Journal of Biomedical Informatics 43 (2010) 24–30

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators

Scott L. DuVall [a,b,*], Richard A. Kerber [c], Alun Thomas [a]

[a] Department of Biomedical Informatics, University of Utah, USA
[b] VA Salt Lake City Healthcare System, USA
[c] Department of Epidemiology and Population Health, School of Public Health and Information Sciences, University of Louisville, USA

ARTICLE INFO

Article history:
Received 26 September 2008
Available online 13 August 2009

Keywords:
Medical record linkage
Probability
Algorithms

ABSTRACT

Probabilistic record linkage is a method commonly used to determine whether demographic records refer to the same person. The Fellegi–Sunter method is a probabilistic approach that uses field weights based on log likelihood ratios to determine record similarity. This paper introduces an extension of the Fellegi–Sunter method that incorporates approximate field comparators in the calculation of field weights. The data warehouse of a large academic medical center was used as a case study. The approximate comparator extension was compared with the Fellegi–Sunter method in its ability to find duplicate records previously identified in the data warehouse using different demographic fields and matching cutoffs. The approximate comparator extension misclassified 25% fewer pairs and had a larger Welch's T statistic than the Fellegi–Sunter method for all field sets and matching cutoffs. The accuracy gain provided by the approximate comparator extension grew as less information was provided and as the matching cutoff increased. Given the ubiquity of linkage in both clinical and research settings, the incremental improvement of the extension has the potential to make a considerable impact.

Published by Elsevier Inc.

1. Introduction

Information in large healthcare databases is entered during the course of patient care, through administration and billing processes, and in research studies. Data may be entered by persons with different roles, for different purposes, and with varying amounts of detail. Records with missing information or with variations of a person's name, address, and other personal information can result in the creation of duplicate records where data for one person is mistakenly placed in records thought to belong to different people. When this occurs, the safety and wellbeing of patients is put at risk. Previously recorded information that is not available during the time of treatment has been associated with a longer hospital stay, delayed care, additional services, and extra costs [1–3]. While duplicate records are not the sole cause of missing information, they can increase the time it takes to retrieve information, increase the risk of providing an incomplete patient history, and ultimately impact patient care [4]. Duplicate records are costly to find and resolve [5]. Correctly identifying which records belongs to which patients is an important part of both care delivery and research. Ensuring that patient information from different records is gathered together correctly and provides an accurate representation of the patient's health history is an issue central to healthcare's current trend toward interoperability and information exchange.

To discover which records are duplicates the demographic information associated with each record is compared. When two records have similar demographic information a determination can be made of how likely it is that the records belong to the same person. Fellegi and Sunter formalized the probabilistic method commonly used to make this determination [5–13]. In the Fellegi–Sunter (FS) method, each demographic field is assigned an agreement weight and a disagreement weight. These weights are log likelihood ratios based on the ability of field values to discriminate between records and the probability that the values contain errors. For example, sex has poor discrimination because there are very few options. Last name has high discrimination because there could be hundreds of thousands of possible values. On the other hand, the reliability of last names suffers because of alternate spellings, misspellings, and typographical errors that would very rarely be found in values of sex. Once field weights are determined, values in each field in one record are compared to the values in another record. When the values match exactly, or are determined sufficiently similar in some cases, the field agreement weight is added to a score. Otherwise the field disagreement weight, which is often a negative number, is added. This comparison is repeated for each field until a final score is calculated for the pair of records. The final score determines whether a

* Corresponding author. Address: Department of Biomedical Informatics, University of Utah, 26 South 2000 East, Room 5775 HSEB, Salt Lake City, UT 84112, USA. Fax: +1 (801) 581 4297.
E-mail address: scott.duvall@utah.edu (S.L. DuVall).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2009.08.004

record pair will be considered a match, not a match, or a possible match. Possible matches can be manually reviewed and reclassified or automatically reassigned as matches for pairs scoring above a threshold [10].

There are many active areas of research surrounding probabilistic linkage [7], but the two addressed in this paper are incorporating approximate field comparators in the FS method and estimating optimal field weights. Linkage methods compare pairs of records to determine whether they match or not, so the ideal output would be a binary decision. In fact, the FS method is optimized when the third category – possible matches – is minimized. A limitation of the FS method, though, is that each field has only two possible weights: one for agreement and one for disagreement. This forces a binary decision for each pair of field values – an all or nothing agreement – in addition to the classification from a final score of all fields.

To allow room for error, approximate comparators are used. An approximate comparator is an algorithm that determines how closely two values match. For example, while *Joe* and *Joseph* do not match exactly, a human reviewer and a good approximate comparator should both be able to tell that these values are much more likely to refer to the same person than values such as *Joseph* and *Bradley*. Taking into account approximate matches is important because fields like first name and last name have been shown to include misspellings and typographical errors in up to 25% of records [8,14]. There are classes of approximate comparators that deal specifically with names, addresses, dates, and other string and numerical values. Instead of binary output, approximate comparators often have a range of output values. Rather than match or not a match, output could be the number of days between two birthdays or the number of letters that are different in two names.

Because the FS method assigns field weights based on binary agreement or disagreement, approximate comparators can only be used when a cutoff is set to classify the comparison's output as a match or not a match. If two values are similar enough to score above the cutoff, the field agreement weight is given. Otherwise the values are not counted as matches and the field disagreement weight is given. Using cutoffs ignores the additional information gained from using approximate comparators because an exact match is given the same weight as a partial match. Rule-based methods can be developed to take advantage of partial agreement, can be easier to implement, and may perform well [13,15,16] – but rule-based systems also have limitations. The rules may be created empirically, require maintenance, have no statistical justification, and be difficult to repeat [9,13]. They may also have less discrimination than probabilistic systems [10,15]. Probabilistic systems are thought to more closely resemble human judgment and skill in evaluating discrepancies [11,13].

Approximate agreement has been addressed recently by expanding the probabilistic method to include a third field weight for close agreement [12]. We propose a method that incorporates the approximate comparator raw scores into the field weights, creating field approximate agreement weight functions. This method avoids the loss of information encountered by classification.

2. Material and methods

2.1. Data

The enterprise data warehouse (EDW) of the University of Utah Health Sciences Center contains demographic and clinical information on 1.8 million patients. The EDW is linked to a genealogical resource called the Utah Population Database (UPDB) that allows researchers to study the heritability of disease [17,18]. The UPDB is also linked to vital records and clinical data sources that provide additional demographic information including a history of name and address changes. The additional information allows duplicate records in the EDW to be identified that may not have enough data to be found otherwise. One example where this commonly occurs is when women marry, which often includes changing a last name, phone number, address, and assigning the new spouse as next of kin. The UPDB may contain links to a marriage certificate and an updated driver license record showing the address change that provide enough information to identify these types of records as belonging to the same people.

In the current EDW–UPDB linkage, 118,404 EDW record pairs were identified as potential duplicates. Records were counted as duplicate when two or more records in the EDW were linked to the same record in the UPDB. Because these records are linked only after a rigorous process of both automated and manual review, the increased certainty of links from the additional information provided by the UPDB, and the careful curation of the UPDB by a professional staff, this duplicate set was used as a reference standard for the study. A sample of 118,404 pairs was randomly selected from the remaining portion of the EDW that did not contain duplicate records.

2.2. Extension of the Fellegi–Sunter probabilistic method

Field weights in the FS method are log likelihood ratios formally specified in terms of m and u, where m is the probability that the field values match in a duplicate pair and u is the probability that the field values match in a non-duplicate pair. The field agreement and disagreement weights are therefore:

$$\log\left(\frac{m}{u}\right) \text{ and } \log\left(\frac{1-m}{1-u}\right) \tag{1}$$

For the approximate comparator extension (ACE) of the FS method, let δ be the difference between field values as measured by an approximate comparator. The probabilities m and u become functions $m(\delta)$ and $u(\delta)$ defined as the probability that the field values differ by δ in a duplicate or non-duplicate pair, respectively. The agreement and disagreement weights are then replaced with an approximate agreement function for each field expressed as:

$$\log\left(\frac{m(\delta)}{u(\delta)}\right) \tag{2}$$

Although not a necessity, a value difference of $\delta = 0$ corresponds to a perfect match in all our measures.

In the process of linking, the approximate agreement weight functions are substituted for the agreement and disagreement weights. When field values are compared, the difference between the values is used to look up the appropriate weight from the field approximate agreement weight function. Weights from each field are summed to a final score for the record pair.

2.3. Expectation maximization

Given a sample of record pairs where it is known which are duplicate and which are not, it would be straightforward to estimate $m(\delta)$ and $u(\delta)$ using the corresponding sample frequencies. For example, we could estimate $m(\delta)$ and $u(\delta)$ for the last name field as:

$$\begin{aligned} m(\delta) &= \frac{\#\text{ duplicate pairs with last name values differing by }\delta}{\#\text{ total duplicate pairs}} \\ u(\delta) &= \frac{\#\text{ non-duplicate pairs with last name values differing by }\delta}{\#\text{ total non-duplicate pairs}} \end{aligned} \tag{3}$$

However, as it is unknown in advance which record pairs are duplicates and which are not, this becomes a classical missing data problem that can be addressed using the expectation maximization (EM) algorithm [19]. EM has previously been used in record