

Final Report

The Wisconsin CVD Surveillance Data Pilot Project

Matching Mortality and Inpatient Records with Two Methodologies

The Heart Disease and Stroke Prevention Program
Wisconsin Department of Health Services, Division of Public Health

Richard Miller
Office of Health Informatics

Herng-leh (Mike) Yuan
Bureau of Community Health Promotion

1. Project Summary

This project pilots a new data resource for the surveillance of cardiovascular disease (CVD) and analysis of treatment outcomes by matching inpatient discharge records to death certificates. Rather than limit the actual linkage to CVD patients or deaths, we worked with all patients and all deaths.

We applied and evaluated two alternative methodologies for record matching: deterministic and probabilistic record linkage methods. We first used deterministic linkage techniques to de-duplicate three years of statewide inpatient discharge records and then to link the resulting patients to mortality records from the same period. We then replicated that effort by developing models and parameter estimates using the advanced probability-based methodology implemented by the *LinkSolv* software product. Finally, we developed analyses that compared and evaluated the results of both methods.

Valid and reasonably complete matches between mortality and hospital patient records may be created by either deterministic or probabilistic linkage methodologies. The results promise new data resources for the surveillance and analysis of CVD and other chronic diseases. For example, survival rates for patients with different demographics, co-morbidities or procedures could be compared and evaluated.

2. Linkage Methodologies

If information uniquely identifying the same individual in two datasets were perfectly captured and recorded, then a “**deterministic** record linkage methodology” would be reliable. Pairs of records could be compared for exactly matching identifying information and exact matches would *determine* true record matches.

In real world data systems, of course, uniquely identifying data elements are not often available. Further, recorded data on names, dates, addresses and other identifying information are subject to many influences that may cause differences between records and missing values in records. A “**probabilistic** record linkage methodology” recognizes that a pair of records with some matching elements – or even nearly-matching elements - has some *probability* of being a “true match.” The methodology uses statistical tools to estimate and evaluate that probability.

Deterministic record linkage. We first asked: How well can “a patient” be defined or “determined” as a unique value of available combined data elements? Before inpatient records can be linked to mortality reports, they must first be de-duplicated so that one record per patient is compared to death reports.

We did that inpatient record linkage with alternative deterministic algorithms and evaluated the results. One combination of elements was found to create reasonably complete and valid de-duplicated patients from the inpatient records. Inpatient records

with exactly matching values of the patient's initials, date of birth, gender and ZIP code were 'determined' to be for the same patient.

Next we applied the deterministic methodology to matching to mortality records. Two sets of iterative deterministic algorithms were developed to link (1) in-hospital death and inpatient records and (2) other deaths and inpatient records. Records with exact matches on the strongest combination were identified and set aside. The residual records then were tested for matches on the next-strongest combination, and so on. Ultimately about 66 percent of 2006-08 Wisconsin resident and occurrence mortality reports were matched to a Wisconsin resident patient in a Wisconsin hospital during the same period.

Probabilistic record linkage. We used the *LinkSolv* product to develop probabilistic linkage models. This software applies and extends rigorous statistical procedures by taking into account the many real-world situations which complicate the assumptions of more elementary probabilistic algorithms, such as missing data, correlated agreements and disagreements between data fields, recording errors, and so on. LinkSolv identifies those record matches with a user-defined high probability of being a true match but also imputes a small number of possible matches with a lower probability due to missing and incorrect data.

Following experimentation with more elaborate models, we settled on a fairly streamlined set of models that used initials, encrypted last name, gender, last four SSN digits, date of birth, and first three ZIP digits. We simplified the task by using only patients whose last hospitalization was in 2006 and by excluding birth-related hospitalizations.

As would be expected, the probabilistic linkage process identified a somewhat greater number of record linkages than did the deterministic approach. However, each method produced generally similar results. Analyses of discrepant results indicate ways that the more-approachable deterministic models could be improved. Notably, further work should evaluate deterministic algorithms using the last four SSN digits instead of the full SSN and first three ZIP digits instead of the full ZIP.

3. Data Analysis and Dissemination Plan

It is tempting to continue to refine and evaluate the linkage procedures. However, the potential utility of linked patient-mortality records as a new data resource argues for moving quickly to exploiting the analytical possibilities.

One of us (Yuan) is the Wisconsin epidemiologist for the Heart Disease and Stroke Prevention Program. We will soon extract the set of records with CVD-related inpatient diagnoses and/or CVD-related deaths. Among the considerable analytic issues, we are first planning to focus on three.

1. Among hospital inpatients with CVD-related diagnoses, what are the survival rates going out up to three years? How does this vary by diagnosis category?
2. Among CVD-related deaths, how many of the deceased were an inpatient within three years of their death? For what conditions were they treated? What co-morbidities are evidenced in their patient records? How do these patterns vary for specific underlying causes of death?
3. Among hospital patients with particular surgical procedures related to CVD diagnoses, what are the procedure-specific survival rates going out up to three years?

The work conducted for this pilot project has been described in a detailed working paper that is available from the authors and will be presented at the annual CSTE conference in June, 2011.

4. Status of All Project Objectives and Lessons Learned

The original project objectives were formulated as follows:

“The Wisconsin Cardiovascular Disease Surveillance Data Pilot Project will match records from three databases:

1. Wisconsin Vital Records’ death-certificate-based mortality files for 2008. There were 11,272 deaths with an underlying cause of some CVD among Wisconsin residents that year.
2. Wisconsin hospital inpatient discharge records for all Wisconsin residents discharged from a Wisconsin hospital during the three years 2006-2008. This well-established, high-quality database has about 63,000 discharges each year with principal diagnoses of heart disease. We will link discharges for the same person.
3. The Wisconsin Ambulance Run Data System (WARDS) has EMS ambulance run records for almost 300 Wisconsin ambulance services with about 250,000 runs for each year from 2007 on. These data are not yet complete statewide, but they do include 60% of Wisconsin’s services. We will link EMS services to the same person over time.”

These objectives have been significantly modified as work on the project advanced our conceptual and practical understanding.

- We dropped our ambitions of working with ambulance run data fairly early in the project. The coverage is still spotty statewide, the program had considerable technical difficulties exporting usable datasets, and the time required to incorporate these data into the other linkage efforts was prohibitive.
- We quickly realized that the most recent year of a patient’s hospitalization is the proper anchor for analyses, rather than the year of death. Therefore, we used deaths over the period 2006-08, which was the most current period when the work

began. We also included all deaths, without any screen for CVD-related mortality.

- We used the same three years of hospital inpatient discharge records in the deterministic linkage process and we used all inpatients. The initial step of de-duplicating patients is critical to both methodologies, so using all three years is important for identifying both the patients and the most recent inpatient stay.
- We decided to take advantage of the smaller subsets in both files that could be identified as in-hospital deaths. These should have high rates of record matches and indeed were found to be readily linked.
- The processing required by probabilistic linkage with *LinkSolv* was simplified in two ways that have little impact on results. First, only patients whose most recent stay occurred in 2006 were matched. Analyses of survival rates and so on require equal exposures to the possibility of death, so 2007's patients would have to be matched to 2007-09 deaths, etc. Second, the large number of patients who are mothers and their newborns may be dropped from the patient records being matched. These simplifications will be applied to deterministic matching as more data years are added.

The final status of our project objectives is summarized by the following:

1. We completed de-duplicating inpatients discharged during 2006-08 using deterministic methods and identified the most effective combination of data elements from a number of alternatives.
2. We examined potential deterministic matches between the most recent stay for each patient and the deaths occurring in 2006-08 and identified a large number of plausible matches.
3. We successfully developed rigorous probabilistic linkage models matching hospital patients to deaths and identifying high probability matches.
4. We analyzed the agreements and discrepancies between the results from each methodology. Generally the agreements were strong and validated our processes. We did draw lessons from the probabilistic results that will strengthen our deterministic processes going forward.