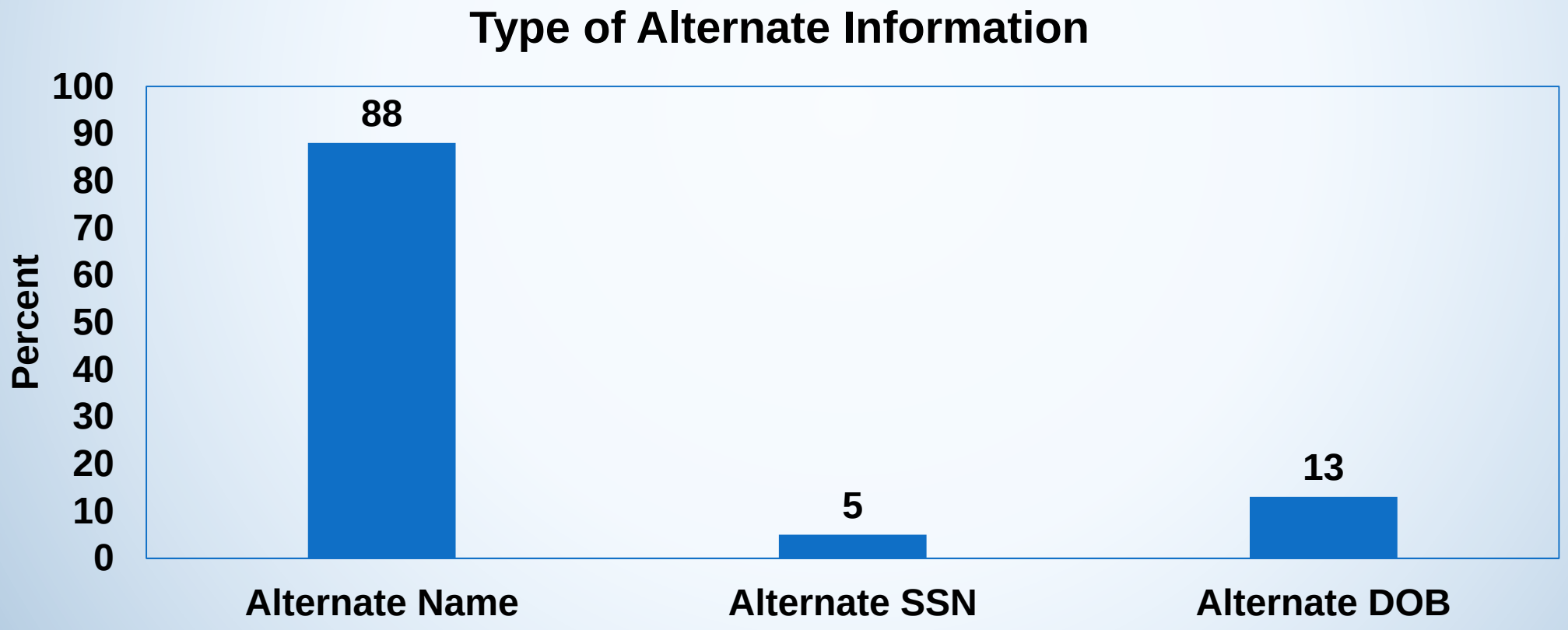
- **Question 3**: How many links were identified with an alternate record that would have been missed otherwise?
  - File 2
    - File with all alternate records and all possible matches with the NDI
      - Select the highest scoring original record and highest scoring alternate record
        - » Based on score cutoffs used to determine links, compare number considered linked for original and alternate records

Study Population

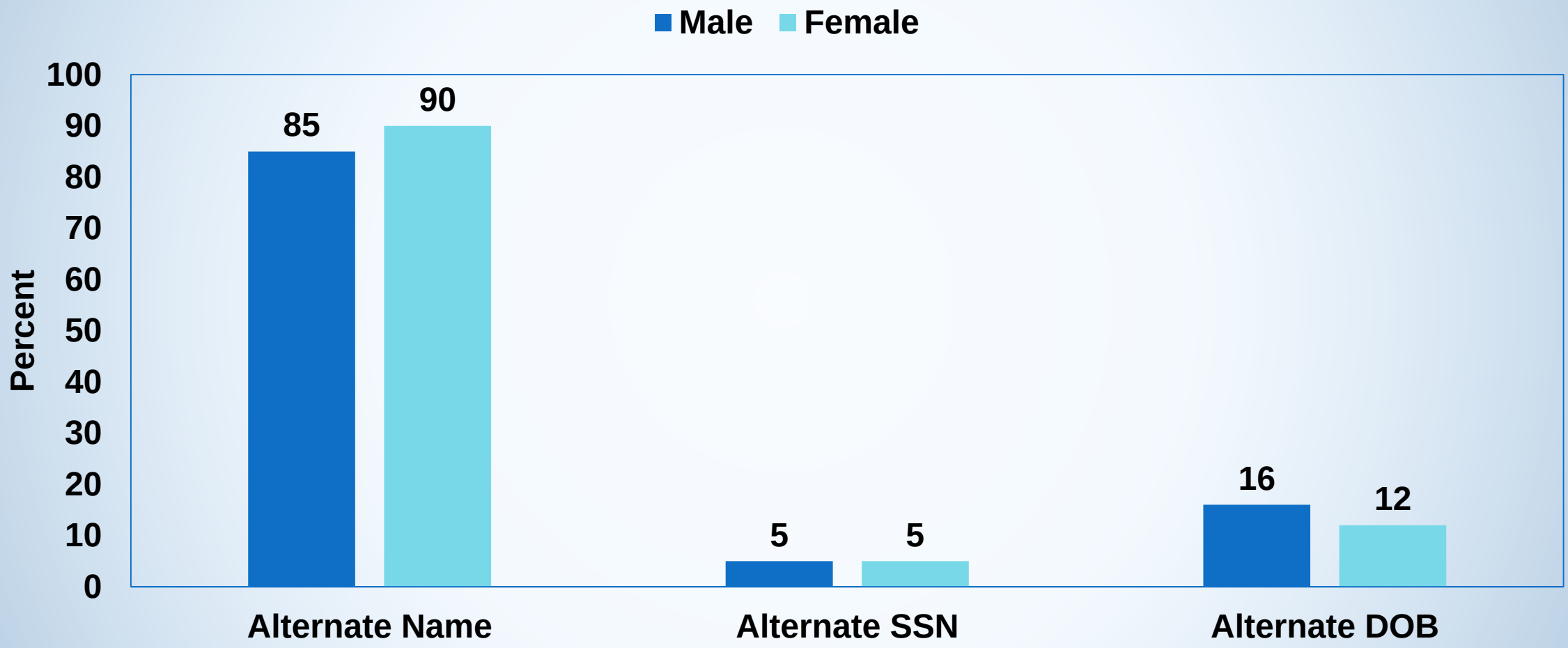
- There were 2,235,985 participants included in this analysis
  - 275,729 (12%) were considered linked to an NDI record
    - Ranges from 23% in 1986 NHIS to 2% in 2009 NHIS

Results – Questions 1 and 2

- Of those considered linked
  - Approximately 23% included some alternate information (i.e. highest scoring record)



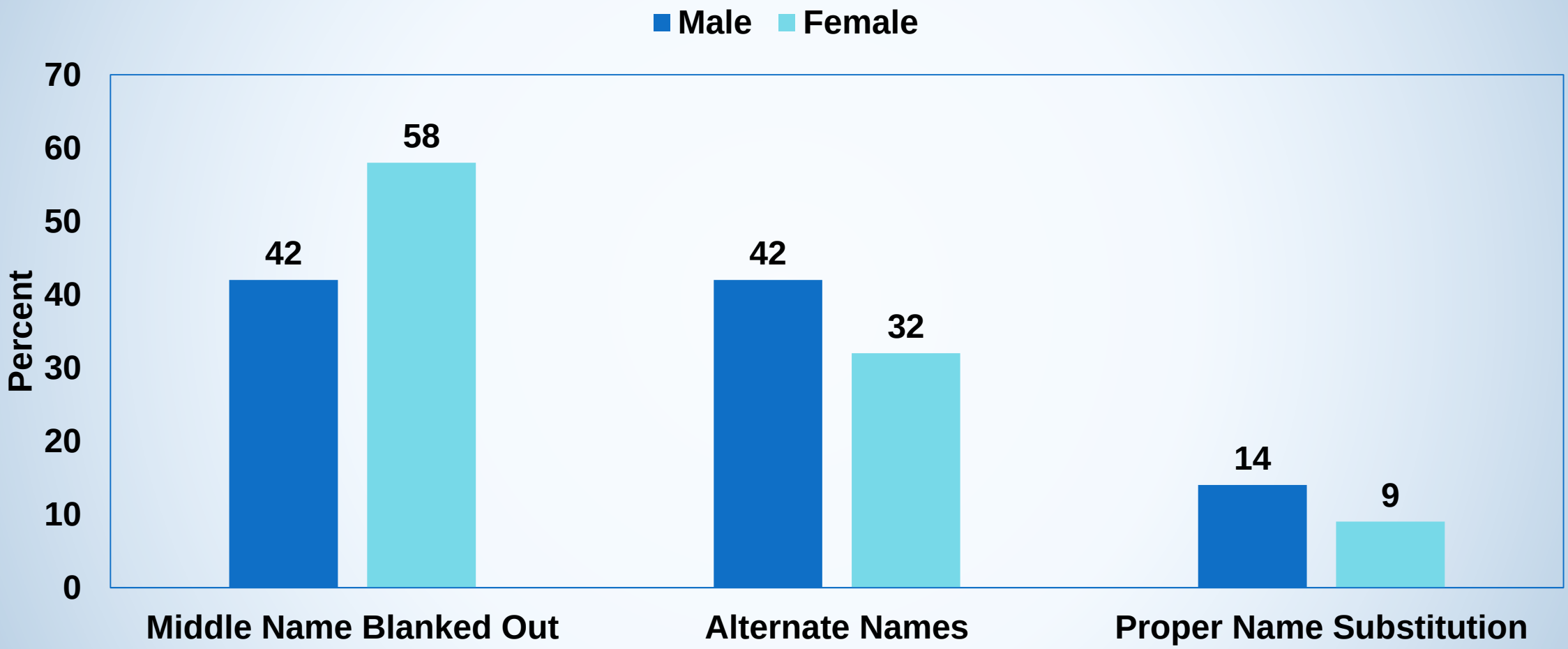
Type of Alternate Information by Sex



Frequency of Alternate Name Records Used

Type of Alternate Name Record	N	Percent
Blank Out Middle Name	28,990	52%
Alternate Name (First, Middle, or Last) from Another Source	19,981	36%
Proper Name to Replace Nickname	6,757	12%
Multi-Part First Name	746	1%
Switch First and Middle Name	477	0.9%
Multi-Part Last Name	300	0.5%
Asian: Switch First and Last Name	115	0.2%
Nickname to Replace First Name	102	0.2%
Hispanic: Switch Middle and Last Name	92	0.2%

Type of Alternate Name Information by Sex



Results – Question3

- How many links were identified with an alternate record that would have been missed otherwise?
  - There were close to 5,000 links that were found with an alternate record that would have been missed with only the original record
    - Translates to ~2% of links

Summary

- Almost one quarter of the highest scoring links used some type of alternate information
- While the percentage of additional links was small (~2%), the method added a meaningful number of links
- For alternate name records, blanking out middle name, replacing nicknames and using names from other sources (if available) were most effective

Limitations

- Used a study population with a relatively high number of participants without SSN
- Were not able to isolate use of ethnic specific name substitution
- There is some potential to create false positive links
  - But with appropriate score cut-offs risk appears small
  - We were not able to fully evaluate this yet

Conclusions

- Use of alternate records was successful at finding links that would have been missed
- May be beneficial for cancer registries to use alternate names (or replace nicknames), SSNs and/or dates of birth for multiple submissions in linkages
- Registries may want to experiment linking with and without middle name/initial

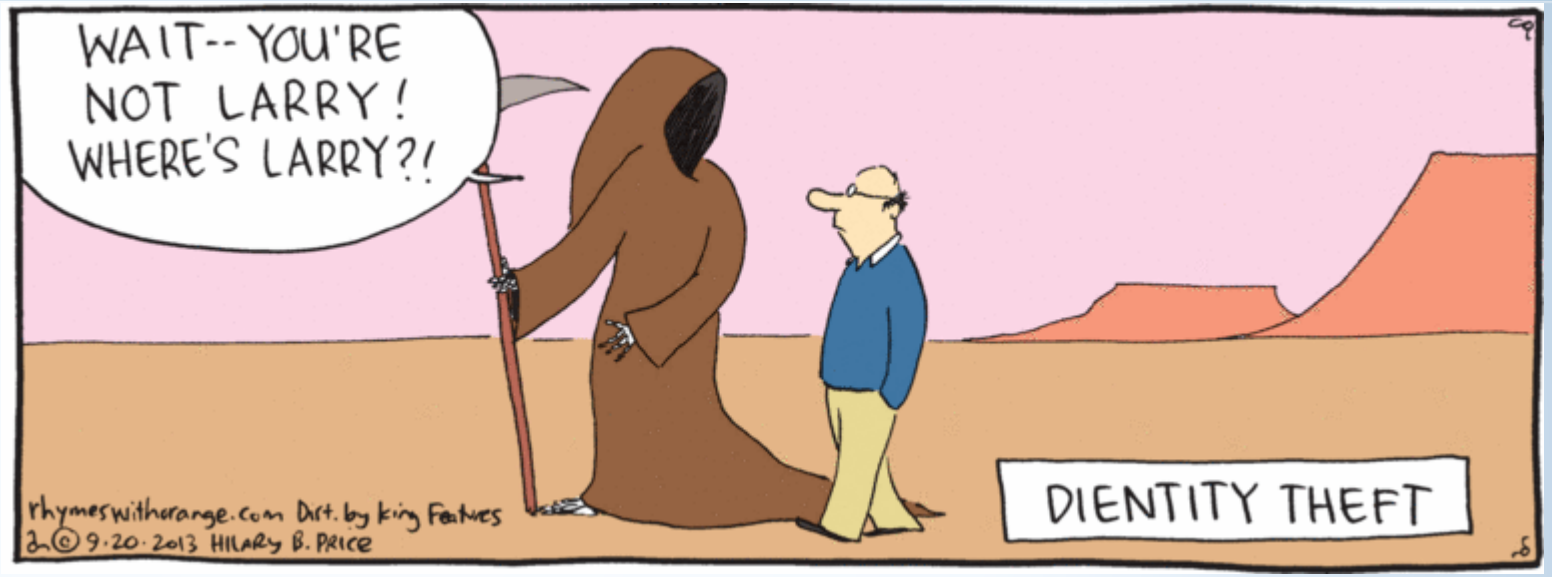
Thanks!

Acknowledgements:

Keith Zevallos

Eileen Call

Jesse Bassich



Contact Info:

Eric A. Miller bwe6@cdc.gov

or datalinkage@cdc.gov