National Center for Health Statistics
Data Linkage

# Incorporating an Enhanced-Linkage Algorithm for Household Survey Data Linked to the NDI

Lisa B. Mirel

Board of Scientific Counselors
January 9, 2020

# Introduction

- Proposing to use an enhanced algorithm and document changes with the next release of the linked mortality data for all NCHS surveys (expected release date Q3 2020)

- New files will include:
  - Detailed information on enhanced linkage algorithm
  - Comparative analyses using the old and new algorithms

# Background I

- NDI algorithm was calibrated using NHANES I Epidemiologic Follow-up study (NHEFS)

- For past linked mortality files, the linkage group used a slightly modified NDI algorithm for linking NCHS survey data (NHIS and NHANES) to NDI
  - Accommodate SSN 4 data collection
  - Hispanic and Asian name alternate records

- 2017 PCORTF project to link NHCS to NDI resulted in development of enhanced linkage algorithm

# Background II

- Enhanced algorithm was developed and applied to the 2014 and 2016 NHCS data linked to the NDI

- Same enhanced algorithm was applied to household survey data linkage to NDI
  - Resulted in changes in assigned vital status for survey participants compared to previously conducted linkages by linkage group
  - Larger number of decedents are no longer deceased

# Enhanced Linkage Approach

- Linkage conducted in two passes:
  1. Deterministic match using SSN collected in the survey
     - Identifier fields such as name, state of residence, and date of birth are compared to validate
     - This dataset becomes the “test deck”
  2. Probabilistic matching techniques used to identify likely pairs using other identifiers (not SSN)
     - SSN is not used to create the match pool instead it is used to measure linkage accuracy

# Specifics: Probabilistic Techniques

- Possible pairs are scored according to Fellegi-Sunter (F-S) paradigm
- For each identifier, first name, year-of-birth, etc., M- and U-probabilities are computed
  - M-probabilities: rate of identifier agreement for matched pairs
  - U-probabilities: likeliness of a spurious agreement
  - Rare values (e.g., unusual names) have lower U-probabilities
- M- and U- probabilities are used to algebraically determine agreement and non-agreement weights according to F-S theory
- Weights for all identifiers summed to produce total pair weight

# Probability of a Match P(Match)

- Pair weights used to estimate P(Match): the probability that a given pair is an actual match (i.e., paired records represent same person)

- Pairs with estimated P(Match) above a threshold were considered matches all those below were assumed alive

# Selection of Best Pair as a Match

- When a survey record has been linked to multiple NDI records
  - The linked pair having the highest probability of being a match is accepted
  - Deterministic links are assumed to have probability of 1 and are always selected over links established from probabilistic search

# Assess Quality of Matches

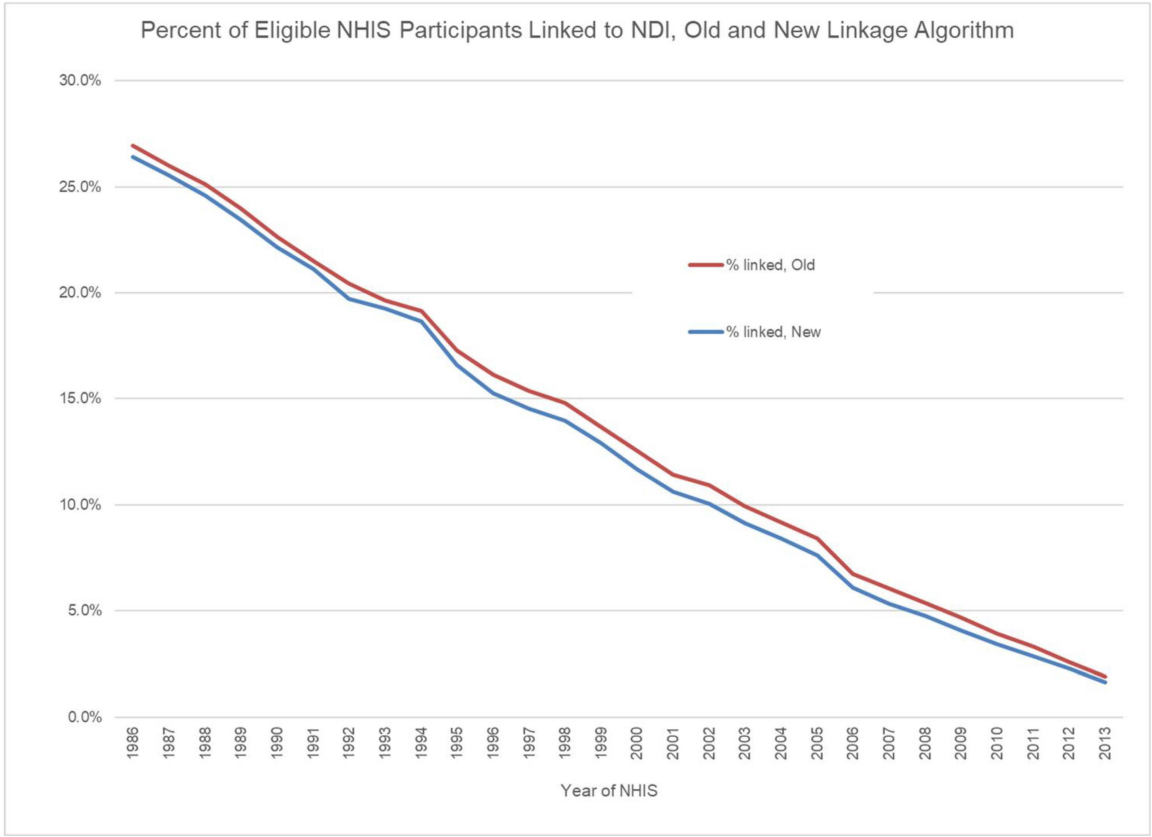
- Type I (false positive) and Type II (false negative) errors were calculated using the test deck in order to assess quality of matches developed

- Results: Highly accurate linkage results (low type I and type II errors)
  - Deterministic matches (pass 1) represent 2/3 of total matches (assume a zero error rate with deterministic approach)

- For all surveys combined:
  - Type I error rate = 1%
  - Type II error rate = 2%

# Comparing the Two Approaches

# Percent of Eligible NHIS Participants Linked to NDI: Old and New Linkage Algorithm



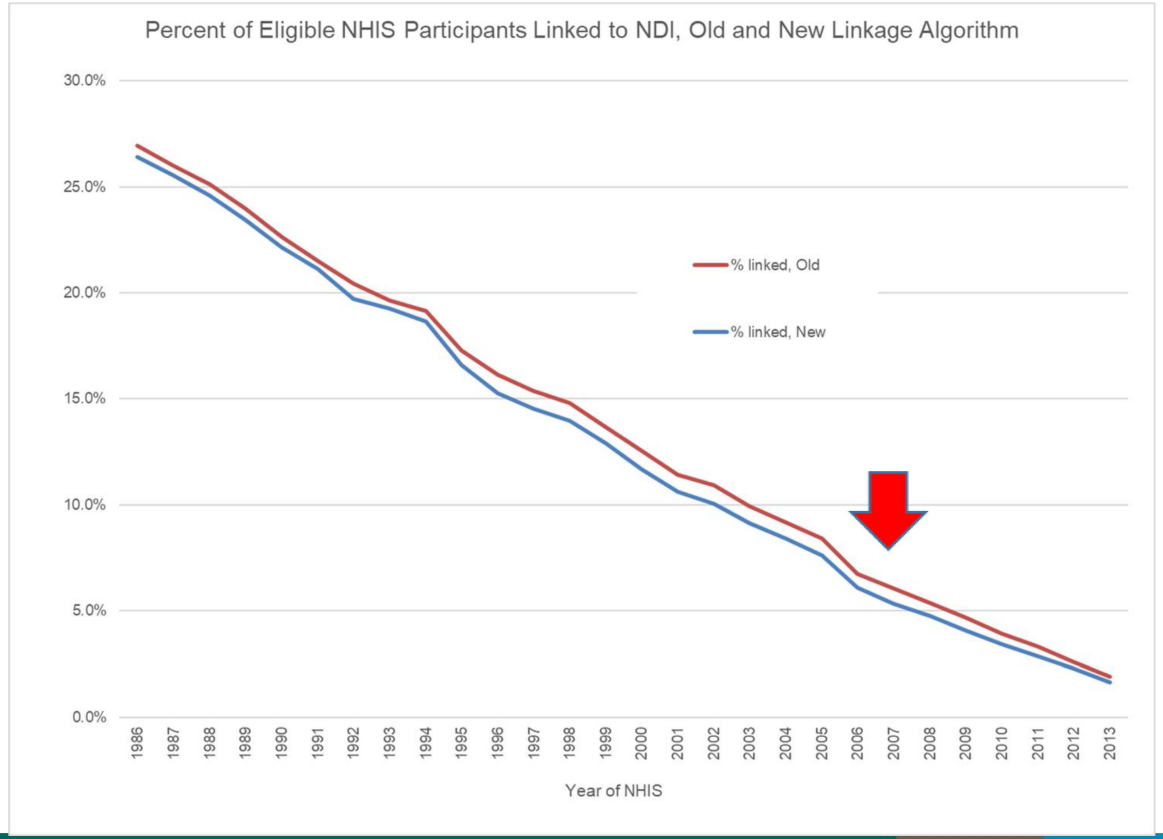
# All NHIS years combined: 1986-2013

**NHIS 1986-2013**

| | | New Linkage Algorithm | | |
|---|---|---|---|---|
| | | Assumed Alive | Assumed Deceased | Total |
| Old Linkage Algorithm | Assumed Alive | 3,892,452 | 19,907 | 3,912,359 |
| | Assumed Deceased | 48,862 | 676,355 | 725,217 |
| | Total | 3,941,314 | 696,262 | 4,637,576 |

- 68,769 out of 4,637,576 (1.5%) will have a different outcome with the new algorithm compared to old
- 93.2 % of old matches (676,355/725,217) in concordance with new matches
- 97.1% of new matches (676,355/696,262) in concordance with old matches

# Percent of Eligible NHIS Participants Linked to NDI: Old and New Algorithm



# NHIS: SSN9 Results

**NHIS 1986-2006**

| | | New Linkage Algorithm | | |
|---|---|---|---|---|
| | | Assumed Alive | Assumed Deceased | Total |
| Old Linkage Algorithm | Assumed Alive | 1,652,177 | 9,375 | 1,661,552 |
| | Assumed Deceased | 22,388 | 328,634 | 351,022 |
| | Total | 1,674,565 | 338,009 | 2,012,574 |

- 31,763 out of 2,012,574 (1.6%) will have a different outcome with the new algorithm compared to old
- 93.6 % of old matches (328,634/351,022) in concordance with new matches
- 97.2% of new matches (328,634/338,009) in concordance with old matches

# NHIS: SSN4 Results

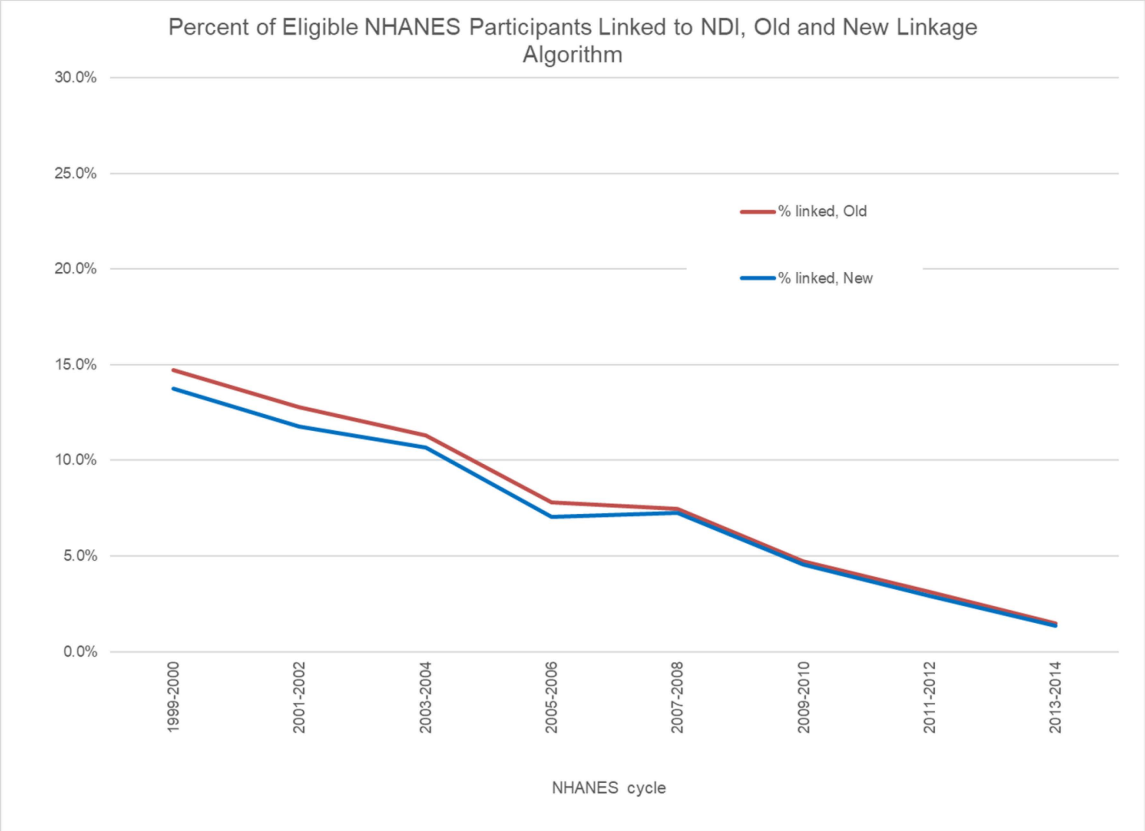
**NHIS 2007-2013**

| | | New Linkage Algorithm | | |
|---|---|---|---|---|
| | | Assumed Alive | Assumed Deceased | Total |
| Old Linkage Algorithm | Assumed Alive | 588,098 | 1,157 | 589,255 |
| | Assumed Deceased | 4,086 | 19,087 | 23,173 |
| | Total | 592,184 | 20,244 | 612,428 |

- 5,243 out of 612,428 (0.7%) will have a different outcome with the new algorithm compared to old
- 82.4 % of old matches (19,087/23,173) in concordance with new matches
- 94.3% of new matches (19,087/20,244) in concordance with old matches

*Note: starting in 2007 only the last 4 digits of SSN were collected from the Sample Adult. All eligible participants were linked.*

# Percent of Eligible NHANES Participants Linked to NDI: Old and New Linkage Algorithm



# Continuous NHANES: 1999-2014

**NHANES 1999-2014**

| | | New Linkage Algorithm | | |
|---|---|---|---|---|
| | | Assumed Alive | Assumed Deceased | Total |
| Old Linkage Algorithm | Assumed Alive | 75,283 | 95 | 75,378 |
| | Assumed Deceased | 513 | 6,013 | 6,526 |
| | Total | 75,796 | 6,108 | 81,904 |

- 608 out of 81,904 (0.6%) will have a different outcome with the new algorithm compared to old
- 92.1 % of old matches (6,013/6,526) in concordance with new matches
- 98.4% of new matches (6,013/6,108) in concordance with old matches

## Class and Score

How were the previous links selected?

Defining Class and Score: Old algorithm

| Class | Match | Score and Determination |
|---|---|---|
| 1 | At least 8 (of 9) or 4 (of 4) digits of SSN, first name, middle initial, last name, birth year (+/- 3 years), birth month, sex, and state of birth. | All are Matches |
| 2 | At least 7 (of 9) or 4 (of 4) digits of SSN at least 5 more of the following items: first name, middle initial, last name, birth year (+/- 3 years), birth month, sex, and state of birth. | Score >= 44 then considered a Match |
| 3 | A: SSN is unknown, but last name matched and at least 7 of the following items agreed: first name, middle initial, last name, birth year (+/- 3 years), birth day, sex, race, marital status and state of birth.<br><br>B: SSN was known but 3 or more (of 9) and 1 or more (of 4) digits did not agree, but at least 8 of the following items agreed: first name, middle initial, last name, birth year, birth day, sex, race, marital status, and state of birth. Switched from Class 5 to Class 3 – SSN was recorded incorrectly or spouse's SSN was recorded. Scores adjusted to reflect that SSN was missing (assigned value of 0). | Score >= 45 then considered a Match |
| 4 | SSN was unknown on either the NCHS survey submission record or the NDI record and fewer than 8 of the items listed in Class 3 matched. | Score >= 42 then considered a Match |
| 5 | SSN was present but fewer than 7 (of 9) or 4 (of 4) digits on SSN agreed | None are Matches |

# Class and Score: All Surveys Combined

| | Class | Number | Total | Percent | Mean Score |
|---|---|---|---|---|---|
| Old only | 1 | 10 | 29,727 | 0.0 | 78.9 |
| Old only | 2 | 679 | 29,727 | 2.3 | 54.3 |
| Old only | 3 | 9,986 | 29,727 | 33.6 | 50.3 |
| Old only | 4 | 17,614 | 29,727 | 59.3 | 42.8 |
| Old only | 5 | 1,438 | 29,727 | 4.8 | 11.5 |
| Old and New agree | 1 | 217,067 | 470,695 | 46.1 | 89.0 |
| Old and New agree | 2 | 139,198 | 470,695 | 29.6 | 75.1 |
| Old and New agree | 3 | 93,808 | 470,695 | 19.9 | 60.1 |
| Old and New agree | 4 | 18,660 | 470,695 | 4.0 | 49.8 |
| Old and New agree | 5 | 1,962 | 470,695 | 0.4 | 14.0 |

~93% of old only are class 3 and 4

*Class 5 deaths indicate death from a non-NDI source*

# Effects on Inference: Old and New Linkage Algorithms

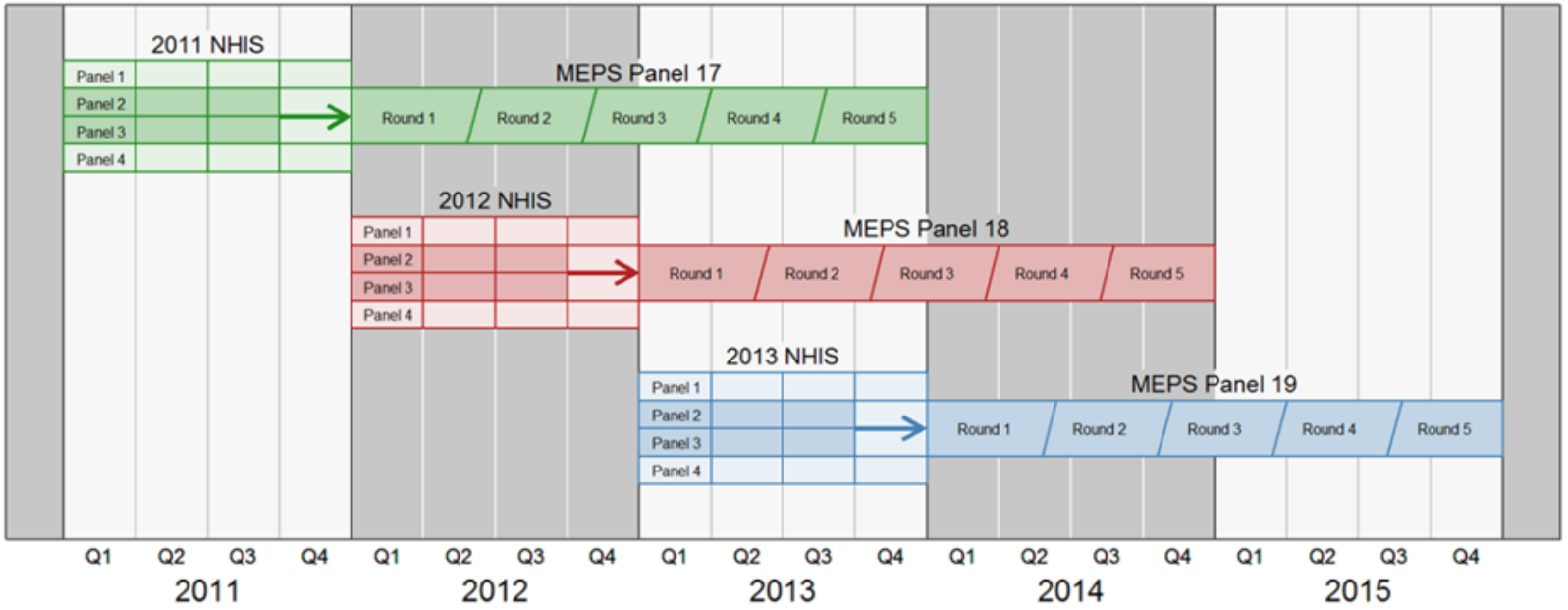
- Survival models run using old and new algorithm
- Models included age, sex, race/ethnicity, education, marital status and region
- Compared hazard rates from two approaches
  - For all cause and cause specific mortality the % differences of hazard rates from the survival models were <=5% except for Hispanics which were greater than 10%

# Validation Checks for New Algorithm

# Medical Expenditure Panel Survey (MEPS) Analysis

- Current NCHS mortality linkage doesn’t have a “gold standard” available for assessment
- Survey-reported death data are an ideal comparator
- Assessment of mortality linkage algorithm possible using the MEPS

# MEPS Analysis (cont.)

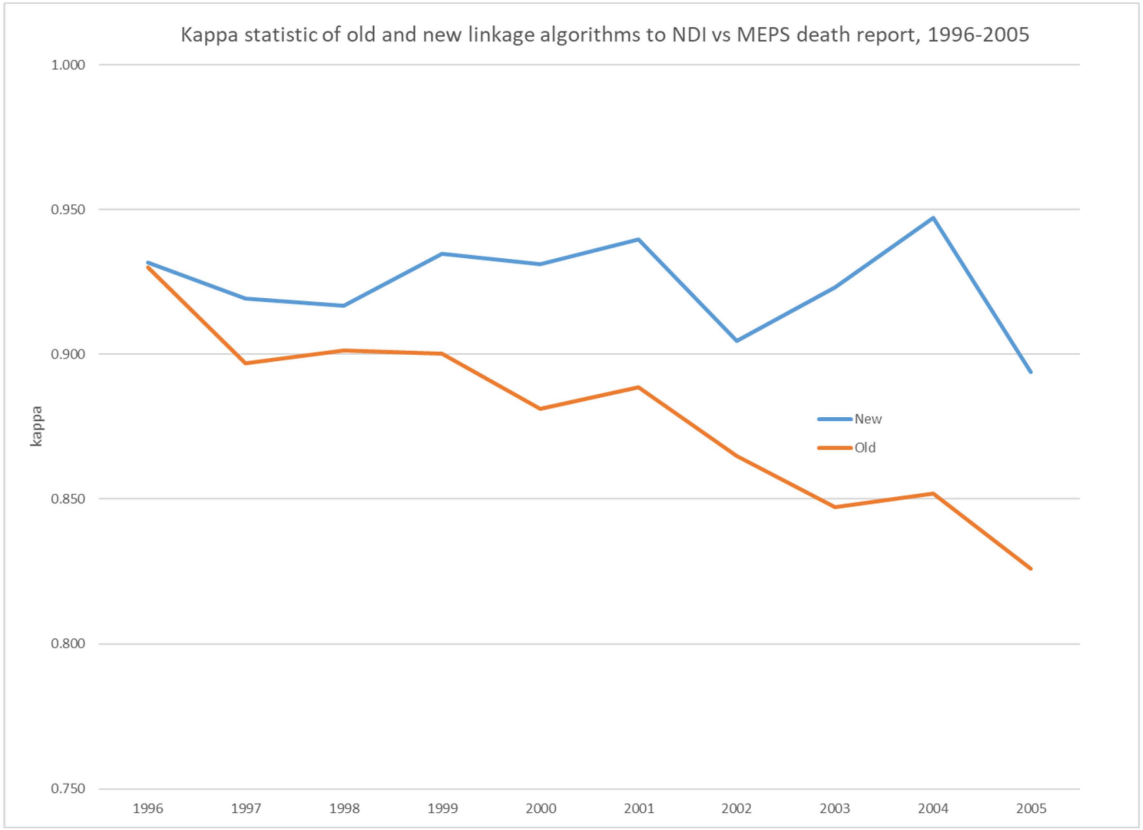



MEPS follows NHIS participants over time; during MEPS data collection may determine that a participant has died and in what year

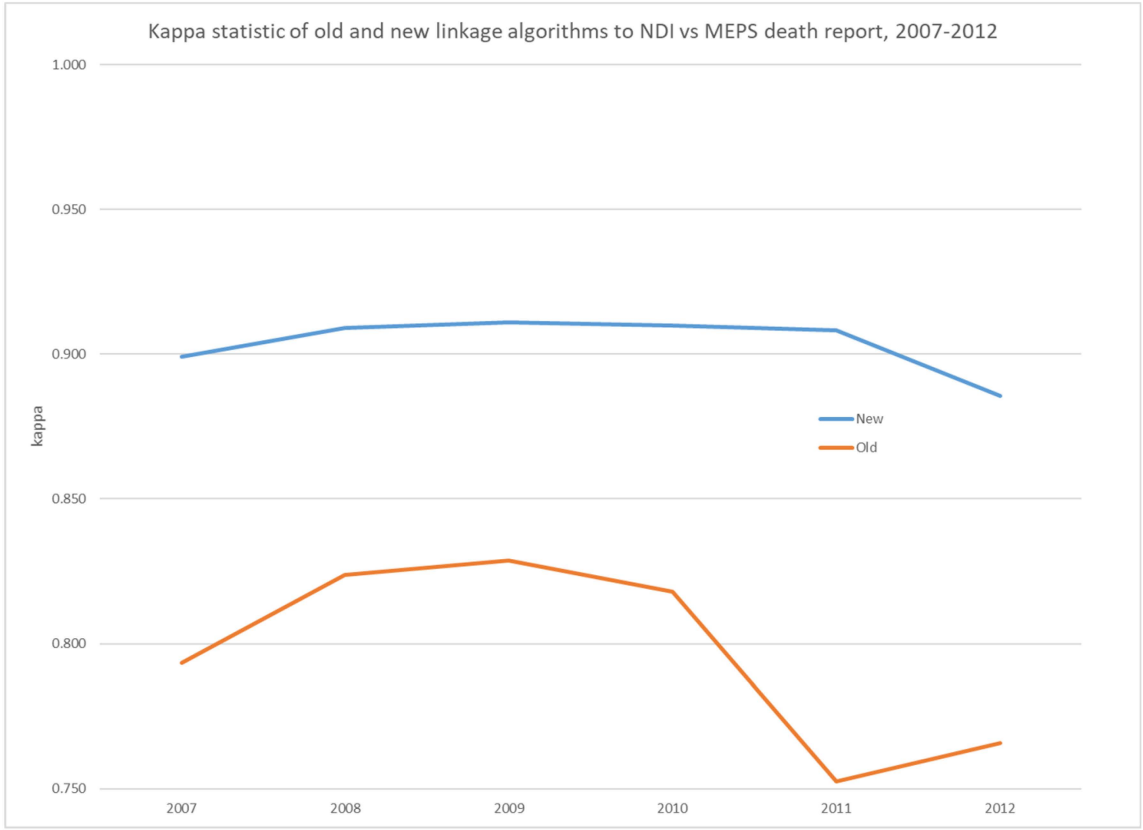
## MEPS Analysis (cont.)

- If the participant died, the mortality status as reported in MEPS, becomes a proxy for “gold standard”
- If the date of death was greater than the MEPS round date the participant was censored (assumed alive for the kappa calculation) for that year of NHIS

# Kappa Statistic of Old and New Linkage Algorithm: NDI vs MEPS death report: 1996-2005



# Kappa Statistic of Old and New Linkage Algorithm: NDI vs MEPS death report: 2007-2012



# 2014 NHCS Linkage: Standard NDI Algorithm* vs New Algorithm

**2014 NHCS**

| NDI Linkage Algorithm | | New Linkage Algorithm: Assumed Alive | New Linkage Algorithm: Assumed Deceased | Total |
|---|---|---|---|---|
| | Assumed Alive | 3,386,651 | 17,737 | 3,404,388 |
| | Assumed Deceased | 3,957 | 149,941 | 153,898 |
| | Total | 3,390,608 | 167,678 | 3,558,286 |

- 21,694 out of 3,558,286 (0.6%) have a different outcome with the new algorithm compared to NDI
- 97.4 % of old matches (149,941/153,898) in concordance with new matches
- 89.4% of new matches (149,941/167,678) in concordance with old matches

*Note: of the 3,957 NDI-only links, 97.6% were class 4 (0.13% were class 3)*

**2014 NHCS was run through the standard NDI algorithm prior to 2017 PCORTF project*

# QC check 2014 NHCS

- 32,763 patients in the 2014 NHCS had a discharge status of deceased on their hospital record
- Of the 32,763 with a discharge status of deceased:
  - New algorithm linked 31,723 (96.8%)
  - NDI algorithm linked 30,530 (93.2%)

**2014 NHCS patients with a discharge status as deceased**

| | | New Linkage Algorithm | | |
|---|---|---|---|---|
| | | Assumed Alive | Assumed Deceased | Total |
| NDI Linkage Algorithm | Assumed Alive | 864 | 1,369 | 2,233 |
| | Assumed Deceased | 176 | 30,354 | 30,530 |
| | Total | 1,040 | 31,723 | 32,763 |

# Conclusions

- For HH surveys: concordance between the two methods is high overall (~94%)

  - New deaths for previously matched years – explained by improved matching techniques. Relatively small numbers when compared with total eligible (1986-2013 NHIS=19,907 (0.4%), 1999-2014 NHANES =95 (0.1%), NHANES III=48 (0.1%))

  - Previous decedents no longer considered deceased – explained by improved matching techniques. Larger numbers when compared with total eligible (1986-2013 NHIS=48,862 (1.1%), 1999-2014 NHANES =513 (0.6%), NHANES III=616 (1.8%))

# Conclusions (cont.)

- Old algorithm was based on what we knew at the time (NHEFS was used for validation)

- New algorithm
  - Aligns with outside sources for validation (MEPS and NHCS discharge status)
  - Improves shortcoming with certain demographic groups

# Implications for Dissemination

- For HH surveys linked to NDI: plan to use newly enhanced algorithm for updated linked mortality file production, beginning in January 2020 with 2018 NDI data

- Mitigate user concern over different results from previous mortality releases by publishing comparative analyses of the two approaches

- Question for the BSC:

  How should we proactively communicate with new and current users about the changes?

# Appendix

# NHANES III: 1988-1994

**NHANES III**

| | | New Linkage Algorithm: Assumed Alive | New Linkage Algorithm: Assumed Deceased | Total |
|---|---|---|---|---|
| Old Linkage Algorithm | Assumed Alive | 25,560 | 48 | 25,608 |
| | Assumed Deceased | 616 | 7,735 | 8,351 |
| | Total | 26,176 | 7,783 | 33,959 |

- 664 out of 33,959 (2.0%) will have a different outcome with the new algorithm compared to old
- 92.6 % of old matches (7,735/8,351) in concordance with new matches
- 99.4% of new matches (7,735/7,783) in concordance with old matches

# NHANES Feasibility Longitudinal Study

- Old algorithm falsely assigned deceased status to 5 people in the NHANES feasibility longitudinal study

  NHANES interviewed these 5 as part of the longitudinal study

- New algorithm assigned assumed alive status to all 5 of these people