Linkage of the Victorian Admitted Episodes Dataset

Vijaya Sundararajan, MD, MPH, FACP, Toni M. Henderson, Michael Ackland, MBBS, FAPHM, Ric Marshall, PhD

Victorian Department of Human Services

Abstract

Background

The Victorian Admitted Episodes Dataset (VAED), the state's hospital morbidity dataset, is an episode-of-care level dataset. Turning the VAED into a case-level dataset has potential benefits in epidemiologic, health services research and quality of care research. However, at this time, there is no unique variable which can be used to separate the dataset into cases.

Methods

Initially, for the fiscal years 1994–2000, we evaluated the quality of data by comparing the agreement of identifiers using well-matched pairs of observations. Next, four linkage variables were created for use: 1) year/day/month of birth/postal code/continent of birth/gender (link1); 2) hospital code/hospital record number (link2); 3) 3-digit medicare suffix/8-digit medicare number (link3); 4) year/day/month of birth/postal code/gender (link4). Link1 was the variable used for the first pass at linkage and was the basis for the newly created variable "caseno". After this first pass, two derivative sets were created: one with observations which did not group with any others based on agreement of their link1 (the "orphans") and one with observations which did group (the "many"). For the second pass, link2 was compared between observations in the orphan set to those in the many set – if there was agreement on link2, the orphan observation was given the "caseno" from the many dataset's observation whose link2 agreed with it. Passes 3 and 4 were conducted in the same fashion after creating the second and third orphan and many sets. The pilot project was conducted using SAS 8.0.

Results

The public VAED had Medicare numbers for 86.6 to 100% of its observations; the private VAED was more limited with only 30–34% of its observations with a Medicare number until fiscal year 2000 when the rate was 100%. The coding error rate was low: most of the potential linkage variables had 98–99.9% agreement in pairs of well-matched observations. These data have been used in a number of studies which will be briefly described.

Introduction

The linkage of administrative data offers great scope for epidemiological and health services research. In Western Australia, such a health services research database has been completed and is proving valuable (Holman, Bass et al. 1999). In Victoria there is no unique identification number that allows the accurate separation of records into case-level data. A linkage process using combinations of several identifying variables must be used. We describe the steps we undertook in a project transforming the Victorian Admitted Episodes Dataset (VAED) from episodes of care level data into case-level data for the calander years 1995 to 2000. The VAED contains episodes of care level data (usually hospitalisations, but at times separate records for care delivered in different specialty areas) from all separations in Victoria, both public and private (Division 2000). Importantly, the VAED contains data on diagnostic related groups (DRGs), ICD–9 and 10 diagnostic and procedure codes associated with each separation.

The steps we undertook to complete the linkage were:

1. Assessment of the quality of the coding of the variables.
2. Development of the linkage algorithm.
3. Assessment of the quality of the data linkage.

Methods

Data source

The VAED is a minimum dataset of acute hospital separations from throughout Victoria (Division 2000). It contains demographic information (date of birth, gender, country of birth, postal code, marital status), Medicare number and suffix, hospital code and hospital record number, 25 diagnostic and 25 procedure codes (ICD–9 until 1998, ICD–10 thereafter), transfer information and length of stay.

Coding quality of the identification variables in the VAED

To assess the quality of coding we assessed the "frequency of agreement in links". This frequency quantifies the level of agreement of a variable in a pair of observations that is otherwise perfectly matched. For instance, for year of birth, we might evaluate pairs of observations that agree on month/day of birth, postal code, country of birth, gender, Medicare number and hospital record number (Table1). The frequency of 'agreement in links' for year of birth in this set of 4 pairs would be 75%.

The frequency was evaluated for the following identification variables:

- 4-digit year of birth
- Month of birth
- Day of birth
- 4-digit postal code
- Country of birth
- Gender
- 8-digit Medicare number (The error rate in the last two digits was high.)
- Hospital record number

Rationale of linkage

The linkage algorithm used multiple cycles to bring together hospitalisations from the same case (Newcombe, Kennedy et al. 1959; Newcombe 1988). The rationale of the linkage process was more deterministic rather than truly probabilistic. Given the large number of observations we were dealing with, greater than one million per calendar year, we decided that this was necessary to minimise the false positive match rate. Initially we completed the algorithm year-by-year and then determined the false positive and false negative rates. Within each year, 3–4 passes were undertaken to bring together observations which belonged together into case-groups. All linkage was conducted on SAS version 8.2 (SAS Institute 1999).

First pass: The initial pass was based on grouping the observations using a combination of date of birth, gender, postal code and country of birth (Link1). The observations were then separated into two datasets, one called "MANY" in which there were at least 2 or more observations with the same Link1, the other called "ORPHAN" which had all the observations which did not group to any other observations based on this Link1. A new variable was created in MANY, called "Newid" which was equal to Link1. Newid was also created in ORPHAN, but was set to missing at the end of the first pass.

Second pass: The next pass used hospital code and hospital record number as its linkage variable. In essence, all of the observations in ORPHAN which had the same hospital code and record number as those in MANY had their missing Newid replaced by the Newid from the MANY dataset. In this way, the ORPHAN observation went to the case-group from MANY with which it agreed on hospital code and record number. The observations from ORPHAN which had a Newid after the second pass were added to MANY. ORPHAN2 was then created from the remaining unmatched observations for the next pass.

Third pass: For the third pass, the link was based on the combination of 8-digit Medicare number and 3-digit Medicare suffix.

Assessment of the quality of the data linkage

Initially in developing the linkage algorithm, the quality of the linkage was assessed in two ways. For the false positive rate of the linkage (1-specificity), the frequency with which observations were put into groups to which they did not belong, we randomly selected 200–300 case groups (approx 1300 observations) and manually evaluated whether all of the observations within case groups belonged together. Two formal rules were used to define a false positive observation: 1) an observation had a Medicare number and suffix which disagreed with the others; and; 2) an observation had the same hospital code as others in the case group but differed in the hospital record number. The false positive rate was then the number of observations grouped incorrectly in ratio to the total number of observations evaluated.

The false negative rate was more challenging to estimate. Our goal was to use this in the process of improving our linkage algorithm. We surmised that a reasonable way to assess whether observations were missing from case-groups was to look at the data on transfers. In essence, if a case were transferred from one hospital to another, both observations should be within the same case-group. We initially took a random sample of transfers, pulled in all of the data on these cases based on the Newid, and then manually assessed how frequently the second hospitalisation was present in the case-groups.

Results

Coding quality of the identification variables in the VAED

There was a change in the proportion of observations with a Medicare number between the years 1995 to 2000 (Table 2). For the public hospital data, Medicare number was present in the majority of observations, and, by 2000, it was present in all records. For the private hospital data, Medicare number was present in approximately one-third of all observations until 1999; here too, by 2000, all observations had a Medicare number. The increase in this proportion is reflective of changing reporting requirements and a general improvement in data quality.

The frequencies of agreement in perfect matches based on hospital code and record number for important linkage variables were generally good. Individually, year of birth, day of birth, month of birth and gender agreed greater than 98% of the time, whereas postal code and country of birth agreed more than 92% of the time. Medicare number and suffix agreed more than 90% of the time. The composite variable made up of date of birth, gender, postal code and country of birth agreed at least 87% of the time. Notably, in pairs of observations matched on Medicare number and suffix, hospital code and record number agreed only 45–50% of the time.

Sensitivity and specificity of linkage

The false positive error rate, that is, the proportion of observations incorrectly classified into case-groups was between 1–2%, remaining stable throughout the years 1995–2000. In comparison, the false negative error rate, based on whether both records from hospital-to-hospital transfers were in the same case group, was 15%.

Discussion

We have presented our initial efforts at linking the Victorian Admitted Episodes Dataset. We focused our efforts at developing a linkage method that would minimise false positive and false negative error rates while still being automated and requiring little manual review.

The three sets of linkage variables available each had their own strengths and weaknesses. The composite variable made of date of birth, gender, postal code and country of birth was present on every observation, but had a high rate of coding errors or, in the case of postal code, had differences reflecting an actual change in postal code for a case. The Medicare number and suffix had very good coding quality when it was present, but was not available on every record. The hospital record number in combination with the hospital code was present on every record, but, reflective of the fact that a case rarely went to only one hospital, had a low frequency of agreement among perfectly matched pairs of observations.

The three passes in our linkage algorithm attempted to maximise the usefulness of each set of linkage variables. Only three passes were used because, after the third pass, further incorporation of observations into case-groups based on less discriminating linkage variables would have increased the false-positive error rate significantly.

Our false positive error rate, the rate at which a record is included into a case-group to which it does not belong, is 1–2% per year. We have yet to combine the years, but do not anticipate that this rate will change. In contrast, our false negative, the rate of missing an observation from the case-group to which it belongs, is 15%. This false-negative rate may be higher than it actually is because of how it was estimated. Because of the fact that there is no reference or "gold" standard to which we can compare the accuracy of our linkage, we have had to develop other ways to estimate our error rates. For the false negative error rate, we have used the fact that if a case is transferred from hospital-to-hospital, both of these records should be within the same case group. Transfer records may not be representative of all records in their false negative rates. By their nature they require identification variables from two different hospitals to match, which may be subject to more error in coding. We use the transfer data because it appears to be the least biased way of estimating the error rate. Using records of cases requiring chemotherapy or hemodialysis may be helpful to refine this error rate, but we believe that rates based on these two diagnoses would tend to falsely lower the false negative error estimate because the coding practices of chemotherapy or dialysis units may be much better than hospitals in general.

How high of a false positive/negative rate is acceptable? No strict thresholds exist for this. At this point, we believe our false negative rate is too high, even assuming that our estimate is higher than the rate is in reality. Our efforts are currently aimed at incorporating date of birth, gender and a diagnostic code (using one of the first three codes listed per record) because most records from case-groups appear to have an overlap of initial diagnostic codes.

Month of birth	Day Of Birth	Postal Code	Country Of birth	Gender	Medicare number	Hospital Record Number	**Year of birth**
12	24	3055	23	1	56386970	94687256	**1923**
12	24	3055	23	1	56386970	94687256	**1823**
8	16	3698	56	2	98867245	35678492	**1945**
8	16	3698	56	2	98867245	35678592	**1945**
7	12	3975	22	1	64582369	75182964	**1926**
7	12	3975	22	1	64582369	75182964	**1926**
6	4	3498	19	1	79685924	65643879	**1948**
6	4	3498	19	1	79685924	65643879	**1948**

Table 1 Assessment of the frequency of agreement in links*

*All data shown are fictitious and do not represent records from the VAED.

Table 2 Proportion of observations with a Medicare number, 1995 to 2000

	Public Hospital VAED		Private Hospital VAED	
Calendar Year	Total number of observations	Observations with a Medicare number, %	Total number of observations	Observations with a Medicare number, %
1995	905621	88	423422	33
1996	936500	88	458635	31
1997	964697	88	484236	31
1998	1007577	89	495700	33
1999	1041835	99	519838	47
2000	1065572	100	580420	100

	Private	Public
Matched on hospital code and record number		
Sex	99-100	99-100
Day of birth	98-100	99-100
Month of birth	99-100	99-100
Year of birth	98-100	99-100
Country of birth	92-97	94-96
Continent of birth	96-98	98-99
Postal code	95-96	92-95
Medicare suffix and 8-digit Medicare number	90-93	92-95
Date of birth||postal code||gender||country of birth	87-93	97-90

Table 3 Frequencies of agreement in perfect matches, 1995–2000

Year	False positive error rate, %
1995	1
1996	1
1997	2
1998	1
1999	2
2000	1

Table 4 False positive error rate (1-Specificity), 1995–2000

References

Victorian Department of Human Services. (2000). The Victorian Admitted Episodes Dataset: An Overview April 2000. Melbourne, Acute Health Division; Victorian Government Department of Human Services: 61.

Holman, C.D., A.J. Bass, et al. (1999). "Population-based linkage of health records in Western Australia: development of a health services research linked database". Aust N Z J Public Health 23(5): 453–9.

Newcombe, H.B. (1988). Handbook of record linkage: methods for health and statistical studies, administration, and business. Oxford, Oxford University Press.

Newcombe, H.B., J.M. Kennedy, et al. (1959). "Automatic linkage of vital records". Science 130(3381): 954–959.

SAS Institute. (1999). SAS 8.2. Cary, North Carolina.