HHS Public Access

Author manuscript

J Public Health Dent. Author manuscript; available in PMC 2015 November 02.

Published in final edited form as:

J Public Health Dent. 2013 ; 73(2): 89–93. doi:10.1111/j.1752-7325.2012.00343.x.

ACCURACY OF RECORD LINKAGE SOFTWARE IN MERGING DENTAL ADMINISTRATIVE DATASETS

Heather Beil, PhD, MPH[1], **John S Preisser, PhD**[2], and **R. Gary Rozier, DDS, MPH**[3]

[1]School of Nursing, University of North Carolina, Chapel Hill

[2]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill

[3]Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina, Chapel Hill

Abstract

Objective—To determine the accuracy of record matching using “Link King” software that uses an ordinal score for the certainty that linked records are valid matches.

Methods—We linked records in North Carolina Medicaid files to public health surveillance files using Link King matching software. We selected a stratified random sample of 230 of 45,295 linked records and 50 of 35,119 non-linked surveillance records, then manually reviewed the records. Sensitivity (Sn) and Specificity (Sp) were calculated based on each cutpoint of the Link King score, using manual review as the gold standard.

Results—The Sn increased from 0.837 (95% CI: 0.785, 0.892) to 0.935 (0.879, 0.994) and Sp decreased from 0.893 (0.816, 0.976) to 0.865 (0.790, 0.947) as cutpoints were varied to widen the scope of declared matches. With a goal of both Sn and Sp being large, accuracy was best when linked record pairs from the top three levels of certainty were included in the match.

Conclusions—This study found that a publicly available software program accurately merged Medicaid and surveillance data. The Link King could be useful to researchers in merging administrative databases.

Keywords

Data Linkage; Medicaid claims; dental care

Introduction

Administrative databases have been used in dentistry to study provider practices, the outcomes of dental interventions, variations in care, and to help establish standards of care (1). Databases created from dental insurance claims are among the most common types of administrative files used in dental research, but unlike medical claims, they do not contain diagnostic codes (2). A few researchers have been able to overcome this problem by linking records from multiple existing datasets, one of which has oral health status information, to create a new merged dataset (3–6). Records are linked using variables common to both datasets when there is not a unique identifier. Linking records without a unique identifier

can be cumbersome and time consuming, however, which can inhibit investigators from engaging in research using merged administrative databases.

Two record linking strategies are generally used to match records with no unique identifier, the deterministic method and the probabilistic method (7–9). Both methods match records based on a group of variables such as name, date of birth, social security number (SSN), and zip code. Simple deterministic algorithms require records to match exactly, while more complex algorithms allow for slight differences by using iterative "fuzzy" matching (9). The probabilistic method relies on statistical analyses to calculate how probable it is that two record pairs are the same person. This method derives a score for each linked record pair that is compared to cut points to indicate if the record pair is a definite, probable or non-match (7), with possibly more refined classifications for 'probable'. The deterministic method typically declares fewer linked records to be matches due to a higher specificity, but with a higher proportion of the declared matches being true matches (10). The probabilistic method, on the other hand, usually has a higher sensitivity meaning that a greater number of true matches are found (10), at the expense of a higher false positive rate, that is, a higher proportion of the declared matches not being true matches.

Several software programs are available that link and eliminate duplicate records in administrative datasets, but little is known about their accuracy (10, 11). To our knowledge, none of these programs have been used in dental research. This study aimed to test the accuracy of "Link King", one such program available in the public domain (12). The Link King uses both the deterministic and probabilistic methods to match and eliminate duplicate records. It was developed at Washington State's Division of Alcohol and Substance Abuse using a probabilistic record linkage protocol that was adapted from the algorithm developed by MEDSTAT for the Substance Abuse and Mental Health Services Administration's (SAMHSA) Integrated Database Project (10,12). If the Link King software matches records accurately, it could facilitate dental public health research that otherwise might not be possible and enhance its efficiency. Researchers have found that Link King accurately eliminated duplicates in a large dataset, but note that the program may not work as accurately if the user chooses the wrong cutpoint for the degree of certainty of whether a link is a match (10).

This study had two main objectives: 1) to determine the accuracy in terms of sensitivity and specificity of the matching program "Link King"; and 2) to identify the cutpoint for the software's user defined "linkage certainty level" that best demarcates matches from non-matches giving balanced consideration for minimizing false positive and false negative declared matches.

Methods

Data sources

This study merged North Carolina Medicaid enrollment and oral health status files derived from the North Carolina Surveillance of Dental Caries (NCSoDC) (13) with the purpose of conducting analyses on the effects of dental utilization on oral health. We evaluated the accuracy of the resulting merge by calculating the sensitivity and specificity of decision

rules using manual review of records as the gold standard. From these results we determined the cut point with the best accuracy. The Medicaid files contained information for children birth to 6 years of age enrolled from October 1999 to June 2006 for all children born on or after January 1, 1998. The Medicaid files were originally obtained from the NC Division of Medical Assistance to evaluate the effectiveness of the "Into the Mouths of Babes" program in which primary care physicians provide oral health services (14, 15).

The NCSoDC provides basic demographic information on each individual child in kindergarten including name, date of birth, sex, race, school name, classroom identification number within school and county of residence. Additionally, the NCSoDC provides a count of decayed, filled, and missing (molars only) primary teeth for each child. Information from the 2005–06 school year surveillance was used for this study because it was the only year of available data that overlapped with our sample of Medicaid children. The NCSoDC used for this study contained 80,414 kindergarten children from 98 of the state's 100 counties, or 70% of the state's public school enrollment for this grade.

Link King program description

The Link King version 6.4.9 allows the user to input one or two datasets to eliminate duplicate records and match records. The software includes five default ordinal "linkage certainty levels" for linked records and the remaining records are considered to be 'non-matches', which we refer to as Level 6. Each successive linkage certainty level has a decreasing likelihood that linked pairs are true matches, therefore Level 1 has the highest possible probability of being a true match and Level 5 has the lowest probability. The user decides how many levels to include in the merge to create the final dataset, however, the user cannot determine how many matches will be in each level of certainty. The user must also input matching variables from a choice of name, social security number, date of birth, gender, race and a 'flexible' variable that can be a geographic variable that the two datasets have in common such as zip code or city name. Link King requires that the two datasets be matched on at least the first and last name and date of birth or social security number. We matched on all of the data elements that were common to both datasets: first name, last name, middle name, date of birth, race, gender and county of residence.

The first objective of the study is to calculate the accuracy of the software-generated matches when different linkage certainty levels are included in the data merge. The second objective of this study is to determine what linkage certainty level would result in the best accuracy of our data merge, giving balanced consideration to sensitivity and specificity.

Sample Size Allocation

Our goal was to have an adequate sample size and allocation of sample size across certainty levels in order to discriminate between the accuracy of the adjacent five cut point-based decision rules. The sampling scheme proportionally over-sampled records from certainty levels corresponding to the greatest potential for uncertainty. This led us to oversample match certainty levels 4 and 5. Link King generated 23 linked pairs at Level 3, so we included almost all of them (20) in the final sample. Additionally, we sampled 50 record

pairs from Levels 1, 2, and 4 and non-linked records, and 60 record pairs from Level 5 for a total sample size of 230 linked pairs and 50 non-linked records (Figure 1).

Matching review process

The accuracy of the match was determined comparing the results to a "gold standard", which was manual review. Using the method used in a previous study on the accuracy of record linking, two reviewers, including one author and another trained reviewer, examined each linked pair of records and classified it into one of five categories: a) definitely not the same; b) probably not the same; c) not enough information to determine whether or not they are the same; d) probably the same; and e) definitely the same person (10). The final decision considered the linked pair as a valid match if both reviewers classified the record pair as "probably" or "definitely" the same. Remaining record pairs were considered invalid links. Similarly, the reviewers sampled 50 non-linked records from the NCSoDC surveillance database and searched the NC Medicaid enrollment data for possible matches using a computer-assisted 'blocking' strategy (9) to make the search feasible. In particular, subsets of the NC Medicaid enrollment data were created based on possible misspellings of first names and last names separately and birthdate data to determine whether the sampled NCSoDC non-linked records were true negatives.

Data Analysis

As in a previous study, observations (selected linked and non-linked surveillance records) were weighted by the inverse of their probability of selection into the sample to produce population estimates of Sensitivity (Sn), and Specificity (Sp) (10), where 'population' refers to the complete set of N=80,414 NCSoDC records. Sn and Sp are defined in greater detail in the appendix.

This study was approved by the Institutional Review Board of the University of North Carolina at Chapel Hill and the North Carolina Division of Medical Assistance.

Results

As expected, the sensitivity of decision rules for declaring matches increased, while the specificity decreased, with the inclusion of linked records of increasing levels of uncertainty. The sensitivity values ranged from 83.7% to 93.5% and the specificity ranged from 86.5% to 89.3% (Table 1). Accuracy was judged to be best when linked record pairs from the top three levels of certainty were included in the match based on the following considerations. First, sensitivity was relatively low (83.7%) if only certainty level 1 was included, whereas inclusion of linked pairs with certainty levels 2 and 3 markedly increased sensitivity (89.2%) while maintaining reasonably high specificity (88.2%). Second, due to the small number (N_3) of linked records pairs with certainty level 3 resulting from the merge, the analysis weight for the corresponding linked pairs in the sample (n_3=20) was very small (value of 1.1; Figure 1) indicating negligible influence of level 3 linked pairs on accuracy. Finally, inclusion of certainty levels 4 or 5 in declared matches would have resulted in notable reductions in specificity.

Discussion

This study demonstrated that publicly available software, the Link King, displayed a high level of accuracy in merging administrative databases. These results indicate that the Link King could be beneficial to researchers in merging administrative databases efficiently and accurately, thus allowing for expansion of the scope and importance of research questions. The software could prove useful to researchers in applications such as eliminating duplicate records in administrative data and merging datasets such as insurance claims files, patient records and public health surveillance data.

Linking datasets such as insurance files and surveillance data allows dental public health researchers and practitioners to examine relationships between the use of dental and other health care and oral health status. Without linked files, it is not possible to determine oral health status from claims files alone, especially since dental claims do not have diagnostic codes. From a public health point of view, linked population databases would permit public health officials to determine use and oral health status for a large geographic area. For example, a previous study that linked Medicaid and surveillance data used the data to produce statewide reports on children's oral health status by public insurance enrollment status by county for the state health department (4,19).

We used the Link King to merge NC Medicaid data to an oral health surveillance database. Both datasets had a very large number of observations representing a census of a subset of an entire state population. Unlike experimental trials, linked datasets offer the advantage of including the population of interest (16). More than half of the observations in the surveillance database matched with records in the Medicaid dataset using a limited number of variables that each database has in common. In addition to matching on name, date of birth, race, gender and a variable of choice (county of residence), the Link King also has options to match on SSN. The Link King requires the data to contain at least a first name, last name and either date of birth or SSN to link or unduplicate data. Data linkage accuracy is dependent on the variables used and type of data that are available. Linkage of other administrative datasets may therefore have more or less accuracy with using the Link King then we found in this study.

Our results indicate that the software matched records at high levels of accuracy for each level of certainty primarily because sensitivity improved significantly while there was little loss in specificity at higher levels. This result is consistent with prior observation that when the number of records truly unmatched is large, specificity will be uninformative (7). When choosing the level of certainty for the cutpoint, however, researchers may want to put differential weight on sensitivity or specificity depending on the study's purpose (17). We used matches from levels 1, 2 and 3 in the study of dental use that resulted from the data merge we performed with the Link King because we judged that it was more important for our research question that linked pairs had a high probability to be valid links rather than having all possible matches. The size of the database and the size of the population at each level should be considered when determining the appropriate number to sample to evaluate the accuracy of matching software. In future studies, determination of sample size and its allocation across certainty levels could be based on considerations of minimizing the

variance of an objective criterion such as the sensitivity or specificity at various cutpoints (18).

Limitations

This study relied on manual review as the gold standard to determine matches. Although that is a standard that has been used in other studies, it is subject to human error in making the choice whether a link is a true match or not (7). Moreover, it was not possible to manually review all records in the surveillance database when searching for false negatives, so records could have been missed, which would overestimate the specificity. We also had a small population size in the third level of certainty, which made it very difficult to distinguish the accuracy of a match rule that included the level relative to one that did not.

Conclusions

Linking administrative databases provides the opportunity for dental researchers to conduct studies that otherwise would not be possible. Linking data by expert review can be time consuming and prone to errors when there are no unique identifiers. In this study merging NC Medicaid data to an oral health surveillance dataset, the Link King proved to be an accurate, efficient way to link the datasets. Accuracy of merging data is contextual, however; it depends on the dataset and variables available. Further testing should be done to confirm the Link King's accuracy with other administrative datasets.

Acknowledgements

Funding for work presented in this paper was provided by Grant No. 1R36HS018076-01 from the Agency for Healthcare Research and Quality. Funding for the acquisition of the data used in this paper was provided by Grant No. R01 DE013949 and Grant No. R03 DE017350, both from the National Institute of Dental and Craniofacial Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDCR or the National Institutes of Health (NIH).

References

1. Leake JL, Werneck RI. The use of administrative databases to assess oral health care. J Public Health Dent. 2005; 65:23–35.
2. Miller CS. Where are the diagnostic codes in dentistry? Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2011 Feb; 111(2):131–132. [PubMed: 21093325]
3. Ismail AI, Brodeur J-M, Gagnon P, Payette M, Picard D, Hamalian T, Oliver M. Restorative treatments received by children covered by a universal, publicly financed, dental insurance plan. J Public Health Dent. 1997; 57:11–18. [PubMed: 9150059]
4. Brickhouse TH, Rozier RG, Slade GD. Effects of enrollment in Medicaid versus the State Children's Health Insurance Program on kindergarten children's untreated dental caries. Am J Public Health. 2008 May; 98(5):876–881. [PubMed: 18382008]
5. Pahel BT, Rozier RG, Stearns SC. Agreement between structured checklists and Medicaid claims for preventive dental visits in primary care medical offices. Health Informatics J. 2010 Jun; 16(2): 115–128. [PubMed: 20573644]
6. Robison VA, Rozier RG, Weintraub JA. A longitudinal study of school children's experience in the North Carolina Dental Medicaid Program, 1984 through 1992. Am J Public Health. 1998; 88(11): 1669–1673. [PubMed: 9807534]
7. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Inter J Epidemiol. 2002; 31(6):1246–1252.

8. Cotter JJ, Smith WR, Rossiter LF, Pugh CB, Bramble JD. Combining state administrative databases and provider records to assess the quality of care for children enrolled in Medicaid. Am J Med Qual. 1999; 14(2):98–104. [PubMed: 10446671]

9. Clark DE. Practical introduction to record linkage for injury research. Injury Prevention : Journal of the International Society for Child and Adolescent Injury Prevention. 2004; 10(3):186–191. [PubMed: 15178677]

10. Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. Health Informatics Journal. 2008; 14:5. [PubMed: 18258671]

11. Jones L, Sujansky W. Patient Data Matching Software: A Buyer's Guide for the Budget Conscious. California Health Care Foundation. 2004

12. The Link King. Record Linkage and Consolidation Software. Camelot Consulting. 2008. Available at: http://www.the-link-king.com/

13. Rozier RG, King RS. Defining the need for dental care in North Carolina: contributions of health surveillance of dental diseases and conditions. N C Med J. 2005 Nov-Dec;66(6):438–444. [PubMed: 16438100]

14. Pahel BT, Rozier RG, Stearns SC, Quiñonez RB. Effectiveness of preventive dental treatments by physicians for young Medicaid enrollees. Pediatrics. 2011 Mar; 127(3):e682–e689. Epub 2011 Feb 28. [PubMed: 21357343]

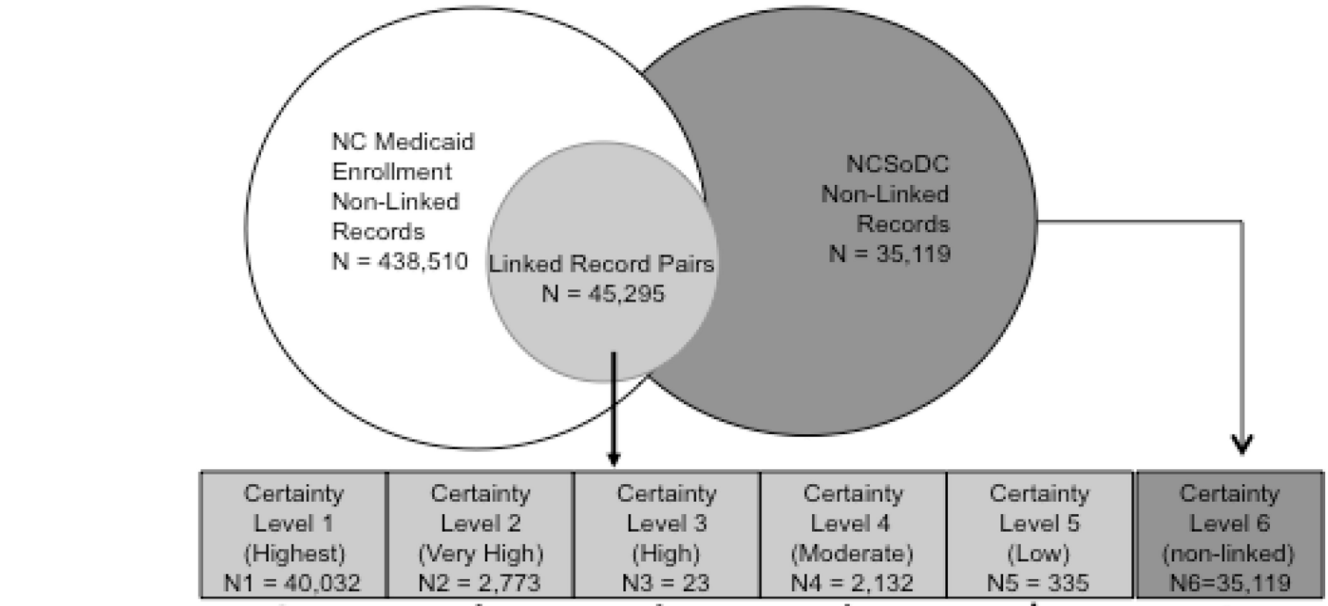
15. Rozier RG, Stearns SC, Pahel BT, Quinonez RB, Park J. How a North Carolina program boosted preventive oral health services for low-income children. Health Aff (Millwood). 2010 Dec; 29(12): 2278–2285. [PubMed: 21134930]

16. Sturmer T, Funk MJ, Poole C, Brookhart MA. Nonexperimental Comparative Effectiveness Research Using Linked Healthcare Databases. Epidemiology. 2011 May; 22(3):298–301. [PubMed: 21464649]

17. Chetty VK, Zellner BB. Use of survey and clinical data for screening and diagnosis. Statistics in Medicine. 2007; 26:3213–3228. [PubMed: 17230454]

18. McNamee R. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. Statistics in Medicine. 2002; 21:3609–3625. [PubMed: 12436459]

19. Brickhouse TH, Rozier RG, Slade GD. The effect of two publicly funded insurance programs on use of dental services for young children. Health Serv Res. 2006 Dec; 41(6):2033–2053. [PubMed: 17116108]



Expert review: Were the two records a match?	Sample n1=50	Sample n2=50	Sample n3=20	Sample n4=50	Sample n5=60	Sample n6=50
yes	45	43	14	37	44	4
no	5	7	6	13	16	46
Analysis weight	800.6	59.0	1.1	46.1	6.2	701.9

Figure 1.

The data merging process resulted in 56% of the NCSoDC records being linked to a NC Medicaid record. Numbers of linked record pairs (N_i; i=1,2,3,4,5) varied widely across certainty levels. We defined non-linked record pairs as certainty level 6 (N=35,119). Analysis weights were equal to the inverse of the fraction sampled (n_i/N_i) for each certainty level.

Table 1

Estimated Sensitivity and Specificity for each certainty level that indicates how certain it is that linked pairs are a true match

Certainty Level	Total N	n sampled	Sensitivity (Sn)	Specificity (Sp)
Level1: Highest	40,032	50	0.837 (0.785, 0.892)	0.893 (0.816, 0.976)
Level 2:	2,773	50	0.892 (0.838, 0.949)	0.882 (0.807, 0.965)
Level 3:	23	20	0.892 (0.838, 0.950)	0.882 (0.807, 0.965)
Level 4:	2,132	50	0.929 (0.874, 0.988)	0.867 (0.792, 0.949)
Level 5:	335	60	0.935 (0.879, 0.994)	0.865 (0.790, 0.947)