MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

Adding Value to Clinical Data By Linkage to a Public Death Registry

Robert D. Pates, Kenneth W. Scully, Jonathan S. Einbinder, Richard L. Merkel, George J. Stukenborg, Thomas A. Spraggins, Calvin Reynolds[a], Ronald Hyman[a], and Bruce P. Dembling

Department of Health Evaluation Sciences, University of Virginia Medical School, Charlottesville, Virginia, USA
[a] Virginia Center for Health Statistics, Virginia Department of Health, Richmond, Virginia, USA

Abstract

We describe the methodology and impact of merging detailed statewide mortality data into the master patient index tables of the clinical data repository (CDR) of the University of Virginia Health System (UVAHS). We employ three broadly inclusive linkage passes (designed to result in large numbers of false positives) to match the patients in the CDR to those in the statewide files using the following criteria: a) Social Security Number; b) Patient Last Name and Birth Date; c) Patient Last Name and Patient First Name. The results from these initial matches are refined by calculation and assignment of a total score comprised of partial scores depending on the quality of matching between the various identifiers. In order to validate our scoring algorithm, we used those patients known to have died at UVAHS over the eight year period as an internal control. We conclude that we are able to update our CDR with 97% of the deaths from the state source using this scheme. We illustrate the potential of the resulting system to assist caregivers in identification of at-risk patient groups by description of those patients in the CDR who were found to have committed suicide. We suggest that our approach represents an efficient and inexpensive way to enrich hospital data with important outcomes information.

Keywords:

Hospital Information Systems; Medical Record Linkage; Death Certificates

Introduction

The ability of researchers and clinicians to observe groups of patients over longer periods of time is recognized as important for the provision of quality health care [1]. During the past five years, we at the University of Virginia Health System (UVAHS) have developed an information system that brings together patient data from multiple sources called the Clinical Data Repository (CDR). To achieve this, we have merged available sources of automated hospital legacy data using Sybase relational database management software [2].

For many clinical studies, patient mortality is a crucial outcome to monitor, but hospital databases such as the CDR only contain in-hospital deaths. Previous research suggests that accurate linkage of hospital data to vital statistics files using both unique and partial patient identifiers is both feasible and advantageous [3,4]. As stated by Rosenberg: "No other health data source exists that is as universal in coverage, as standardized, uniform and timely as mortality data from the vital statistics system" [5]. Thus, we proposed to link the CDR with mortality information from the statewide death registry at the Virginia Center for Health Statistics (VCHS), thereby enabling the users of the CDR (researchers and clinicians) to perform a wide variety of longitudinal clinical studies that would not otherwise be feasible.

Since the VCHS data are subject to certain confidentiality restrictions, we were unable to obtain a direct copy of the statewide death files. However, we were authorized to request VCHS staff to perform a series of broadly inclusive queries of these data in an attempt to identify all CDR patients. We established a transferable operational protocol to detect the deaths of our patients recorded by the VCHS, and then we incorporated date, cause of death and other VCHS death data elements into our database. Finally, we evaluated the utility of the enhanced CDR by an assessment of those patients found to have committed suicide between 1992 and 1999.

Materials and Methods

Hardware and Software

The CDR is housed on a Dell PowerEdge 1300 (dual 400MHz Intel processor, 512MB RAM) running Linux operating system and Sybase (11.9.1) relational database management system. The machine is equipped with a Dell Powervault 201S RAID disk array system (capacity 236GB). Initial linkage steps were performed on an IBM 3090 mainframe computer using SAS (SAS Institute, Cary,

NC) software version 6.12. All subsequent data manipulations were performed on a Dell GX300 Desktop computer (600 MHz) using SAS version 8.1 for Windows 2000.

Data Linkage Process

Initial Match Process

Owing to privacy restrictions, we do not have direct access to the death files, so the following procedure was developed to retrieve broadly inclusive fractions of mortality data from VCHS, which were then refined.

The master patient index from the CDR was extracted from Sybase into SAS using SAS/ACCESS software routines. There were a total of 509,434 patients who had received care at the UVAHS since CDR records began (1/1/92). Of this total 439,006 (86.2%) were associated with a value for Social Security Number (SSN), and 509,162 (99.9%) had values for first name, last name, and birth date (272 patients were only identified by Medical Record Number only, and could not be included in the link process). Three files in SAS transport format were prepared from these data: a) CDR patient identifier (a unique integer internal to and assigned by the CDR system) and SSN (439,006 observations); b) CDR patient identifier, patient last name and birth date (509,162 patients); and c) CDR patient identifier, patient first name and last name (509,162 patients).

These 3 files were sent by File Transfer Protocol (FTP) to the VCHS mainframe. Three data runs were then made to link each of these files with VCHS mortality files recording deaths in Virginia from January 1st, 1992 to December 31st, 1999. The result of this process was 3 files, one from each of a), b), and c) above, containing 25,615, 30,512, and 171,022 rows respectively. These returned data were concatenated into a single file and linked back to the CDR identifying information by the CDR patient identifier. Each row was then assigned a weighted score according to a methodology described below. The file was then sorted by CDR patient identifier and weighted score, and the highest scoring row for each patient was selected. The result was a file containing 132,438 patients.

Refinement of Initial Match Fractions

A weighted scoring methodology described in [6] and [7] was employed. Briefly, each of seven identifiers was associated with probability m (the probability that the identifier agrees given that the pair is a match), and probability u (the probability that the identifier agrees given that the pair is not a match). Since this method is a reflection of the error rates in each of the identifiers, the first run was performed using reasonable estimates and subsequent runs were made with values calculated from the data. The seven identifiers (SSN, last name, middle initial, first name, birth date (birth year, month, and day), sex, and zip code) were each assigned a weighted score for each observation (according to $\log_2(m/u)$ in the case of agreement between the value from the CDR and that of VCHS, and $\log_2((1-m)/(1-u))$ for disagreement). Missing data contributed 0. The sum of the partial contributions (plus scaling factors, see below) was computed for a total weighted score. For example, in the case of last name, repeated runs indicated that agreement was 98.4% within matching pairs. Chance agreement was estimated at 1 in 900. Thus, this variable contributed a partial score of +9.8 ($\log_2(0.984/0.0011)$) for agreement while disagreement scored –6.1 ($\log_2(1-0.984)/(1-0.0011)$).

For those identifiers that were unevenly distributed (last name, first name, middle initial, birth year, and zip code), scaling factors were calculated [7]. The scaling factors adjusted the weighted scores to account for the relative frequency of particular values of the identifier. The scaling factor for a particular value of last name was calculated as $\log_2(\text{sqrt}(N/(Q*F)))$ where N is the total number of individual patients in the dataset, Q is the number of unique values of last name, and F is the frequency of the particular value of last name. In the case of last name the scaling factor ranged from +1.9 (for a name occurring once) to –4.1 (for the most frequently occurring name in the database).

Figure 1 displays the distribution of total weighted scores for all those pairs scoring 10 and above (n=38,138). Once scored, the rows were divided into one of three groups – "matches", "non-matches", and "uncertain" (see Table 1). The upper and lower thresholds of the "uncertain" matches were set at 2 standard deviations from the means of reasonable approximations of the "matches" (weighted score >= 24) and "non-matches" (weighted score < 24) populations. In this way, the "matches" and "non-matches" populations were defined by mean values of 54.3 (standard deviation = 8.7) and 4.8 (standard deviation = 3.9) weighted points, respectively.

Accordingly, the thresholds of the "uncertain" matches were set formally at >= 14.0 and <= 39.9. This range (which included 5,056 pairs) provided a useful starting point for identification and examination of "uncertain" matches. Following extensive manual review, the "uncertain" range was narrowed considerably (see Table 1). The final data fraction selected for inclusion in the CDR consisted of all pairs scoring 24 and above, with the following exceptions. In the range 24 – 40 inclusive we were careful to exclude apparent matches between family members – particularly in those cases where family members shared a single SSN. This was achieved by paying particular attention to the contributions to the total weighted score from sex and birth date variables.

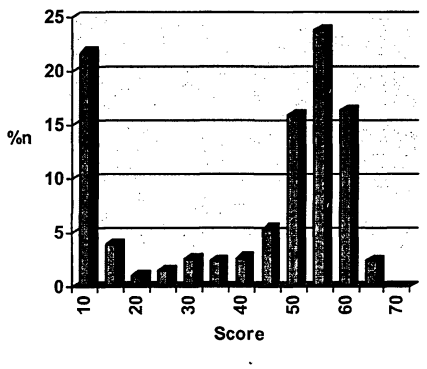

Figure 1- Distribution of weighted scores for those matches whose scores were 10 and above (n=38,138)

Results

To validate the scoring process, we examined those scores obtained by patients reported by UVAHS to have died over the eight year test period. Table 1 shows scored and grouped data for both the total sample (n=132,438), and for UVAHS-reported deaths (n=5,220).

Table 1 – Results of grouping the data (date range 1/1/1992-12/31/1999) by weighted score

Weighted Score	Total N (%)	UVAHS N (%)
>24 ("matches")	28,008 (21.2)	5,072 (97.2)
20 – 24 ("uncertain")	336 (0.2)	31 (0.6)
<20 ("non-matches")	104,094 (78.6)	51 (0.9)
[Not returned in initial match]	N/A	66 (1.3)
Total	132,438 (100.0)	5,220 (100.0)

For the total sample, 28,008 (21.2%) patients achieved a score of 24 or above, and were deemed to be "matches". Of these, 24,602 (87.8%) were derived from patients originally linked by SSN, while 3,172 (11.3%) were derived from the initial last name / birth date match, with 234 (0.8%) being derived from last name and first name match. Thus, while SSN proved to be by far the most powerful identifier, the overall result is enhanced significantly by the additional inclusive initial match steps. This is particularly important in our case since for many patients we cannot rely on SSN for a match. Only 86% of the rows in our patient index bear a valid SSN (and approximately 10% of these may be erroneous).

Further reference to Table 1 enables assessment of the overall quality of the matching and scoring scheme employed here. For example, of those patients known to have died in the hospital (n=5,220), 5,072 (97.2%) scored a total of 24 points or more. This figure may be a conservative estimate, since of those that were missed by the scoring system (n = 31 + 51 = 82), 54 (65.9%) were infants born after 1990 and consequently poorly identified in our database. Accordingly, of the 66 not returned by the initial match process, 35 (53.0%) were infants.

Richness of Mortality Data in the CDR

The VCHS data included in the CDR upon linkage comprise some 51 data elements in addition to date of death, including ICD codes for underlying cause of death, plus up to 20 contributing causes of death (average approximately 3), and 3 accident codes. Additional clinical data elements include flags indicating whether the deceased was pregnant at time of death, and whether an autopsy was performed. The remaining data elements are geographic (place of death, hospital of death, place of birth, place of residence at time of death), administrative (date death filed), and demographic (father's, mother's names) in nature.

Discussion

Quality of the Method

We demonstrate a rather effective and inexpensive method for linkage of a hospital database to public death files. For example, our internal controls indicate that we may be able to detect over 97% of the total in-state deaths. This figure is very similar to that reported by Newman and Brown in 1997 who were able to link 96.5% of their in-hospital deaths using commercially-obtained software ("Automatch"). Among the initial SSN match (n = 25,280) -- often the sole linkage employed for many purposes -- we detected and eliminated 678 (2.7%) false positive matches. In addition, we detected and included 3,406 (12.2%) true positive matches that would have been missed using SSN matches alone.

Since approximately 95% of Virginia residents actually die in-state [8], we further estimate that for the time period under consideration (i.e. from 1992 until the conclusion of the previous calendar year), that we are able to include 92% (i.e. 95% of 97%) of the total mortality among those patients in our hospital database. With much extra effort, the number of false negatives could be reduced further by linkage of death records from surrounding states (WV, TN, KY, NC, MD, and Washington DC). An alternative solution would be to link to the National Death Index (NDI) of the National Center for Health Statistics. However, not only is this an expensive service ($350 plus $0.30 per person per year searched), but also the data are less timely than those provided directly by the state.

Clinical Advantages for the CDR

Mortality is a very important outcome that ideally must be measured accurately to assess quality of care provided. For example, for clinical practices in which patients are at immediate risk of death, the mortality outcomes of treatment can perhaps be monitored adequately in datasets that only include in-hospital deaths. However, the risk of death in the period immediately following discharge is substantial for many clinical practices.

Slightly fewer than half of all deaths occur in hospitals [8]. Consequently, the combined clinical and cause of death data for this population would provide a substantial increase in epidemiological knowledge over that contained in either record alone. Those who do not die in hospitals frequently have hospital treatment histories related to their causes of death. While hospitals have detailed knowledge of the patients who die in their facilities, they may have little or no knowledge of the health status of patients after they leave the facility.

Now that the CDR is updated with date of death as described above, the extent of such mortality may be monitored more accurately. In addition, clinical practices for treating patients with chronic conditions (cancer, kidney disease, heart disease) have the objective of extending life, and this outcome cannot be assessed without longitudinal information regarding patients with chronic disease diagnoses and dates of death. Thus, questions regarding patient longevity, life expectancy, geographic variations, and comorbid conditions contributing to death may now be addressed using the CDR. Conversely, for those patients who died at UVAHS, we would be able to verify the reliability and validity of data comprising the death records.

Table 2 shows the contribution of in-hospital deaths to the linked total by cause of death category. The clinical classifications in Table 2 are a result of cross-walking the ICD codes to the Clinical Classifications for Health Policy Research (CCHPR, maintained by the Agency for Healthcare Research and Quality (AHRQ)) [9].

The data indicate that in no cause of death category does in-hospital death account for any more than 40% of the total for the category. So, despite a marked variation across categories, the inclusion of mortality data from the state source is clearly advantageous for the study of all causes of death. For example, while suicides in the US account for less than 2% of all US deaths each year, they disproportionately affect young age groups. Consequently, suicide has become a major public health priority for the Surgeon General [10]. The updated CDR provided us the opportunity to assemble and characterize those patients who between 1992 and 1999 were shown to have committed suicide.

A Closer Look at Suicides Among UVAHS Patients

We selected all those patients whose underlying cause of death ICD-9 code fell between "E950" ("Suicide-Analgesics") and "E959" ("Late effects of self-injury")

Table 2 – Contribution of in-hospital deaths to the total deaths by category in the CDR following linkage to the VCHS death registry.

Category of cause of death	In-Hospital deaths in CDR as a percentage of total for category
Infectious Disease	40.0
Neoplasms	12.7
Endocrine	17.5
Circulatory	17.5
Respiratory	20.0
Digestive	35.7
Accident	33.3
Suicide	15.9
Homicide	18.9
Other	20.8
Total (5,220/28,008)	18.6

inclusive. For the eight year period, a total of 336 suicide cases retrieved from the CDR and were classified based on a review of their clinical histories. One hundred and twelve cases (33.3%) received minimal care at UVAHS, including 29 (8.6%) whose only contact was treatment for the suicide event. The remaining cases included 136 (40.0%) with lengthy medical histories but no psychiatric contacts, 36 (10.7%) with both medical and psychiatric care, and 52 (15.5%) with psychiatric care only. The population was found to be predominantly white, 77% male and of median age 45 years. On average, firearms were used in 63% of the suicides. Approximately 50% died within 6 months of last contact for care.

Health providers have no practical means to reduce risk for those patients with no diagnostic history. However, post mortem reviews of other clinical categories in view of the suicidal outcome could reveal signs of depression or other risk factors with potential preventive value. Thus, linkage with state death records enables accuracy in quality assessment and suicide detection and prevention efforts that would not be possible using hospital records alone.

Conclusion

The method described here provides a database linked with mortality data that are acceptable for research applications at UVAHS. Since all states produce death records according to standards dictated by the WHO and the National Center for Health Statistics, this approach may be applicable in other US states, in addition to Virginia.

Acknowledgements

The authors would like to thank Jane Schubart of the UVA Dept. Health Evaluation Sciences and Susan K.R. Heil and her colleagues at the Oklahoma Department of Mental Health and Substance Abuse for helpful discussions.

References

[1] Schubart J. Improving the quality of medical care: A case for the hospital-based clinical data repository. *HIMMS News,* May 1998.

[2] Scully KW, Pates RD, Desper GS, Connors AF, Harrell FH, Pieper KS, Hannan RL and Reynolds RE. Development of an enterprise-wide clinical data repository: merging multiple legacy databases. *Proc AMIA Fall Symposium*, 1997: pp. 32-36.

[3] Herrchen B, Gould JB and Nesbitt TS. Vital statistics linked birth/infant dealth and hospital discharge record linkage for epidemiological studies. *Comput Biomed Res.* 1997: 4 pp. 290-305.

[4] Newman TB, and Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. *Journal of the American Medical Informatics Association* 1997: 4 pp. 233-237.

[5] Rosenberg HM. Cause of death as a contemporary problem. *Journal of the history of Medicine* 1999: 54 pp. 133-153.

[6] Jaro MA. Probabilistic linkage of large public health data files. *Statistics in Medicine* 1995: 14 pp. 491-498.

[7] US Department of Health and Human Services, Center for Substance Abuse Treatment. Integrated database project. Linking methods. 2000: URL: http://sg.gov/library/calltoaction/default.htm.

[8] National Center for Health Statistics. Multiple cause of death computer file and documentation for 1993. ICSPR 6546, 1995.

[9] Clinical Classifications for Health Policy Research, Version 2: Hospital Inpatient Statistics. AHCPR publication no. 96-0017. Rockville, MD: US Department of Health and Human Services, Agency for Health Care Policy and Research; February 1996.

[10] Satcher D. The Surgeon General's call to action to prevent suicide 1999. URL: http://sg.gov/library/calltoaction/default.htm.

Address for Correspondence

Robert D. Pates
Department of Health Evaluation Sciences
University of Virginia
Charlottesville
Virginia 22908
USA
Rpates@virginia.edu