Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Improving record linkage performance in the presence of missing linkage data

Toan C. Ong [a,b,d,*], Michael V. Mannino [a], Lisa M. Schilling [b], Michael G. Kahn [c,d]

[a] University of Colorado, Denver, Business School, Denver, CO, USA
[b] Department of Medicine, School of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA
[c] Department of Pediatrics, School of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA
[d] Colorado Clinical and Translational Sciences Institute, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA

ARTICLE INFO

Article history:
Received 13 July 2013
Accepted 24 January 2014
Available online xxxx

Keywords:
Record linkage
Missing data
Data quality
Comparative effectiveness research
Quasi-identifiers

ABSTRACT

Introduction: Existing record linkage methods do not handle missing linking field values in an efficient and effective manner. The objective of this study is to investigate three novel methods for improving the accuracy and efficiency of record linkage when record linkage fields have missing values.
Methods: By extending the Fellegi–Sunter scoring implementations available in the open-source Fine-grained Record Linkage (FRIL) software system we developed three novel methods to solve the missing data problem in record linkage, which we refer to as: Weight Redistribution, Distance Imputation, and Linkage Expansion. Weight Redistribution removes fields with missing data from the set of quasi-identifiers and redistributes the weight from the missing attribute based on relative proportions across the remaining available linkage fields. Distance Imputation imputes the distance between the missing data fields rather than imputing the missing data value. Linkage Expansion adds previously considered non-linkage fields to the linkage field set to compensate for the missing information in a linkage field. We tested the linkage methods using simulated data sets with varying field value corruption rates.
Results: The methods developed had sensitivity ranging from .895 to .992 and positive predictive values (PPV) ranging from .865 to 1 in data sets with low corruption rates. Increased corruption rates lead to decreased sensitivity for all methods.
Conclusions: These new record linkage algorithms show promise in terms of accuracy and efficiency and may be valuable for combining large data sets at the patient level to support biomedical and clinical research.

1. Introduction

Electronic health records (EHRs) are being adopted across diverse clinical practice settings, enabling clinical investigators to access detailed longitudinal patient- and practice-level data not previously available [1–3]. Rapidly evolving sources of rich health and wellness data include personal medical records, electronic diaries, online social media, disease-specific virtual communities, registries, and real-time personal health monitoring devices. Important data for research also exists in operational, administrative, and financial systems, hence, relevant clinical and financial data often exist in many independent organizations [4,5] and these data sources represent both enormous opportunities for and significant challenges to clinical practice and research. Without an accurate and universal patient identifier, the full spectrum of available patient data is not easily linked, creating barriers to an integrated, comprehensive view of treatments, outcomes, and costs.

Record linkage methods combine independent data sources so that data belonging to the same patient are assigned a common identifier. Current record linkage methods use one or more non-unique fields, called quasi-identifiers, to link two records belonging to the same individual [6]. Quasi-identifiers are defined as fields that, when combined, may be able to uniquely identify an individual, such as date of birth and last name [7]. In medical settings, missing data, including quasi-identifiers, can occur due to multiple reasons, creating challenges for record linkage. For instance, patients may not provide required information or clinical workflows may not ensure complete and accurate data collection and documentation. In a study about data quality in electronic medical records of HIV patients, Forster found that the median missing data

* Corresponding author at: Biomedical Informatics Core, Colorado Clinical and Translational Sciences Institute, University of Colorado, Anschutz Medical Campus, 12401 E. 17th Avenue, Campus Box B141, Aurora, CO 80045, USA.
E-mail address: toan.ong@ucdenver.edu (T.C. Ong).

http://dx.doi.org/10.1016/j.jbi.2014.01.016
1532-0464/

rate across six observed variables – age, sex, CDC or WHO clinical stage at baseline and follow-up, CD4+ lymphocyte (CD4) counts and year of ART initiation – was about 10.9% [8].

Current record linkage methods determine match results based on the calculated similarity between two linking fields' values and a set of weights which determines the relative contribution of each linking field's similarity or dissimilarity to a final match score [9]. A number of methods for calculating distance measures that have different properties or optimizations for specific data types can be used to calculate similarity scores. However, it is not possible to calculate a distance if either of the two values is missing.

While multiple methods have been proposed to solve the problem of missing data in traditional statistical analytic settings [10], much less research has focused on solving missing-data problems in fields that are used to perform record linkage. A common approach is to remove record pairs that have any missing data in any record linking field. Another approach is to simply ignore the field with missing data in the linkage-scoring algorithm. In both cases, valid record pairs may be missed due to the removal of information available for linkage determination.

We have developed novel algorithms with the objective to correctly identify matching records despite the occurrence of missing data in record linkage fields. We sought to accomplish two key goals: (1) maintain computational efficiency and (2) maximize the accuracy (sensitivity and positive predictive value (PPV)) of the linkage mechanism. We adapted solutions used to resolve missing data in standard classification methods to the problem of missing data in record linkage [11,12]. The three novel approaches: Weight Redistribution, Distance Imputation, and Linkage Expansion, better leverage the data available and discard less data, thereby preserving more information for record linkage. *Weight Redistribution* removes fields with missing data from the set of quasi-identifiers and redistributes the weight from the missing attribute based on relative proportions across the remaining available linkage fields. *Distance Imputation* imputes the *distance* between the missing data fields rather than imputing the missing data *value*. *Linkage Expansion* adds previously considered non-linkage fields to the linkage field set to compensate for the missing information in a linkage field. This study implements and compares the performance of all three approaches.

2. Background

In a relational database, two records are linked using a common primary key that must be unique for every distinct object and can never be missing. An always-present universal patient identifier would represent a common primary key to link patient-related data across relational tables and different data sources. However, in the United States, a universal patient identifier is not available so a combination of quasi-identifiers is used to link records across different data sources.

2.1. Record linkage approaches

There are two main approaches to matching two records using quasi-identifiers: deterministic and probabilistic. Deterministic record linkage methods establish the linkage between two records based on the exact agreement/disagreement of a combination of fields [13]. The strength of the deterministic approach is simplicity, transparency, and acceptable results [13,14]. The pitfall of the deterministic approach is its inability to account for the similarity between quasi-identifier values during field comparison [15]. Deterministic approaches are unable to match records with typographical or phonetic errors.

Probabilistic methods determine the likelihood two records refer to the same person. The most widely used probabilistic record linkage method was initially proposed by Fellegi and later extended by Sunter [6,16]. The Fellegi–Sunter (FS) method requires each linkage field be assigned both a match and an unmatch weight, numeric values, which represents the ability of that field to discriminate correctly matched from correctly nonmatched records. In its original formulation, FS examines the two field values in a record pair, determines if the values are a match or unmatch, and assigns either the full match weight or the full unmatch weight for that linkage field. The same binary determination (match or unmatch) and assignment of the full match/unmatch weight is performed for all pairs of values for all linkage fields in a record pair. The final FS score is the sum of the assigned matched and unmatched weights. This final score is compared to arbitrarily set thresholds, based on linkage purpose, to determine matched, possibly matched, and unmatched record pairs. Possibly matched record pairs usually require human review and adjudication. A recent addition to the FS method determines optimal match and unmatch weights for linkage variables using the Expectation Maximization (EM) algorithm, replacing tedious manual methods for determining these critical values [17].

The original FS method considered each pair of quasi-identifier in a record pair to be either a match or a non-match and assigned the full match or unmatch weight accordingly [9]. Over the past 20 years, a number of distance measures for comparing strings and dates have been developed which have been used to calculate similarity scores for a pair of quasi-identifier values used in record linkage fields [18,19]. The original FS method has been extended to include distance methods, such as edit distance [20] and dice-coefficients [21,22] allowing quasi-identifiers to be considered a *partial* match if they are *approximately similar* [9]. In many record linkage algorithms, similarity measures are normalized to a 0–100 scale where 0 represents no similarity (infinite distance) and 100 represents identical values (zero distance). For a pair of variable values in a record-linking field, the similarity score is combined with normalized field weights, which map the original FS match and unmatch weights, onto a 0–1 continuous scale to calculate a field's relative contribution to the total linkage score. Using normalized similarity measures and normalized field weights, a perfect match on all record-linkage fields results in a match score = 100. A similarity score less than 100 reduces a field's normalized weight contribution, yielding a final match score less than 100. Including similarity measures into record linkage algorithms creates flexibility for errors such as typographical and phonetic errors [23]. Methods that combine field similarity (distance) measures with probabilistic scoring have been found to have better performance in comparison to the deterministic methods [9,15].

In addition to the probabilistic methods, more complex methods using naïve Bayes classifier have been developed for record linkage [24]. However, similar to probabilistic methods, naïve Bayes-based methods depend on the assumption that the linking fields are independent [25]. An advanced record linkage method using neural network and complex features rather than individual fields was proposed by Wilson [26]. A complex feature is formed by considering multiple fields simultaneously. For instance, instead of comparing only birth dates of the two records, death dates can be used to identify if a person in one record died before the person in the other record was born. Wilson claims that using both complex features and a complex classifier (e.g. neutral network) outperforms the traditional probabilistic method. Because the focus of this study is on approaches for improving the performance of existing record linkage methods in the presence of missing data rather than on developing completely new record linkage methods, we opted to use the most commonly implemented record linkage method (FS).

Please cite this article in press as: Ong TC et al. Improving record linkage performance in the presence of missing linkage data. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.016