MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1122

The Impact of a Growing Minority Population on Identification of Duplicate Records in an Enterprise Data Warehouse

Scott L. DuVall[a-c], Alison M. Fraser[d], Richard A. Kerber[e], Geraldine P. Mineau[d,f], Alun Thomas[c]

[a] *Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA*
[b] *VA Salt Lake City Health Care System, Salt Lake City, Utah, USA*
[c] *Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA*
[d] *Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA*
[e] *Department of Epidemiology and Population Health, School of Public Health and Information Sciences, University of Louisville, Louisville, Kentucky, USA*
[f] *Department of Oncological Sciences, University of Utah, 2000 Circle of Hope, Salt Lake City, Utah, USA*

Abstract

Patient medical records are often fragmented across disparate healthcare databases, potentially resulting in duplicate records that may be detrimental to health care services. These duplicate records can be found through a process called record linkage. This paper describes a set of duplicate records in a medical data warehouse found by linking to an external resource containing family history and vital records. Our objective was to investigate the impact database characteristics and linkage methods have on identifying duplicate records using an external resource. Frequency counts were made for demographic field values and compared between the set of duplicate records, the data warehouse, and the external resource. Considerations for understanding the relationship that records labeled as duplicates have with dataset characteristics and linkage methods were identified. Several noticeable patterns were identified where frequency counts between sets deviated from what was expected including how the growth of a minority population affected which records were identified as duplicates. Record linkage is a complex process where results can be affected by subtleties in data characteristics, changes in data trends, and reliance on external data sources. These changes should be taken into account to ensure any anomalies in results describe real effects and are not artifacts caused by datasets or linkage methods. This paper describes how frequency count analysis can be an effective way to detect and resolve anomalies in linkage results and how external resources that provide additional contextual information can prove useful in discovering duplicate records.

Keywords:

Record linkage, Subpopulations, Minority population

Introduction

It is common in large healthcare databases for information to be collected at different times in different places by different people. The disparate and sometimes inconsistent manner in which information is collected can lead to fragmented pieces of a person's medical information being persisted. Multiple records belonging to the same person, but mistakenly thought to belong to different people are called duplicate records. Having a single person's medical information spread across multiple records increases the time it takes to retrieve information, increases the risk of providing an incomplete patient history, and ultimately can impact patient care [1]. It therefore is important to find and eliminate duplicate records. Duplicate records are found by comparing pairs of records in a process called record linkage. The dominant method for linkage is the probabilistic approach formalized by Fellegi and Sunter [2]. This method is used in the non-trivial case where record identifiers do not match perfectly, but are close enough that they may be identified as duplicates.

Duplicate records are often found by comparing pairs of records within a single database. This paper describes an alternative situation where a second, external database exists which can be used to help identify duplicates. The enterprise data warehouse (EDW) of the University of Utah Health Sciences Center is an aggregate of medical records generated from inpatient and outpatient settings. It is routinely examined internally for duplicate records and is also linked to the Utah Population Database (UPDB), an external resource containing family history and vital records. In this second comparison, when two or more records in the EDW link to the same UPDB record, they are marked as potential duplicates. The EDW staff is notified of potential duplicates, verifies and resolves them if needed. All records, even known duplicates, are linked to the UPDB as an additional check to EDW internal deduplication processes.

In this study, we compared the frequency of name values in records in the duplicate subset with records in the full EDW and UPDB and describe instances where records in the duplicate subset are not typical of the database at large. We provide considerations for others looking at duplicate records in healthcare databases that help detect and resolve anomalies in

linkage results with population characteristics and linkage methods that are applied.

Materials and Methods

Data Sources

The University of Utah Health Sciences Center maintains an EDW that contains records for more than 1.8 million people resulting from all inpatient and outpatient visits to its hospitals and clinics since 1993. The demographic data in the EDW comes from both patient administration systems and physician billing systems. The EDW maintains a demographic record for each patient that contains fields describing a person's names, date of birth, sex, Social Security Number, addresses, phone numbers, and information about spouse and next of kin.

The UPDB is a research resource administered by the Utah Resource for Genetic and Epidemiologic Research. It was created in the mid-1970's using family histories from the Utah Genealogical Society containing the genealogy of the descendents of the Utah Pioneers [3]. The UPDB has since added records of Utah births, marriages, divorces, and deaths along with diagnosed cancers and driver license records. Each of these data sources gives extra information that can aid in the matching process. Records for each individual are grouped together into a person record - the composite of the best information available about a single person from one or more UPDB records. The more than 7 million person records in the UPDB contain demographic and family history information about individuals. Because of the scale and diversity of sources used to create the UPDB, most families living in Utah are represented in it. Birth and marriage certificates are used to expand the genealogy records and some families span as many as eleven generations. These data can only be used for biomedical and health-related research; the privacy of individuals represented in these records and confidentiality of the data is strictly protected [4]. The ability to correlate genealogy, medical, and demographic information makes the UPDB a valuable resource that has been used in many research studies [5]. For example, the UPDB was instrumental in discovering genes related to breast cancer [6,7], melanoma [8], colon cancer [9], and several other diseases.

Linkage Methods

In the interest of investigating the heritability of disease, the medical records available in the EDW are regularly linked with UPDB person records. The staff that manages the UPDB complete this activity using software that implements probabilistic record linkage.

First, middle, and last names of a patient are compared directly to the first, middle, and last name fields in a UPDB person record. Names for spouse and next of kin are compared to records linked through genealogy with the respective relationship to a particular person record. Because the EDW contains the patient's mother's maiden name, it is compared with the record linked through genealogy that is the mother of a particular person record. Both addresses in the EDW are compared with the address histories in the UPDB. Although the UPDB does not contain a history of phone numbers, the home and work phone numbers in the EDW are included since phone numbers are common identifiers used in linking at other institutions. Both the EDW and the UPDB contain more than Male and Female values for sex, including Unknown and a few other medical classifications.

Statistical Analysis

The top 2,500 most common last name values in the EDW were empirically categorized as *Founder* for Northern and Western European names; as *Traditionally Hispanic* for names typical of Latin and South America; or as *Other Ethnicities* as a collective group of Asian, Middle Eastern, and Native American names. Categorization of ethnicity based on last name was determined based using lists of names common in countries and by searching the origin of the name.

We compare frequency counts of demographic field values in the set of records identified as duplicates with the EDW generally. If duplicates occur at random within the EDW we would expect that values in these two sets would have the same relative frequencies. Comparisons that reveal notable deviations from this expectation may indicate possible areas where the matching process can be improved. Regression lines were calculated for each category comparing frequencies in the duplicate set with the EDW and the EDW with the UPDB.

Results

Of the 1,850,683 demographic records in the EDW, 1,375,704 were linked to UPDB person records. Of those, 209,852 EDW records linked to UPDB records that were simultaneously linked to by other EDW records; these were marked as potential duplicates and were used in this analysis.

Figure 1 shows the frequency of the 2,500 most common last name values in the EDW and duplicate set.

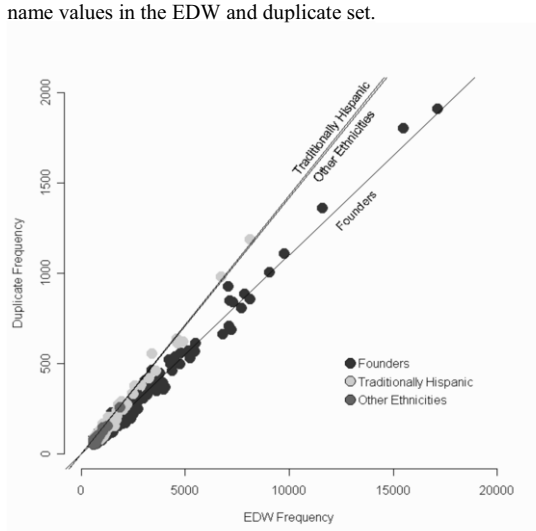

Figure 1 - Comparative frequency of ethnicity in the EDW and duplicate set

The name values are separated by assigned ethnicity with linear regression trend lines for each showing that Traditionally Hispanic names and names of Other Ethnicities are overrepresented in the duplicate set.

Figure 2 shows the frequency of the same 2,500 most common last name values in the EDW and UPDB.

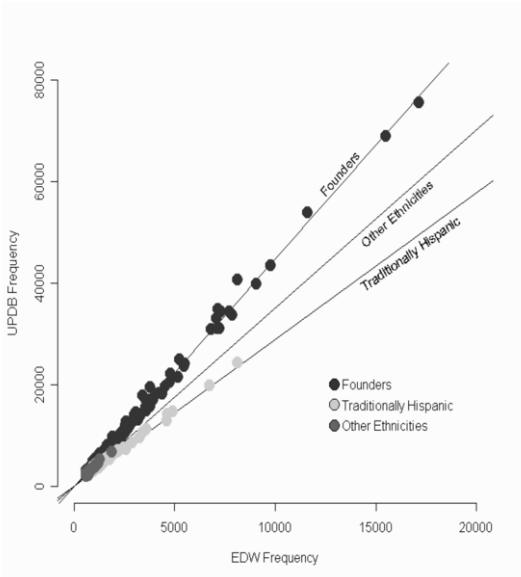

Figure 2 - Comparative frequency of ethnicity in the EDW and UPDB

The separation between Traditionally Hispanic names and those common among Founders in this comparison is even greater.

Discussion

An interesting artifact revealed in the linkage process was a division between names common among Utah founders and those common in the state today. First identified with the value Maria, a first name common in the overall population that is even more common among Hispanics, it became apparent that many of the names overrepresented in the duplicate set were traditionally Hispanic. Although less pronounced because much less frequent in the population, common Vietnamese, Korean, Chinese, Navajo, and Arabic names exhibited the same overrepresentation in the duplicate set. It was important to understand whether the higher proportion of Hispanic names existed in the duplicate record as an artifact of the linkage process or a real effect of the changing population.

The UPDB historically contains information that reflects the mostly North and West European background of the initial settlers of Utah [10,11]. The cultural face of Utah is changing, however, including a recent increase in the Hispanic population of Utah. In 1970, 95% of the Utah population was white and non-Hispanic compared to 85% in 2000 [10]. During the decade between 1990 and 2000, the Hispanic Population in Utah increased by 138% while the overall population increased by only 30% [12,13] This trend has continued since 2000 as the Hispanic population in Utah has increased at least three times faster annually than the overall population of Utah [13]. In addition to the recent increase, more than half of the Hispanic population lives in Salt Lake County, the area serviced by the University of Utah and thus represented in the EDW [14]. The recent demographic changes may not be adequately reflected in the UPDB. Temporary migrants who receive care at a hospital or clinic, but do not remain in Utah long enough to have a life event recorded in the UPDB will not have a person record created.

It is possible that by performing a single linkage with records of persons from all ethnicities, that a record linking process may not appropriately weight name frequencies. For instance, Martinez is the most common Hispanic last name in the duplicates, but only the 5th most common last name overall in the EDW and only the 19th most common name in the UPDB; Torres is the 10th most common Hispanic name in the duplicates, 59th in the EDW, and 203rd in the UPDB; and so on. As many of the common names in the EDW are found much less frequently in the UPDB, the likelihood of records being classified as duplicates when values match may artificially be inflated. Performing linkage on individual subpopulations may produce a more accurate result set, though it may be difficult to correctly decide how to classify records, particularly in cosmopolitan populations with diverse and inter-marrying ethnic populations. Other work suggests that it is not the ethnicity of an individual that causes linkage issues but the characteristics of names of certain origins that follow different naming conventions and phonetic rules than linkage tools are designed to consider [15]. Our work suggests that the uneven distribution of names between the datasets also affects linkage. Linkage artifacts caused by such a discrepancy may be less of an issue when values in both datasets are more balanced.

On the other hand, it is possible that a group may be legitimately overrepresented in the duplicate set. It is likely that names unfamiliar to registration clerks and other hospital staff would have an increased occurrence of misspellings. This could happen either during transcription, when a clerk enters information into a computer record from a paper sheet the patient filled out, or dictation, when a clerk writes or types information that a patient speaks. It may be the case for persons who do not speak English as their first language that information presented at different times and places may contain inconsistencies because of confusion, miscommunication, different traditions and feelings about record keeping, or the possible use of translation services.

Additionally, many database designs, including the EDW and the UPDB, hold that a person's name consists of a first name, a middle name, and a last name. In many cultures, this is not the case. Hispanic names often include more than one first or middle name, and it may be appropriate to use different last names in different situations. Asian names often have the last or family name presented before the first or given name. While these format variations may fool a hospital system in initially creating duplicate records, many commercial linkage systems

contain algorithms to recognize and correct these name differences. The ability of the EDW-UPDB linkage to classify these types of records as duplicates may account for their overrepresentation.

The overrepresented Hispanic names do not necessarily mean that the entire Hispanic population is overrepresented, but could be restricted to further subpopulations. The undocumented Hispanic population of Utah was estimated at between 55,000 and 85,000 in 2005 [16]. It may be less common that this group discloses complete and consistent demographic detail during medical visits [17]. Undocumented workers are more likely to be uninsured than either Hispanic or non-Hispanic legal residents and may receive care at the University of Utah which, as a state funded hospital, may have more flexibility to fund care for individuals who do not qualify for either federal or private reimbursement [18]. Care at the University of Utah indicates inclusion of records in the EDW and inconsistent information increases the likelihood of creating duplicate records. Recent immigrants are another subpopulation that may be less settled or more mobile. This may result in records being created at a number of different clinics that are later resolved as duplicates.

We found that the use of an external resource for discovering duplicate records in a healthcare database did affect which records were identified. We present the strengths and limitations of such a process along with considerations for those attempting such a linkage.

Strengths

Duplicate records are usually found by comparing sets of records within a single database and both the EDW and the UPDB undergo internal de-duplication as new records are added. Additionally, a linkage is made where the EDW is used as a database of interest and the UPDB as an external reference standard. Mistakes are made in de-duplication when dissimilar records are not matched, but are actually duplicates and when similar records are matched when they are not really duplicates. Using an external resource representing the population in the target dataset can provide the extra information and context needed to distinguish pairs that are truly duplicates and those that are not. The UPDB is such a resource that contains the majority of the population that receives healthcare from the University of Utah.

The additional information provided in links to family members and demographic field histories found in the UPDB allows duplicate records to be identified in the EDW that may not be found by other methods. For example, twins often have similar names, share a birth date, have the same parents, and may have the same address. Despite how similar these records are, the UPDB would show multiple births on each person's birth certificate and the two records could match properly to siblings instead of each other. As a further example, a woman who has recently married may have different last names and addresses on two records. Despite the records being dissimilar, UPDB person records would list her maiden name, her new last name and her husband's identity – obtained from a marriage license – a history of her known addresses, and a history of addresses for her husband. The two records could then be matched to the same person.

Limitations

Using an external reference for de-duplication may not eliminate the need for other methods of de-duplication. Duplicate records in the EDW cannot be found for individuals who do not have a record in the UPDB or where duplicates exist in the UDPB itself. Although the UPDB represents the population served by the University of Utah, it is not a true super-set. As a large academic research hospital, individuals may be referred from other states for specialized care. Others may receive care while visiting, but not living in the state. In addition, even for persons living in Utah to be included in the UPDB, they must have a life event that triggers the creation of a record.

Conclusion

The EDW and UPDB have different record characteristics and forces acting on them. Information is collected independently and for different purposes. When two different datasets are used for linkage, especially when they are collected at different times and for different purposes, a portion of the results may be explained by dataset differences. It is important to know when anomalies occur and if they describe real effects or artifacts caused by the datasets.

The changing face of the population represented in these datasets shows how subpopulations and changes in demographic trends may affect linkage. It is possible that segmenting data into homogenous demographic groups may lessen the impact that minority populations have on linkage results.

Understanding the impact of dataset characteristics and record linkage methods is a first step in improving duplicate record detection. We suggest the use of frequency count analyses as an effective way to detect anomalies in linkage results and as a tool for validating records identified as duplicates.

Acknowledgements

SLD was funded for this work by training grant # LM007124-11 from the National Library of Medicine and Robert Wood Johnson Foundation. The authors wish to thank Laverne A Snow and Reed M Gardner for collaboration on related projects, David E Avrin, Matthew H Samore, and Kerry G Rowe for graduate committee oversight. Partial support for all datasets with in the Utah Population Database (UPDB) was provided by the University of Utah Huntsman Cancer Institute. Support for this project was also provided by the Pedigree and Population Resource Group at the University of Utah Huntsman Cancer Institute and the Division of Genetic Epidemiology in the University of Utah Department of Biomedical Informatics. This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding support from the Veterans' Informatics, Information and Computing Infrastructure (VINCI), VA HSR HIR 08-204; the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08- 374; and the CDC-Utah Center of Excellence in Public Health Informatics, CDC 5P01HK000030.

References

[1] Mays S, Swetnich D, Gorken L: Toward a unique patient identifier. Florida IDN attacks duplicate records with MPI software, consultation and a shift in organizational philosophy. Health Manag Technol 2002, 23:42-44.

[2] Fellegi IP, Sunter AB: A Theory of Record Linkage. J Am Stat Assoc 1969, 64:1183-1210.

[3] Skolnick M, Bean L, Dintelman S, Mineau G: A computerized family history database system. Sociol Social Res 1979, 63:506-523.

[4] Cannon Albright LA: Utah family-based analysis: past, present and future. Hum Hered 2008, 65:209-220.

[5] Wylie JE, Mineau GP: Biomedical databases: protecting privacy and promoting research. Trends Biotechnol 2003, 21:113-116.

[6] Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, et al.: A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994, 266:66-71.

[7] Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, et al.: Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science 1994, 265:2088-2090.

[8] Cannon-Albright LA, Goldgar DE, Meyer LJ, Lewis CM, Anderson DE, Fountain JW, Hegi ME, Wiseman RW, Petty EM, Bale AE, et al.: Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. Science 1992, 258:1148-1152.

[9] Groden J, Thliveris A, Samowitz W, Carlson M, Gelbert L, Albertsen H, Joslyn G, Stevens J, Spirio L, Robertson M, et al. Identification and characterization of the familial adenomatous polyposis coli gene. Cell 1991, 66:589-600.

[10] Perlich PS: Immigrants Transform Utah: Entering a New Era of Diversity. Utah Economic and Business Review 2004, 64:5-6.

[11] Cannon-Albright LA, Farnham JM, Thomas A, Camp NJ: Identification and study of Utah pseudo-isolate populations-prospects for gene identification. Am J Med Genet A 2005, 137A:269-275.

[12] 1990 Census of Population, General Population and Housing Characteristics: 1990, Geographic Area: Utah. (1990 DP-1). U.S. Bureau of the Census.

[13] 2000 Census of Population, Profile of General Demographic Characteristics: 2000, Geographic Area: Utah. (2000 DP-1). U.S. Bureau of the Census.

[14] Schaub K, Lund M, Abebe B, Maloney T, Jameson K, Holzner C: Mexico and Utah: A Complex Economic Relationship. Institute of Public and International Affairs, University of Utah; 2006.

[15] Branting LK: Inducing Search Keys for Name Filtering. In Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 906–914. Prague; 2007:906–914.

[16] Passel JS: Estimates of the Size and Characteristics of the Undocumented Population. Pew Hispanic Center 2005.

[17] Berk ML, Schur CL: The effect of fear on access to care among undocumented Latino immigrants. J Immigr Health 2001, 3:151-156.

[18] Goldman DP, Smith JP, Sood N: Legal status and health insurance among immigrants. Health Aff (Millwood) 2005, 24:1640-1653.

Address for correspondence

Scott L. DuVall, scott.duvall@utah.edu