Case Study of Linking Dental and Medical Healthcare Records

Mary Kay Theis, MA, MS; Robert J. Reid, MD, PhD; Monica Chaudhari, MS; Katherine M. Newton, PhD; Leslie Spangler, VMD, PhD; David C. Grossman, MD, MPH; and Ronald E. Inge, DDS

Objective: To link the administrative data of a large dental carrier and an integrated health plan in Washington State to conduct an observational study of diabetes and periodontal disease.

Study Design: Evaluation of variable suitability, testing of linkage variables, and performing an *n* – 1 deterministic linkage strategy.

Methods: We examined a variety of administrative data variables for their consistency over time and their information richness to use as matching variables. After choosing social security number, date of birth, first name, and last name, we tested their reliability as linking variables among a population with dual dental and medical insurance. Lastly, we performed four *n* – 1 deterministic linkage steps to obtain our study population.

Results: With a success match rate of more than 96% with the 4 test variables, we extracted the entire population who met the study criteria with the understanding that only a subset would successfully link. We linked 78,230 individuals (55.2% of the Group Health Cooperative population). Of these matches more than 50% occurred within a last name–first name–birth date deterministic match.

Conclusions: Employer groups who provide dental-medical benefits for their employees send identical administrative data to dental and healthcare plans. The *n* – 1 deterministic linkage was accomplished by using a relatively straightforward approach because these data were fairly homogeneous and of high quality. Until medical care and dental care are integrated, it is possible to link these data to assess the impact of oral disease on overall health.

(Am J Manag Care. 2010;16(2):e51-e56)

For author information and disclosures, see end of text.

An emerging area of research interest is the intersection of medical and dental services, where the goal is to understand the potentially bidirectional relationships between dental and medical health, treatments, and outcomes. Evidence is accumulating on the association between dental conditions and diverse medical conditions including diabetes, cardiovascular disease, and chronic renal disease.[1-7] As a result, expert groups and advisory bodies are recommending better integration of medical and dental care.[8,9] However, medical and dental insurance and care systems have evolved separately with little integration, making it difficult to study the joint effects of medical and dental conditions. To conduct observational studies of the joint effects of medical and dental conditions, record linkage is required. Record linkage is a common tool for health researchers but has been infrequently used to link medical and dental care data.[10]

When no common identifier such as a medical history number is present to link data from unrelated organizations (as is the usual case for medical and dental records), the ability to successfully match individuals from the databases of interest is strongly dependent on the amount, accuracy, and overlap of personal identifying information in the data files.[11] Combining information in multiple, partially identifying data fields (eg, surname, first name, date of birth) can create linking algorithms with high sensitivity and specificity.[12] Creating a successful linkage strategy depends on the number of linking variables, the information richness of those variables, and the accuracy with which they have been recorded. Having too few variables increases the likelihood of duplicate matches, and coding errors result in lost true matches. If substantial, such errors can result in misclassification bias in observational research studies.[12]

Deterministic matching and probabilistic matching are the 2 major methods of record linkage. At its simplest, deterministic linking requires an exact match on 1 or more identifiers between 2 databases. This linkage technique uses a reliable identifier such as social security number (SSN) and then performs verification using additional parameters. If records do not link, other relevant identifiers (eg, date of birth, name, sex) are used to iteratively link and verify the linkage. When reliable identifiers are available, deterministic approaches are usually preferred because they achieve ac-

In this article
Take-Away Points / e52
Published as a Web Exclusive
www.ajmc.com

Take-Away Points

Social security number, date of birth, first name, and last name were tested for reliability as linking variables among a population with dual dental and medical insurance.

- A simple *n* – 1 deterministic merge was used successfully to link the dental and medical records of individuals within a defined geographic region and a 5-year time interval, and with a permanent unique identifier in 1 of the datasets.
- Because insurance systems have generally evolved separately, record linkage is an important step to study the joint effects of medical and dental services on health and healthcare use.

ceptable results, are less cumbersome, and require less development time than probabilistic matching.[13-15] The challenge is that across healthcare entities, identifiers (SSN, name, and address) may be coded incorrectly, different data standards or structures may be used, and different rules for data retention or updating may be applied, resulting in high nonmatch rates. For instance, in a Dutch study of perinatal care and pregnancy outcomes with 8 linking variables, complex probabilistic linkage techniques resulted in 80% more links than a deterministic linkage requiring exact agreement.[16] To allow for minor discrepancies, some researchers allow for a less restrictive "*n* – 1" deterministic linkage where 1 of the linkage variables is allowed to differ.[15,16] Deterministic linkage rates also can be improved by simple methods to reduce uncertainty such as eliminating erroneous values, invalid dates, nicknames, prefixes, and suffixes.[17]

When linking data are poor or deterministic linkage achieves poor results, more complex probabilistic linkage techniques are needed. Probabilistic matching uses information from many variables and allows for disagreement. Agreement and disagreement weights are assigned to all possible record pairs depending on the degree of agreement and the number of different values possible. Records are declared linked if the total agreement weights achieve a predetermined threshold.[18-20] Software packages are available to perform probabilistic matching when simpler deterministic matching strategies fail.

The objectives of our study were to (1) quantify the prevalence of treated periodontal disease among diabetic patients age 40 to 74 years; (2) examine the association between periodontal disease control and diabetes control; and (3) quantify the difference in medical costs for diabetic patients who received periodontal services versus those who did not receive periodontal services. Our ability to conduct this study pivoted on our ability to link medical and dental records. We describe in this article the strategy used to link health and dental plan data from a large dental carrier and a large integrated health plan in Washington State. All study procedures were approved by the institutional review board at Group Health Cooperative (Group Health).

METHODS

Study Setting

Group Health is a nonprofit, integrated healthcare system that provides care to approximately 570,000 residents of Washington Sate and northern Idaho. Enrollment is through employer-sponsored programs, self-funded plans, individual and family plans, and Medicare and Medicaid plans. Washington Dental Service (WDS) is a dental benefits carrier in the state of Washington and provides comprehensive dental insurance to more than 2 million adults and children.

Data Sources

For more than 20 years, Group Health has collected monthly administrative, financial, and clinical data formatted in SAS and relational databases. At Group Health, detailed demographic data are collected directly from enrollees or from their employers. We identified the possible linkage variables of sex, SSN, name, date of birth, residential address, and telephone number because these variables were common to both Group Health and WDS. At Group Health, enrollment information is updated by verifying the enrollee's address and phone number at patient encounters, and changes in the enrollee's name or SSN are submitted periodically by their employers. To maximize linkage rates, we extracted all variants of each linkage variable that were retained by Group Health over the 60-month study period.

Washington Dental Service maintains 30 years of computerized dental history for all enrollees, which includes information on demographics, insurance coverage, providers, procedures, and claims. These data reside in an enterprise-wide warehouse.

At the time of enrollment, both Group Health and WDS assign a unique patient identifier to each enrollee. At Group Health, this identifier is permanently retained for each individual across all enrollment-disenrollment cycles, whereas at WDS a new unique identifier is assigned when an enrollee changes employer group. The SSN, while collected, is not used as the principal personal identifier in either plan and is missing for nearly 100% of dependents of the WDS subscriber.

Linkage Variables

To explore the linkage utility of personal identifiers in both databases, we initially examined 10 potential linkage variables for missing or invalid values, and consistency over time. For the 200,864 Group Health enrollees born

Table 1. Percent Change in Coding of Group Health Demographic Fields From January 2004 to January 2005 Among Enrollees Who Were Members in Both Months (N = 200,864)

Linkage Variable	Final Denominator[a]	No. of Variants	No. (%) Changed From January 2004 to January 2005
Sex	200,861	2	116 (0.1)
Date of birth[b]	200,864	10,958	753 (0.4)
Cleaned last name[c]	200,864	48,048	1098 (0.6)
Cleaned first name[c]	200,862	15,263	1116 (0.6)
SSN	197,576	197,608	447 (0.2)
Zip code	200,860	1889	12,441 (6.2)
State	200,861	49	633 (0.3)
City	200,860	1727	9381 (4.9)
Telephone number	194,001	152,538	34,653 (17.7)
Middle name initial	174,172	26	1471 (0.8)

Group Health indicates Group Health Cooperative; SSN, social security number.
[a]Excludes missing data.
[b]Birth date is not missing because analysis was restricted to persons with a known birth date in 2004 and 2005.
[c]Prefixes, suffixes, spaces, and punctuation marks were removed.

between January 1, 1932, and December 31, 1961 who were enrolled in both 2004 and 2005, we examined the availability and appropriateness for each field, and the percentage that changed from 2004 to 2005 (Table 1). Last and first names were "cleaned" by removing spaces, suffixes, prefixes, and punctuation.

None to very few missing or erroneous values were detected on identifiers that are verified at healthcare visits, including sex, last name, date of birth, first name, and address fields. Small numbers of missing or invalid values (<5%) were found for SSN and telephone contacts. The variables with the greatest number of unique values, and thus potentially the most discriminating, were SSN, telephone number, last name, first name, and date of birth. We found 49 variants for state, including military and other district codes (eg, Puerto Rico; Washington, DC) and 2 for sex, making them less useful as primary linkage variables.

As expected, we found significant year-to-year variability in telephone number (17.7%), city (4.9%), and zip code (6.2%), making these variables less suitable for deterministic linkage. Enrollee first name, last name, date of birth, and SSN had less than 1% year-to-year variability, suggesting these variables were of high quality, and thus we chose them as our 4 linking variables. Social security number was included as it has evolved as a proxy universal identifier in insurance and healthcare systems and was expected to have high concordance between Group Health and WDS datasets.

Test Linkage

The second step was to test our confidence in the 4 linkage variables. After removing duplicates, we performed 4 deterministic linkages with 4 insured populations known to be dually insured by Group Health and WDS: Group Health's Medicare Advantage plan, Group Health's individual and family plan, and 2 large commercial accounts. We limited the study analysis to persons who were enrolled in 2004 or 2006 when the insurance products were active and who were born between January 1, 1932, and December 31, 1961. To improve matching precision, the Group Health and WDS datasets were standardized and names were cleaned. The linking results show 96% or higher exact matches among the 4 tests (Table 2).

Comfortable with the high degree of matching within the 4 fields we selected (SSN, last name, first name, and birth date), we developed the entire linked dataset for our study of diabetes mellitus and periodontitis. We included all individuals who met the study criteria from WDS and Group Health with the understanding that only a subset of the population would successfully link. First, we performed 2 exact deterministic linkages as a comparison group. An $n - 1$ deterministic linkage strategy was then performed on 4 sets of merges with 3 of the 4 merge variables: (1) SSN, last name, first name; (2) last name, first name, birth date; (3) SSN, first name, birth date; and (4) SSN, last name, birth date. To obtain our final linked population, we grouped the four $n - 1$ deterministic merges into 1 dataset and removed all duplicates using Group Health's unique identifier.

Data Security

To protect confidentiality, all personal identifiers were stored on a computer file that only the Group Health ana-

Table 2. Test Match Merges of WDS Data to 4 Group Health Plans With Known WDS Dental Coverage[a]

Linkage Variables	WDS Records (n = 35,893)	Group Health Records (n = 43,735)	No. of Unique Matches (% Agreement With WDS)
SSN[b]	35,893	42,973	35,058 (97.7)
SSN + last name[c]	35,893	43,253	34,773 (96.9)
SSN + birth date[d]	35,893	43,163	34,907 (97.3)
Last name + first name + birth date[e]	35,884	43,458	34,433 (96.0)

Group Health indicates Group Health Cooperative; SSN, social security number; WDS, Washington Dental Service.
[a]Coverage was in 2004 for 3 health plans and in 2006 for a Medicare health plan.
[b]Duplicate SSNs were deleted from both datasets.
[c]Duplicate SSN–last name sets were deleted from both datasets.
[d]Duplicate SSN–birth date sets were deleted from both datasets.
[e]Duplicate last name–first name–birth date sets were deleted from both datasets.

lyst could access and were destroyed at the end of the study. Washington Dental Service sent SSNs separately from names and birth dates through a Web-based application (Secure File Transfer [SFT]). Access to this folder was restricted to the study analysts. Secure File Transfer uses the 128-bit Secure Sockets Layer encryption protocol that is the standard transmission protocol for secure information transfer over the Internet. User accounts within SFT are given specific and limited access to only the upload/download area associated with 1 study. The visibility or accessibility of files other than those of the specific study is prevented by the SFT design. Once the linkage was complete, personal identifiers were destroyed and replaced by study identifiers. To protect confidentiality and privacy, the study identifiers were used to combine medical and dental information on the same individuals. We used SAS software in all analyses (SAS Version 9.1 for Windows, SAS Institute, Cary, NC).

RESULTS

Group Health has fewer enrollees than WDS, which is reflected in the total number of records pulled for our final merge: 413,954 for WDS versus 155,625 for Group Health. These totals are the entire population that met our study criteria.

Among the 2 exact merges, SSN alone resulted in 41.4% linked records whereas the SSN, last name, first name, and birth date resulted in 36.3% linked records (Table 3). Among the four $n - 1$ deterministic linkages, the combination of last name, first name, and birth date had the highest linkage rate (50.6%). This result is explained by the nearly 32% (n = 132,032) of WDS enrollees whose SSN was missing because they were dependents of the primary subscribers; the SSN is not consistently collected for dependents. Fewer than 3% (n = 4035) of Group Health enrollees were missing the SSN. Although SSN is a unique identifier and is fairly accurate for linkage, if we relied on SSN alone we would have dropped 17,278 study subjects. The addition of SSN to our $n - 1$ deterministic model added an additional 2298 (3%) study subjects. Our final linked population (n = 78,230) is the result of grouping the four $n - 1$ deterministic merges into 1 dataset and removing all duplicates.

The final linked population was 52% female, compared with 41% and 56% of the unlinked populations in WDS (18.3% had missing sex information) and Group Health, respectively. Similarly, the linked population had a mean (±SD) age of 52.1 years (7.5 years) compared with 49.9 years (6.7 years) and 52.9 years (8.0 years) for the WDS and Group Health unlinked populations, respectively.

Although Group Health's unique identifier allowed us easily to perform four $n - 1$ deterministic merges, it was not critical to the matching success. Fewer than 0.1% (n = 64) of the last name–first name–birth date merge were duplicates. Of these duplicates, 14 enrollees had 2 different last names, 27 had different first names, and 23 had different birth days. If we restricted Group Health data to the last enrolled date, there were only 3 last name–first name–birth date duplicates. However, we would have lost 226 matches (78,230 − 78,004).

DISCUSSION

In this study we demonstrated a simple and successful approach to linking medical and dental records. The success of our linkage was based on a number of factors. First, we examined the reliability of many variables in the Group Health and WDS datasets and restricted matching to 4 identifiers with minimum errors or missing data. Second, we conducted test merges with these 4 selected variables among individuals known to have both WDS and Group Health coverage. Third,

■ **Table 3.** Exact Match and *n* – 1 Deterministic Linkage Results for WDS and Group Health Data Among Persons Age 40 to 74 Years Who Were Continuously Enrolled for 5 Years[a]

Linkage Variables	WDS Records (N = 413,954)	Group Health Records[b] (N = 155,625)	Number of Individuals Linked (% Agreement With Group Health)
Exact merge			
SSN	281,781	141,764	58,654 (41.4)
SSN + last name + first name + birth date	410,591	153,995	55,922 (36.3)
***n* – 1 deterministic merge[c]**			
SSN + last name + first name	403,844	151,575	56,523 (37.3)
Last name + first name + birth date	398,517	150,032	75,932 (50.6)
SSN + first name + birth date	405,855	150,972	56,212 (37.2)
SSN + last name + birth date	410,053	151,178	57,602 (38.1)
Total number of linked unique individuals			**78,230[d]**

Group Health indicates Group Health Cooperative; SSN, social security number; WDS, Washington Dental Service.
[a]Linkage variables were SSN, last name, first name, and birth date.
[b]From a total of 141,709 unique individuals.
[c]Duplicate SSNs, last names, first names, or birth dates were deleted from each merge group in both datasets.
[d]Among the Group Health population, 55.2% were linked (78,230/141,709).

our time frame was short (2002-2006) and within a relatively small region (ie, Washington State). Thus, personal identifiers that are subject to change over time remained relatively constant. Fourth, Group Health maintains monthly administrative records for all enrollees. Thus, Group Health had records of all name variations, and all their variants were used to merge with the WDS dataset, increasing the likelihood of name matching. The Group Health unique identifier that is permanently assigned to each member was used to ensure no duplicate matches.

Many studies that link disparate datasets have found that an individual's first name varies considerably. Differences may present as a nickname, abbreviation, different spelling, or initial. Hence, first name often is an unreliable variable for linked datasets. Our insurance data were fairly homogeneous and of high quality and completeness. We did not experience much variance in the first name between WDS and Group Health datasets. We believe employer groups send identical administrative data to dental and healthcare insurers. For example, some first names were structured with an initial followed by a first name (eg, M Kay) in both datasets. Matches were high even in the absence of SSN. There are numerous resources that describe methods to handle problematic data.[21]

Medical care and dental care in the United States have evolved as separate entities, each with its own governing bodies, professional terminology, research interest, and insurance realms. As a consequence, integration in patient care is lacking, especially with respect to preventive and chronic care. Because a growing body of research associates chronic health conditions to chronic dental conditions, these heterogeneous databases need to be successfully linked to enable further research. Until medical and dental healthcare systems are integrated into a single model with 1 health data system, we will need to rely on successful computerized record linkage methods to further our understanding of the health implications of dental and medical care. In this article we show that a relatively simple deterministic approach, within a narrow time frame and with demographic variables commonly collected by dental and medical insurance carriers, results in high-quality record linkage. Further research is required to determine whether similar success rates can be obtained with enrollment data from other dental and medical care plans.

Author Affiliations: From Group Health Research Institute (MKT, RJR, KMN, LS, DCG), Seattle, WA; and Washington Dental Service (MC, REI), Seattle, WA.

Funding Source: This study was funded by the Washington Dental Service and Group Health Cooperative.

Author Disclosures: Ms Theis and Drs Reid, Newton, Spangler, and Grossman report receiving research grants from Washington Dental Service. Dr Inge and Ms Chaudari are employees of Washington Dental Service, which provided funding for this study.

Authorship Information: Concept and design (MKT, RJR, MC, KMN, LS, DCG, REI); acquisition of data (MKT, MC, KMN); analysis and interpretation of data (MKT, RJR, MC, KMN, DCG, REI); drafting of the manuscript (MKT); critical revision of the manuscript for important intellectual content (MKT, RJR, MC, KNM, LS, DCG, REI); statistical analysis (MKT); obtaining funding (KMN, LS, DCG); and supervision (REI).

Address correspondence to: Robert J. Reid, MD, PhD, Group Health Center for Health Studies, 1730 Minor Ave, Ste 1600, Seattle, WA 98101. E-mail: reid.rj@ghc.org.

REFERENCES

1. Saremi A, Nelson RG, Tulloch-Reid M, et al. Periodontal disease and mortality in type 2 diabetes. *Diabetes Care.* 2005;28(1):27-32.

2. Bahekar AA, Singh S, Saha S, Molnar J, Arora R. The prevalence and incidence of coronary heart disease is significantly increased in periodontitis: a meta-analysis. *Am Heart J.* 2007;154(5):830-837.

3. Borrell LN, Kunzel C, Lamster I, Lalla E. Diabetes in the dental office: using NHANES III to estimate the probability of undiagnosed disease. *J Periodontal Res.* 2007;42(6):559-565.

4. Scannapieco FA. Role of oral bacteria in respiratory infection. *J Periodontol.* 1999;70(7):793-802.

5. Craig RG. Interactions between chronic renal disease and periodontal disease. *Oral Dis.* 2008;14(1):1-7.

6. Tomar SL, Asma S. Smoking-attributable periodontitis in the United States: findings from NHANES III. National Health and Nutrition Examination Survey. *J Periodontol.* 2000;71(5):743-751.

7. Wysen KH, Hennessy PM, Lieberman MI, Garland TE, Johnson SM. Kids get care: integrating preventive dental and medical care using a public health case management model. *J Dent Educ.* 2004;68(5):522-530.

8. Centers for Disease Control and Prevention. *Working Together to Manage Diabetes: A Guide for Pharmacists, Podiatrists, Optometrists, and Dental Professionals.* Atlanta, GA: National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, US Dept of Health and Human Services; 2007.

9. Powell VJH, Din FM. Call for an integrated (medical/dental) health care model that optimally supports chronic care, pediatric care, and prenatal care as a basis for 21st century EHR standards and products. Pittsburgh, PA: Robert Morris University; 2008. Electronic Health Record Position Paper.

10. Saver BG, Hujoel PP, Cunha-Cruz J, Maupomé G. Are statins associated with decreased tooth loss in chronic periodontitis? *J Clin Periodontol.* 2007;34(3):214-219.

11. Roos LL Jr, Wajda A, Nicol JP. The art and science of record linkage: methods that work with few identifiers. *Comput Biol Med.* 1986;16(1):45-57.

12. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int J Epidemiol.* 2002;31(6):1246-1252.

13. Liu S. Development of record linkage of hospital discharge data for the study of neonatal readmission. *Chronic Dis Can.* 1999;20(2):77-81.

14. Gill L, Goldacre M, Simmons H, Bettley G, Griffith M. Computerised linking of medical records: methodological guidelines. *J Epidemiol Community Health.* 1993;47(4):316-319.

15. Roos LL, Wajda A. Record linkage strategies. Part I: estimating information and evaluating approaches. *Methods Inf Med.* 1991;30(2):117-123.

16. Méray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *J Clin Epidemiol.* 2007;60(9):883-891.

17. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Proc AMIA Symp.* 2002:305-309.

18. Victor TW, Mera RM. Record linkage of health care insurance claims. *J Am Med Inform Assoc.* 2001;8(3):281-288.

19. Adams MM, Wilson HG, Casto DL, et al. Constructing reproductive histories by linking vital records. *Am J Epidemiol.* 1997;145(4):339-348.

20. Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc.* 1969;64(328):1183-1210.

21. Federal Committee on Statistical Methodology. *Record Linkage Techniques, 1997: Proceedings of an International Workshop and Exposition.* Washington, DC: Office of Management and Budget; 1997.