University of Pennsylvania
ScholarlyCommons

2018 ADRF Network Research Conference Presentations

ADRF Network Research Conference Presentations

11-2018

# Simulation Approach to Assess the Precision of Estimates Derived from Linking Survey and Administrative Records

Dean M. Resnick
*NORC*

Lisa B. Mirel
*National Center for Health Statistics*

Follow this and additional works at: https://repository.upenn.edu/admindata_conferences_presentations_2018

Resnick, Dean M. and Mirel, Lisa B., "Simulation Approach to Assess the Precision of Estimates Derived from Linking Survey and Administrative Records" (2018). *2018 ADRF Network Research Conference Presentations*. 33.
https://repository.upenn.edu/admindata_conferences_presentations_2018/33

DOI https://doi.org/10.23889/ijpds.v3i5.1062

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/admindata_conferences_presentations_2018/33
For more information, please contact repository@pobox.upenn.edu.

# Simulation Approach to Assess the Precision of Estimates Derived from Linking Survey and Administrative Records


## Abstract

Probabilistic record linkage implies that there is some level of uncertainty related to the classification of pairs as links or non-links vis-à-vis their true match status. As record linkage is usually performed as a preliminary step to developing statistical estimates, the question then is how does this linkage uncertainty propagate to them? In this paper, we develop an approach to estimate the impact of linkage uncertainty on derived estimates by using a re-sampling approach. For each iteration of the re-sampling, pairs are classified as links or non-links by Monte-Carlo assignment to model estimated true match probabilities. By looking at the range of estimates produced in a series of re-samples, we can estimate the distribution of derived statistics under the prevailing incidence of linkage uncertainty. For this analysis we use the results of linking the 2014 National Hospital Care Survey to the National Death Index performed at the National Center for Health Statistics. We assess the precision of hospital-level death rate estimates.



## Comments

DOI https://doi.org/10.23889/ijpds.v3i5.1062



This presentation is available at ScholarlyCommons: https://repository.upenn.edu/admindata_conferences_presentations_2018/33

# Simulation Approach to Assess the Precision of Estimates Derived from Linking Survey and Administrative Records

Dean M. Resnick – NORC
Lisa B. Mirel – National Center for Health Statistic (NCHS)

2018 Administrative Data Research Conference
November 13, 2018

NORC at the UNIVERSITY of CHICAGO

# Record Linkage

*Considerations*

- Generally, the purpose of record linkage is to enable the computation of estimates not possible in each data source alone
- Linkages incur two basic types of errors
  - Type I: Links are made which are not true matches
  - Type II: True matches are not represented among links
- Naturally, the question is how much uncertainty do these errors engender in derived estimates.

# Overview

- To address this uncertainty in derived estimates we ran a Monte-Carlo Simulation: akin to jackknife
- The data source is the National Hospital Care Survey linked with the National Death Index data*
  - Derived estimate will be mean number of deaths within 30 days of discharge
  - Estimate is unweighted and not nationally representative

*see: https://www.cdc.gov/nchs/data/datalinkage/NHCS14_NDI14_15_Methodology_Analytic_Consider.pdf



NORC at the UNIVERSITY of CHICAGO

# Overview

*Additional Details*

- Multiple simulations of the record linkage process
  - Compute variances of derived estimates over full-set of iterations
- The basis for simulations is the probability of being a true match: P(Match)
  - Calculated by logistic regression on true match status
    - Indicated by agreement on Social Security Number (SSN)

# Methodological Approach

*Basis of Simulation*

- The simulation determined which candidate pairs were selected as links
- Two pathways for accepting links:
  - Pathway I: Links meeting threshold: P(Match) > 0.9
  - Pathway II: Links to fill match quota remaining after Pathway I links are identified

# Methodological Approach

*Pathway I*

- For candidate pairs scoring above the linkage acceptance threshold, we included that pair into the set of potential links
  - In each iteration, for each potential link, a standard uniform random variable (RV) was drawn
    - if RV < P(Match) it was selected into accepted links

# Pathway I - Illustration

## *Example of Linkage Assignment Simulation**

| | | Iteration | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | 2 | | 3 | | 4 | | 5 | |
| Patient-NDI Pair | **P**(Match) | RV | Accepted Link: (RV < P(Match)) | RV | Accepted Link: (RV < P(Match)) | RV | Accepted Link: (RV < P(Match)) | RV | Accepted Link: (RV < P(Match)) | RV | Accepted Link: (RV < P(Match)) |
| Patient 1 – NDI 9312 | **0.9135** | *0.4170* | TRUE | *0.1198* | TRUE | *0.5835* | TRUE | *0.0867* | TRUE | *0.1324* | TRUE |
| Patient 8 – NDI 3421 | **0.9350** | *0.3853* | TRUE | *0.7207* | TRUE | *0.9804* | FALSE | *0.0201* | TRUE | *0.6378* | TRUE |
| Patient 23 – NDI 2704 | **0.9850** | *0.3849* | TRUE | *0.6738* | TRUE | *0.7100* | TRUE | *0.2355* | TRUE | *0.0060* | TRUE |
| Patient 31 – NDI 8728 | **0.9005** | *0.7176* | TRUE | *0.5331* | TRUE | *0.4372* | TRUE | *0.9180* | FALSE | *0.3635* | TRUE |
| Patient 52 – NDI 5216 | **0.9250** | *0.9863* | FALSE | *0.5712* | TRUE | *0.0888* | TRUE | *0.3243* | TRUE | *0.8421* | TRUE |

*These are hypothetical pairs—not actual data



NORC at the UNIVERSITY of CHICAGO

# Pathway II

- Generally, N(Links) < N(Matches)
- Since we are computing mean number of deaths based on matches to NDI, this would lead to downward bias:

$$\widehat{DeathRate} < DeathRate$$

- Can Estimate $N_{Matches}$ = ($N_{Links}$ / Linkage Sensitivity)
  - Linkage Sensitivity estimated using test deck
  - Correct estimate by drawing from unlinked candidate pairs, PPS on P(Match) until quota filled.
    - $Quota = \widehat{N_{Matches}} - N_{Links}\ (Pathway\ 1)$



NORC at the UNIVERSITY of CHICAGO

# Calculation of Statistics from Iterations of Simulation

- Pathway I and Pathway II results can then be combined to calculate standard. deviation of the mean number of deaths 30 days post discharge

| Hospital | Iteration (1 – 200) | | | | | | Mean | Std. Dev. | 95% Conf. Int. Lower Bound | 95% Conf. Int. Upper Bound |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | … | | | | |
| 1 | 858 | 848 | 853 | 851 | 856 | | 850.8 | 3.9 | 844.0 | 857.0 |
| 2 | 963 | 966 | 963 | 964 | 955 | | 960.3 | 5.6 | 951.0 | 969.5 |
| 3 | 2,199 | 2,195 | 2,204 | 2,195 | 2,211 | | 2198.7 | 6.4 | 2187.0 | 2210.0 |
| 4 | 578 | 579 | 573 | 576 | 578 | | 575.9 | 3.0 | 571.0 | 580.5 |
| 5 | 1,412 | 1,414 | 1,418 | 1,417 | 1,422 | | 1415.5 | 4.8 | 1407.5 | 1423.0 |
| 6 | 658 | 657 | 663 | 661 | 662 | | 660.3 | 2.8 | 656.0 | 665.0 |

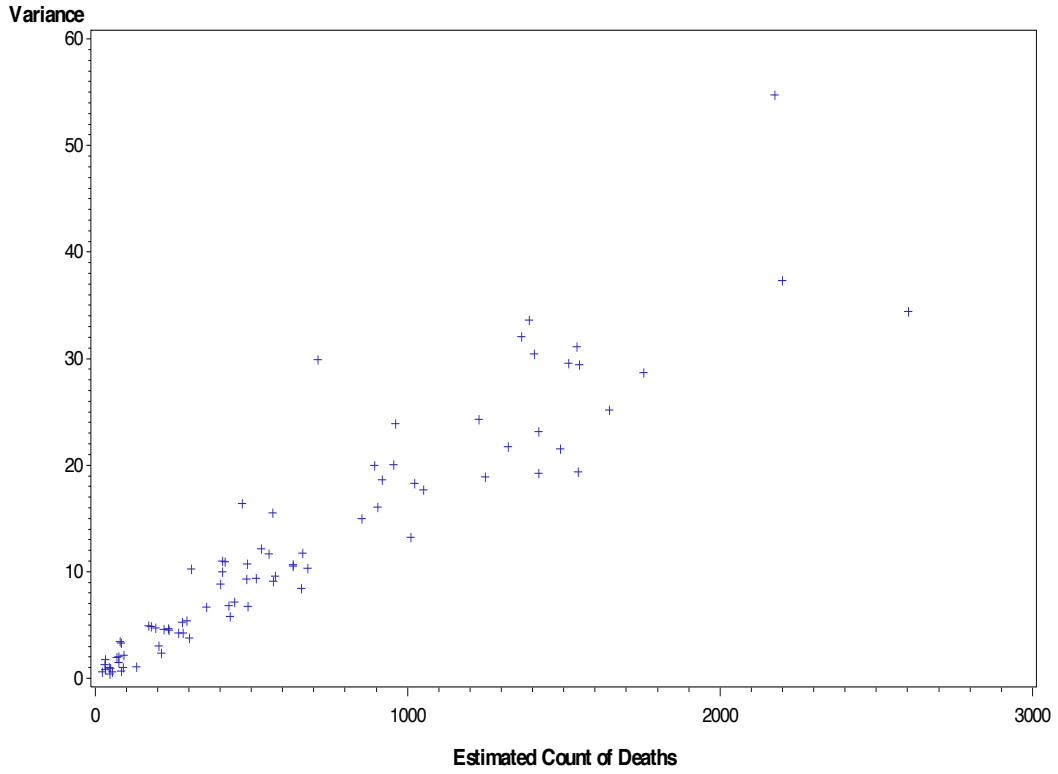


NORC at the UNIVERSITY of CHICAGO

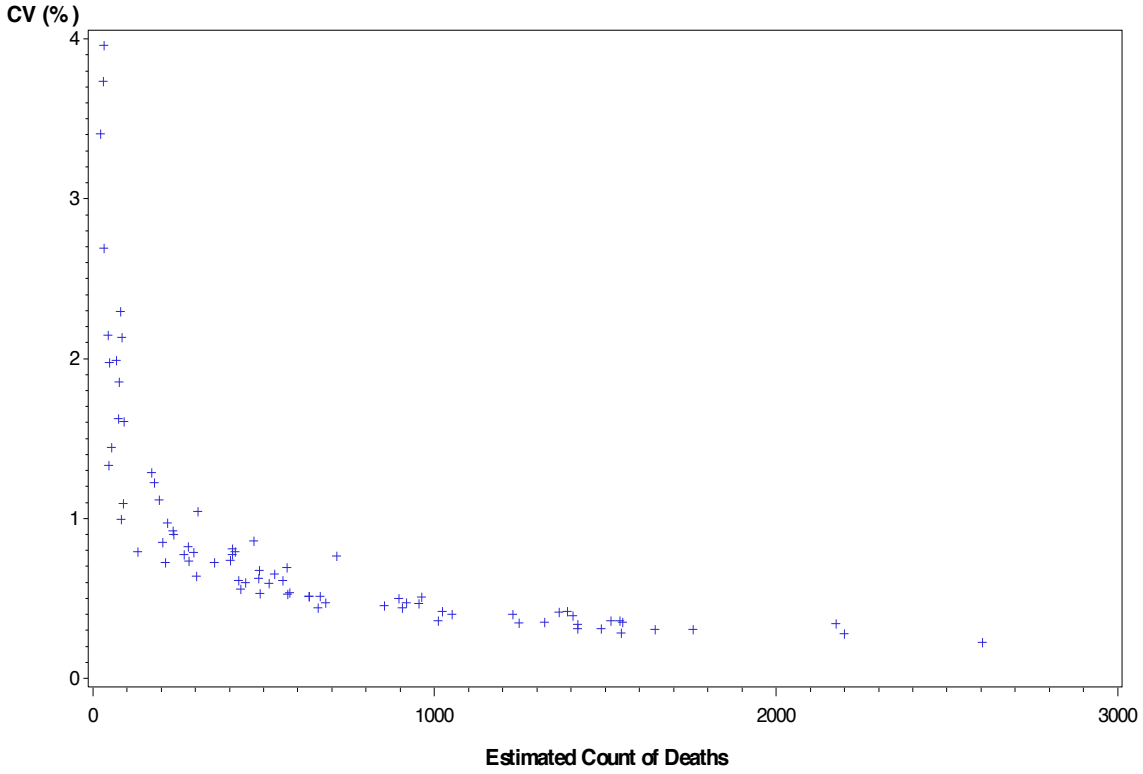
# Results

*Estimated Error*

## Variance vs. Mean # of Deaths
### Within 30 Days of Discharge




## Coeff. of Variation vs. Mean # of Deaths
### Within 30 Days of Discharge






NORC at the UNIVERSITY of CHICAGO

# Findings

- Variance increases proportionally with size of estimate
  - Appears to be some heteroscedasticity
- Coefficient of Variation decreases asymptotically
- Low Variability
  - 67.7% of links made deterministically in NHCS-NDI linkage
  - Cost of a linkage error is not high in this application

# Conclusions

- Simulation is a practical way to estimate effect of linkage uncertainty on derived estimates.
  - Can be combined with replicate weights to estimate total error.
- Mechanics of simulation need to be thought through carefully.
- Results may differ depending on number of links from deterministic approach
- Potentially, optimal cutoffs can be determined where variance is minimized.
  - Depends on estimates to be computed.

**Dean M. Resnick**
*Senior Data Scientist*
NORC at the University of Chicago
Resnick-dean@norc.org
301-634-9481

# Thank You!

NORC at the UNIVERSITY of CHICAGO

insight for informed decisions™