Using an Automated Computer-Based Algorithm to Increase the Efficiency of Selecting the Best Match From National Death Index Linkage Results

Kirk D Midkiff, Brian Calingaert, Patricia Tennis, David Harris, Elizabeth B Andrews

RTI Health Solutions, Research Triangle Park, NC, United States

CONFLICT OF INTEREST STATEMENT

- The authors are employees of RTI Health Solutions. The parent study, from which data were gathered and analyzed for this poster, is funded by GlaxoSmithKline (WEUSRTP4942). The authors have an independent right to publish.

INTRODUCTION

- The National Death Index (NDI) maintains a national, central, computerized repository of death records via collaboration with state vital statistics offices.
- NDI was established as a resource for epidemiologists and other investigators interested in mortality, for deaths occurring after 1978.
- Deaths are added to the NDI master file annually; they are typically available approximately 12 months after the end of a calendar year.[1]
- Large cohorts (occupational, research) can be linked with the NDI to determine fact, date, and cause of death. Each NDI death record that matches the user record on at least one of seven minimum matching criteria is returned to the researcher as a possible match, often resulting in many possible matches.
- For each possible matching NDI death record, NDI indicates which variables matched between the two records, provides a probability score based on the number of matching variables, and indicates whether the possible match is considered a "true match; assumed dead" ("true match") by NDI criteria.
- NDI may return more than one "true match" for a single individual, and many of the possible matching NDI death records returned may not be true matches. Therefore, researchers often must manually review possible matching death records to determine the match.
- Use of an automated algorithm for evaluating NDI results has been recommended by other researchers when manual review and adjudication of multiple possible matching records is not possible.[2] NDI provides cause of death information for each death record they consider a true match to a user record or for a death record ranked first in the list of possible matches to a user record.
- As part of an ongoing drug safety study (see blue insert), data files containing information on hundreds of thousands of patients are linked with the NDI to determine the fact, date, and cause of death.

Asthma Safety Observational Study (ASSESS)

- Objective: To assess the available sample size and precision for evaluating whether long-acting beta-agonist use in combination with an inhaled corticosteroid is associated with an increased risk of asthma mortality
- Design: Retrospective cohort study using claims data or electronic medical records from multiple health insurers analyzed under a distributed data approach to identify the study population and characterize person-time of exposure
- Population: Patients aged 4 years or older fulfilling a study definition of persistent asthma
- Endpoint: Asthma death identified via linkage with NDI

- RTI Health Solutions serves as the coordinating center for ASSESS, providing oversight to 11 participating data partners in preparation and submission of data files to NDI for the study linkage.

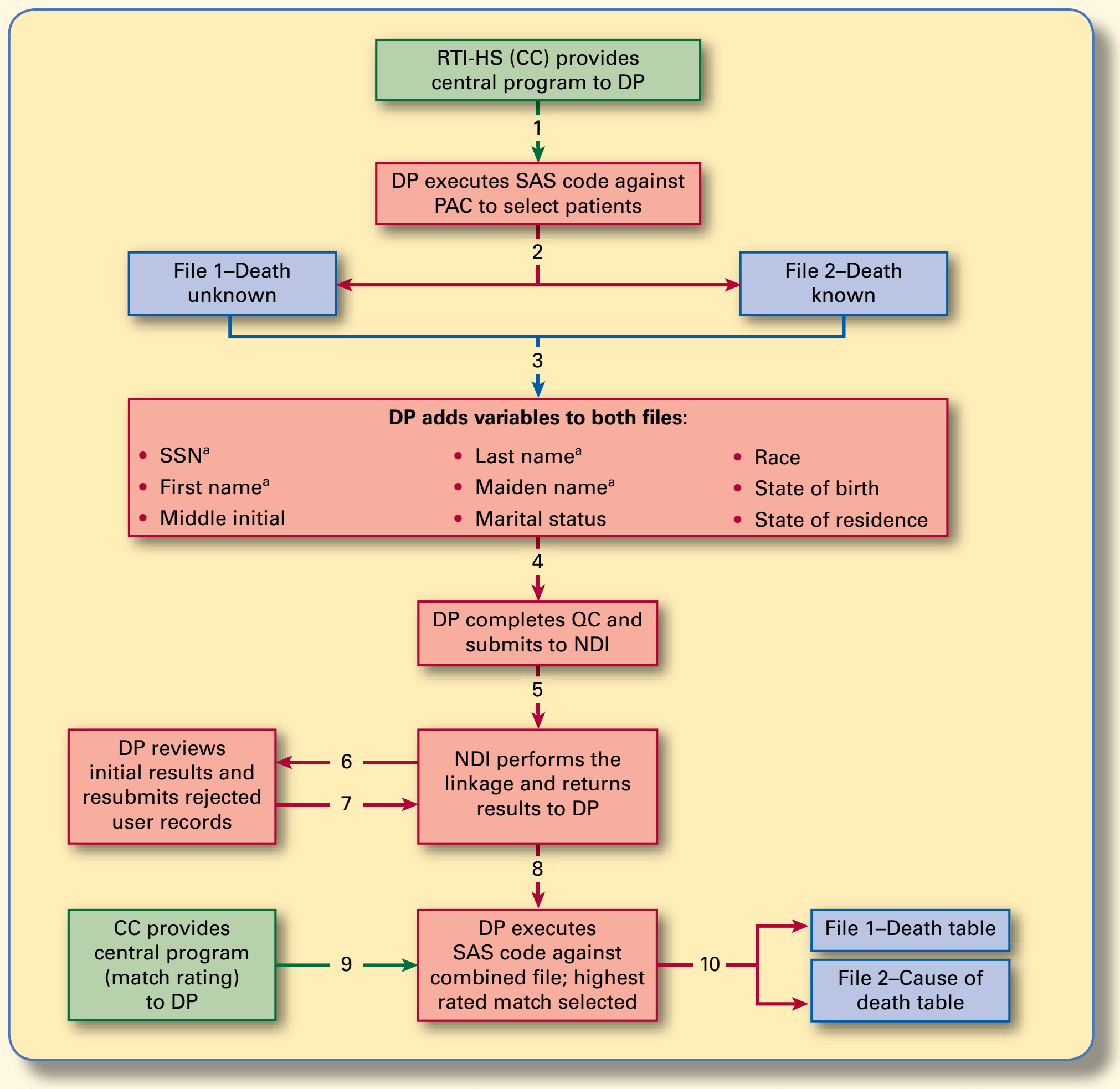
OBJECTIVE

- To describe and evaluate an automated algorithm to be applied to the results returned from NDI to select the single most likely true match.

METHODS

- Each data partner constructed a patient-level general asthma cohort dataset from claims or electronic medical records data, using a central common format (common data model).
- The coordinating center developed and distributed central programs to select the persistent asthma cohort from the general asthma cohort and to select patients from the persistent asthma cohort whose information would be submitted to NDI.
- The coordinating center developed a process (Figure 1) for preparation and quality control of the NDI submission files, obtained administrative approvals, created a procedure guide, and conducted training with staff from collaborating data partners.

Figure 1. NDI File Preparation and Linkage Process



RTI-HS (CC) provides central program to DP

1

DP executes SAS code against PAC to select patients

2

File 1–Death unknown

File 2–Death known

3

DP adds variables to both files:

- SSN[a]
- First name[a]
- Middle initial
- Last name[a]
- Maiden name[a]
- Marital status
- Race
- State of birth
- State of residence

4

DP completes QC and submits to NDI

5

NDI performs the linkage and returns results to DP

6

7

DP reviews initial results and resubmits rejected user records

8

DP executes SAS code against combined file; highest rated match selected

9

CC provides central program (match rating) to DP

10

File 1–Death table

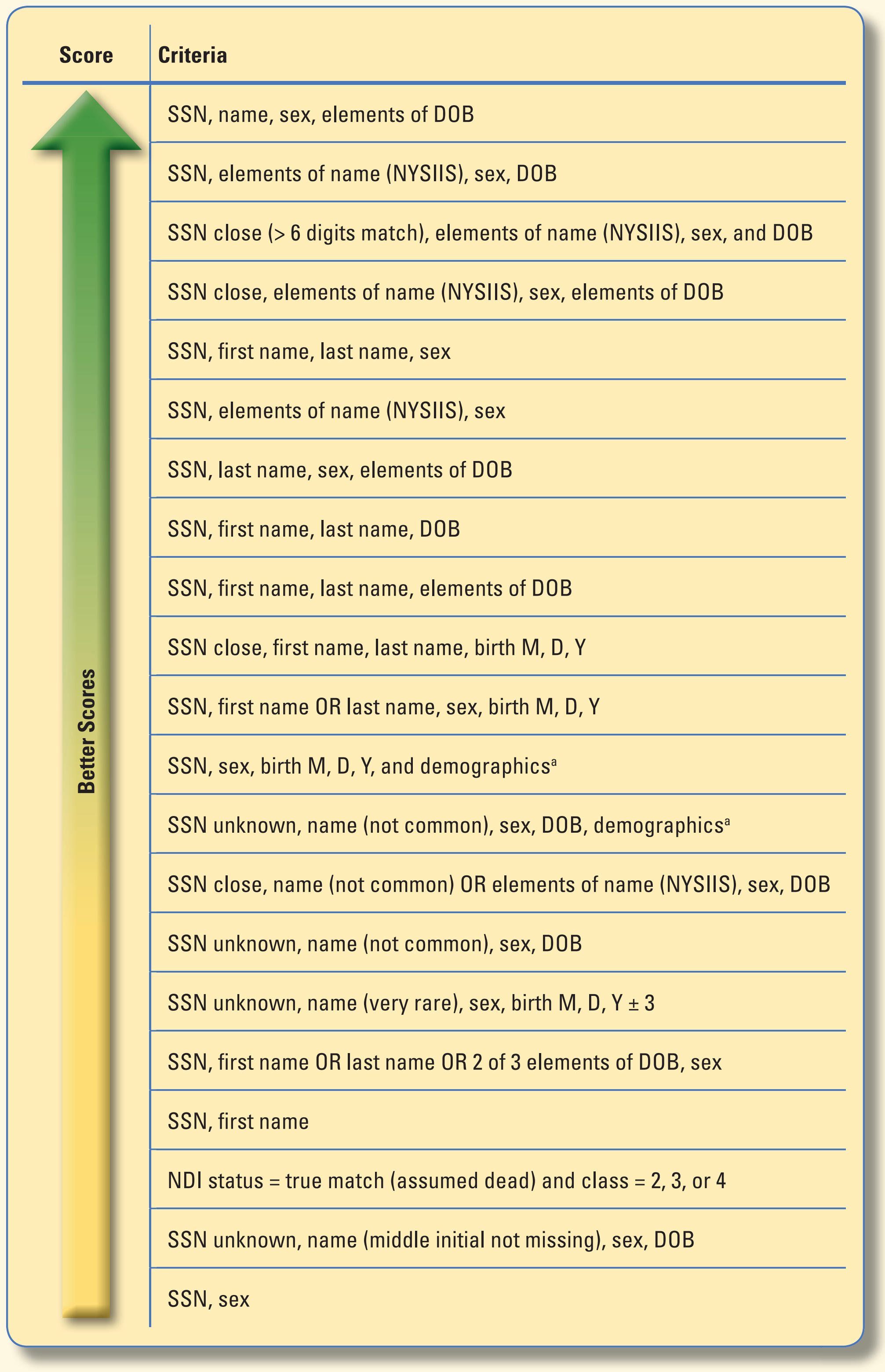
File 2–Cause of death table

CC = coordinating center; DP = data partner; PAC = persistent asthma cohort; QC = quality control; SSN = social security number.

[a] Variables important for NDI to establish a link between a user record and a death record; these include both date of birth (DOB) and sex, which were already part of the file(s) generated during step 2.

- After completing the linkage, NDI returned a file (combined file) to each data partner with possible matching NDI death records. The file may have included more than one possible matching record for each individual user record submitted.
- The coordinating center adapted a common automated algorithm from one used by state cancer registries, to select the single most likely match that met specific minimum criteria from among NDI death records contained in the NDI combined file.
  - The algorithm evaluated all possible matching records returned by NDI to select the single most likely match for each patient. This process was accomplished by assignment of a score, based on different combinations of variables that agree between the user record and NDI death record. The score indicated the likelihood that the returned death record is the true death record.
  - The algorithm was tested on simulated data and then on actual NDI results from two data partners prior to full implementation, to determine if results created by the algorithm were scored as expected.
  - Manual review of NDI results was not performed by the data partners; the algorithm categorized all matches or nonmatches.
- Figure 2 displays a summary of the criteria used in the automated algorithm for assigning the match-rating score from better scores to worse scores. If more than one NDI record was provided for a single user record, only the NDI record with the best score was kept.

Figure 2. Overview of Criteria Used for Selecting the Single Most Likely Match



Score	Criteria
	SSN, name, sex, elements of DOB
	SSN, elements of name (NYSIIS), sex, DOB
	SSN close (> 6 digits match), elements of name (NYSIIS), sex, and DOB
	SSN close, elements of name (NYSIIS), sex, elements of DOB
	SSN, first name, last name, sex
	SSN, elements of name (NYSIIS), sex
	SSN, last name, sex, elements of DOB
	SSN, first name, last name, DOB
	SSN, first name, last name, elements of DOB
	SSN close, first name, last name, birth M, D, Y
	SSN, first name OR last name, sex, birth M, D, Y
	SSN, sex, birth M, D, Y, and demographics[a]
	SSN unknown, name (not common), sex, DOB, demographics[a]
	SSN close, name (not common) OR elements of name (NYSIIS), sex, DOB
	SSN unknown, name (not common), sex, DOB
	SSN unknown, name (very rare), sex, birth M, D, Y ± 3
	SSN, first name OR last name OR 2 of 3 elements of DOB, sex
	SSN, first name
	NDI status = true match (assumed dead) and class = 2, 3, or 4
	SSN unknown, name (middle initial not missing), sex, DOB
	SSN, sex

D = day; M = month; NYSIIS = New York State Identification and Intelligence System phonetic code; Y = year.

[a] Demographics = race, marital status, state of birth.

- Table 1 displays the number of patients in the study cohort during the period of interest (2001-2010) and, of those, the number anticipated to be submitted to NDI (by vital status type).
- More than 1 million patients are in the study cohort; of those, 60% will be submitted to NDI to obtain fact, date, and/or cause of death.

Table 1. Patients in the Study Cohort and, of Those, the Number Submitted to NDI by Type of Vital Status

Data Partner	Patients in Study Cohort	Patients Submitted to NDI by Vital Status	
		Unknown	Known Deceased
1	348,477	198,006	0
2	311,281	204,661	0
3	191,670	76,340	3,489
4	46,136	29,141	1,209
5	43,002	42,671	0
6	28,568	17,621	0
7	15,675	14,788	0
8	13,156	7,231	225
9	11,525	10,471	0
10	8,977	8,721	0
11	5,164	3,516	0
Total	**1,023,631**	**613,167**	**4,923**

RESULTS

- Results for this analysis include data from two data partners (numbers 1 and 11) and are the most recent data available (July 31, 2012). Results were not available for patients submitted to NDI with a vital status of "known deceased."
- A total of 201,522 patients were submitted to NDI for death tracing with an unknown vital status (Table 2). Of the 61,249 patients returned by NDI with at least one possible match, there were almost three possible matches on average per patient, and a similar percentage had at least one true match according to the NDI (4.3%) and the automated algorithm (4.4%) criteria.

Table 2. Number of Patients and Possible Matches Returned by Algorithm Match Status

Item	Patient Vital Status Unknown
Total patients with information submitted[a]	201,522
Patients with at least 1 possible match returned by NDI	61,249 (30.4%)
Of those with at least one possible match:	
Average number of possible matches returned per patient	2.7 (range 1-50)
Number of patients with at least 1 true match according to NDI criteria[b]	8,722 (4.3%)
Number of patients with at least 1 true match according to automated algorithm criteria[c]	8,909 (4.4%)
Number of patients with a possible match that had more than one match in the same match-rating scoring strata (using automated algorithm criteria)	49 (0.6%)

[a] Results not available for all data partners due to delay in data availability of 2010 deaths by NDI.

[b] NDI status code = 1 (true match).

[c] Automated algorithm match-rating score from 1-15.

- Table 3 displays the results of the 61,249 patients who had at least one possible match returned by NDI. There was almost complete agreement between those selected as the best match by the automated algorithm and those selected as the best match based on NDI criteria (99.4%).
- The cause of death was unavailable for only 0.1% of the automated algorithm selected best matches.

Table 3. Level of Agreement Between the Automated Algorithm Selected Best Match and the NDI Selected Best Match, and the Availability of Cause of Death Information

Item	Patient Vital Status Unknown
Percentage of automated algorithm selected best matches[a] that agreed with the best matches according to the NDI criteria[b]	99.4%
Among automated algorithm selected matches (n = 8,909), percentage where cause of death was provided	99.9%
If cause of death was not provided for the algorithm selected match, percentage where cause of death was provided for a lower rated match	0.1%
Average number of ICD-10 cause-of-death codes returned for each patient selected by the automated algorithm	3.2 (range 0-14)

[a] Automated algorithm best match = lowest match-rating score from 1-15 (i.e., 1 is best possible match-rating score).

[b] NDI best match = NDI status code = 1 (true match); if more than one NDI record was returned having a status code = 1 for the same user record submitted, then the NDI record with the highest NDI probabilistic score was used as the best match.

CONCLUSIONS

- The agreement between the algorithm selected best match and the NDI selected best match was extremely high for patients with at least one possible matching NDI death record.
- In the current study involving submission of over 600,000 patients to the NDI, the burden imposed by manual review of results would be prohibitive.
- Application of a standard automated algorithm is preferable from a resource standpoint to manual adjudication of NDI results when there are large numbers of individuals with multiple possible matches.

REFERENCES

1. National Center for Health Statistics. About the National Death Index. Available at: http://www.cdc.gov/nchs/data_access/ndi/about_ndi.htm. Accessed October 5, 2011.
2. Wojcik NC, Huebner WW, Jorgensen G. Strategies for using the National Death Index and the Social Security Administration for death ascertainment in large occupational cohort mortality studies. Am J Epidemiol. 2010 Aug 15;172(4):469-77.

CONTACT INFORMATION

Kirk Midkiff, MPH
Director, Epidemiology

RTI Health Solutions
200 Park Offices Drive
Research Triangle Park, NC 27709

Telephone: +1.919.541.6638
Fax: +1.919.541.7222
E-mail: kmidkiff@rti.org

Presented at: 28th International Conference on Pharmacoepidemiology & Therapeutic Risk Management
August 23-26, 2012
Barcelona, Spain