Embracing the Sparse, Noisy, and Interrelated Aspects of Patient Demographics for use in Clinical Medical Record Linkage

Stephen M. Ash, MS1,2, King Ip-Lin, PhD1
1University of Memphis, Memphis, TN; 2ARGO Data, Richardson, TX

Abstract

Duplicate patient records in health information systems have received increased attention in recent time due to regulatory incentives to integrate the healthcare enterprise. Historically, most patient record matching systems have been limited to simple applications of the Fellegi-Sunter theory of record linkage with edit distance based string similarity measurements. String similarity approaches ignore the rich semantic information present by reducing it to a simple syntactic comparison of characters. This work describes an updated approach to building clinical medical record linkage systems, which embraces the unavoidable problems present in real-world patient matching. Using a ground truth dataset of a real patient population, we demonstrate that systems built in this fashion improve recall by 76% with little reduction in precision. This result empirically demonstrates the size of the gap between sophisticated systems and naïve approaches. Additionally, it accentuates the difficulty in estimating the false negative error in this setting as previous research has reported much higher levels of recall, due, in part, to measuring from biased samples.

Introduction

Despite over 40 years of research, the problem of patient matching is far from solved due to a number of factors: (1) there are no broadly accepted, real world datasets to compare the performance of different approaches. Research from the healthcare community is either based on synthetically generated data or internal data sets which cannot be released; (2) much research has focused on the general problem of record linkage as it applies to any domain using publically accessible data sets such as business names, addresses, and bibliographic citations. Healthcare data differs in some ways from these data sources, and it is unclear how reliably results will transfer across domains.

If a practitioner wanted to build a new patient record linkage system, a survey of the literature would suggest a particular methodology, characterized by features such as: a Fellegi-Sunter1 based system using (1) edit distance based string similarity measurement to compare demographic attributes; (2) the assumption of conditional independence in order to reduce the count of model parameters; (3) a maximum likelihood approach to estimating parameters. Recent work shows the value of adding interaction terms to combat the negative impact of assuming conditional independence2. Other work has demonstrated the feasibility of using supervised learning techniques over maximum likelihood3,4. However, in both cases the inputs into their matching systems were either exact string matching or simple string edit distance based similarity. Previous work incorporating partial demographic agreement 5,6,7 is limited to utilizing the edit distance values (in one way or another) to augment field agreement or disagreement weights.

The main contributions of this paper are (1) a description of an effective approach to incorporate sophisticated techniques to measure demographic similarity and combat error due to conditional independence; (2) performance analysis using these techniques on real-world patient data, which quantifies the gap between systems using only simple string similarity versus systems using more sophisticated means. We feel that this is our most important contribution.

Figure 1: Typical processing pipeline for patient matching
as we believe that the false negative error has been substantially underestimated in past studies due to the inherent difficulty in determining the true duplicates in a large clinical patient population.

**Background**

In the clinical setting, matching occurs at the point of registration for a patient encounter. The input is patient demographics, and the output is a ranked list of possibly matching records. Figure 1 shows the typical processing pipeline for an industrial patient matching system. Previous work describes each step in this pipeline in greater detail. The overall quality of the patient matching system depends on the quality of each step in this pipeline. Here we focus on the weigh agreement vector step, which has received significant attention from the research community.

Probabilistic patient matching systems are built upon the Fellegi-Sunter theory of record linkage (FS), which describes an optimal decision rule to classify a record pair-wise agreement vector into match (M), maybe match (C), and non-match (U) classes. The agreement vector contains one dimension for each demographic attribute (e.g. Given name, Family name, Sex), with the value of the dimension showing the level of matches for that attribute.

Traditionally, string similarity algorithms such as Damerau-Levenshtein edit distance, Jaro-Winkler similarity, or Longest Common Subsequence are used to compare each demographic attribute, using a threshold to determine if the comparison is considered similar. However, we believe the quality of patient matching can be greatly impacted by the system’s ability to extract information from the data and semantically compare the information to produce agreement vectors. In the previous example, ‘Stephen’ and ‘Steve’ are semantically similar (common name aliases) despite being different string values. In addition, researchers have devised schemes to scale the agreement/disagreement weight scores based on partial string similarity or to use partial agreement as states in the agreement vector, increasing model complexity. However, previous approaches only handle scaling individual field weights based on a single definition of partial agreement per field. The kinds of noisy patterns that occur across multiple fields benefit from a more comprehensive approach.

**Design of the FS-Plus Patient Matching System**

FS-Plus follows the same general schematic as shown in Figure 1. At each stage of the pipeline, specialized algorithms were developed to extract as much information from the noisy data as possible. In the parsing, standardization, and filtering stages, sophisticated parsing methods were created using techniques described in data mining literature. Large dictionaries and gazetteers were hand curated to take advantage of real-world knowledge when determining the semantic meaning of particular tokens in the input records.

When calculating the overall agreement or disagreement weight for a record pair, FS-Plus augments the typical FS calculation in multiple ways—two of which are presented here: first, the individual field’s agreement and disagreement weights \((w_a, w_d)\) in the log-scale are interpreted to be maximum agreement and maximum disagreement with zero meaning no information. A field’s entry in the comparison vector is a set of semantic edit operations. Each edit operation has a discount associated with it. A discount is a value in \([-1, +1]\), which represents the normalized partial weight that should be applied if this edit operation were the only change required to go from the value in one record of the pair to the other. A discount of -1 corresponds to complete disagreement weight; +1 corresponds to complete agreement weight. When multiple edit operations are needed to transform one record to another, they are combined multiplicatively such that as more discounts are accumulated, the final discount rapidly approaches -1.

The semantic edit operations that are modelled in the system represent variance in the demographic fields introduced by mechanical, perceptual, or cultural phenomena. For example, accidentally hitting the ‘R’ key instead of ‘T’, because they are adjacent to each other on a QWERTY keyboard, is a mechanical error. Misreading an ‘o’ instead of a ‘p’ on a handwritten form is a perceptual error.

There are many intuitive heuristics a designer can implement as semantic edits described above. In many cases these operations are common across data sets and implementations. The cultural phenomenon of female’s changing their last name is as pervasive across North America as are QWERTY keyboards. By modelling and estimating the discount factors independently on a normalized scale, we allow these factors to be used across multiple data sets. In FS-Plus, we use profiling to determine a good starting point for each discount value, and then use a stochastic optimization procedure to refine the discount factor. This is a pragmatic way to incorporate partial agreement factors that can be refined over time across many data sets without having to dramatically expand the number of parameters which need to be estimated in the FS probabilistic model. FS-Plus models 50+ semantic edit operations including common name aliases, four different phonetic encoding algorithms, multiple common date of birth perception errors, and multiple name token alignment and segmentation phenomena.
The second way in which FS-Plus augments FS is to introduce configurable patterns of specific observation values that, when matched, reward or penalize the overall match weight. These act like interaction terms in a regression equation, allowing the combination of particular values of input variables to affect the overall score. Patterns are specified using a simple syntax in an XML file. Each pattern is a conjunction of simple Boolean predicates. This allows human knowledge to be added to the equation in a targeted way. For example, a pattern rule to describe a female of typical marriage age, who changed her last name, with no other contradictory information can be expressed as: \((\text{gender} = \text{F}, \text{age} > 18, \text{age} < 40, \text{first name weight} > 0, \text{last name weight} < 0, \text{SSN weight} \geq 0, \text{date of birth weight} > 1)\). Each pattern rule has a discount factor associated with it that is derived in a similar fashion as described for the semantic edit operations. Since patterns match multiple demographic fields, they present a pragmatic way to overcome some limitations from the conditional independence assumption, which is typical in FS implementations.

**Experimental Setup**

In order to compare the FS-Plus system to other FS approaches described in previous literature, we curated a high quality ground truth dataset of real-world patient records using a methodology adapted from the NIST Text Retrieval Conference (TREC)\(^ {13} \). We partnered with Just Associates, Inc. (JA), a consulting firm with more than 12 years of experience helping hospitals cleanup duplicate patient records. We used a complete dataset of ~530,000 patient records from a multi-facility, regional integrated health system. The dataset contained birth records, pediatric records, adults, and records of patients residing in retirement homes. Matching used the common patient demographic elements: Name (First, Middle, Last, Suffix), DOB, SSN, Address (Line1, Line2, City, State, Zip), Phone, Emergency Contact (Name, Phone Number, and Relationship). Many of these fields were only sparsely populated, and thus only provided value when present and valid. Given that the dataset is large, there is not enough time to evaluate all 140 trillion possible record pairs for duplication. Therefore, we pooled the results of multiple systems employing different algorithms, at lower than normal thresholds, in order to attempt to identify all of the possible duplicate patient records. Four systems were used to pool results, two of which were commercial EMPIs. This resulted in ~37,000 records involved in possible duplicate groups that needed to be human adjudicated. Between the authors and JA’s professional human adjudicators, all of the pooled adjudication tasks were evaluated, taking over 300 hours of human review. In some cases, the original medical records were consulted to disambiguate record pairs. The final ground truth revealed an overall duplicate rate of 6.9%, which is consistent with previous real world evaluations\(^ {14} \).

A number of different metrics have been presented in the literature to evaluate performance of matching algorithms. In this work, we calculate the pairwise Precision, Recall, F1-score, and Average Precision. Pairwise precision is analogous to **positive predictive value** and pairwise recall is analogous to the **sensitivity** metric in the domain of statistics.

We evaluated the following experimental setups: (1) **FS-Exact**: FS using exact match to determine attribute agreement. (2) **FS-BinaryThreshold**: FS using Jaro-Winkler on alphabetic fields and Levenshtein on numeric fields with a threshold to determine attribute agreement or disagreement. (3) **FS-Partial**: FS using the piecewise scaling function described by Winkler\(^ 9 \) to give partial agreement based on string similarity. This system is equivalent to the probabilistic matching approach available in open source EMPIs and to the probabilistic system evaluated by Joffe\(^ 3 \). (4) **FS-Plus**: Our system, utilizing all of the enhanced methods described above. In each experimental setup, we began with the same initial parameters and used bootstrapping to estimate model parameters. We used 26 different blocking schemes, utilizing combinations of phonetic encodings, name aliases, prefixes, substrings, and ordered bigrams of demographic attributes in order to cover typical demographic variance.

**Test Results**

|          | Precision | Recall  | F1      | AvgPrec |
|----------|-----------|---------|---------|---------|
| FS-Exact | 0.75002   | 0.41640 | 0.53549 | 0.37725 |
| FS-Binary| 0.73784   | 0.52331 | 0.61233 | 0.48081 |
| FS-Partial| 0.77205  | 0.55728 | 0.64731 | 0.51718 |
| FS-Plus  | 0.75432   | 0.97952 | 0.85230 | 0.93938 |

Table 1 Results showing increase in recall

Table 1 contains the test results for each setup. As indicated in the previous literature, using string similarity to scale the agreement weight (FS-Partial) increases recall. However, by utilizing specialized approaches in FS-Plus, we were able to improve recall by ~75% with little cost to precision.
Figure 2 shows the precision-recall curves for each level of recall. This chart is particularly important for the patient matching problem, in which there are not enough human resources to adjudicate the entire output of the matching process. Thus, it is important to ensure high precision at all levels of recall so that human resources are not wasting time on low quality match work. FS-Plus outperforms other approaches at all levels of recall. The Average Precision metric captures the area under the precision-recall curve. FS-Plus improves the Average Precision by ~76%.

![Precision-Recall curve comparing FS-Plus to others](image)

**Discussion**

These experiments set out to quantify the gap between the *information* extracted using traditional string similarity approaches versus more sophisticated approaches. FS-Plus finds more relevant pairs over previous approaches, because the semantic edit operations are tuned to specific cultural, perception, and data-entry phenomena. In order to narrow down which demographic elements contained the most *information* that is missed by traditional similarity measurement, we evaluated FS-Plus against itself, holding out the specialized algorithms for each demographic attribute in turn. Instead of the sophisticated semantic edit matching, we used string edit matching in the same way as the FS-Partial setup. Holding out sophisticated name matching, AvgPrec drops 22% to 0.7311; holding out the sophisticated DOB matching, it drops 20% to 0.7496; holding out fancy address matching, AvgPrec drops 4% to 0.9043; holding out fancy SSN matching drops AvgPrec 3% to 0.9083; holding out fancy phone matching drops AvgPrec 3% to 0.9101.

FS-Plus’s name measurement algorithm has the largest impact to recall, followed closely by DOB. These are often the most present data elements, but they also exhibit variance that is difficult to detect by only string similarity approaches. Investigating output of the runs, there were some distinct patterns that are pathological cases for string similarity algorithms. In this dataset the most common cases were: nickname tokens, transliteration and phonetic variations, truncation problems, and structural variations, where the name particles were permuted (e.g. Joan Smith-Cherry vs Joan Cherry vs Joan C Smith). In addition, it appeared as if some health record/registration systems truncated name fields after some maximum number of characters. This was particularly difficult when combined with hyphenated last names such as ‘Maria Hernandez’ versus ‘Maria Garcia-Hernan’. There were also a number of *fused* name tokens such as ‘Maria Hernandez’ versus ‘Maria Garcia Hernandez’. FS-Plus’s name agreement logic splits the name strings into tokens using character features (whitespace), linguistic features (‘de la’, ‘del’ particles), and name dictionary informed features. These algorithms, being specialized to the kinds of name variance observed in real patient data, increase recall substantially.

The second most powerful demographic was date of birth. Investigating cases in which FS-Plus bested other setups, we observed a number of DOB patterns that are hard to detect by Levenshtein edit-distance alone. Overall, 34% of the relevant pairs contained some variance in the DOB field. However, this variance appears to be disproportionately distributed to patient records of elderly and child patients, which is explainable by the fact that these classes are often less-likely to self-report their DOBs.

Through our partnership with Just Associates, FS-Plus has been run on dozens of real-world clinical patient datasets. Data quality and regional specific phenomena changes from dataset to dataset. However, our experience is that the biggest difference is in the distribution of values and proportions of phenomena. This variance is adequately addressed through data profiling and optimization. The overall heuristics (e.g. hyphenated last names, the general structure of addresses and DOBs, etc.) that are targeted in FS-Plus are common across most datasets. Thus, there is a positive
economy in adding this additional complexity, given the patient safety and cost-sensitive demands of clinical patient matching.

Researchers have run similar experiments in the past, trying to quantify the precision and recall of probabilistic matching algorithms\textsuperscript{15}. Some have published estimates that probabilistic approaches reach >95% recall in healthcare settings\textsuperscript{16}. Our results differ from these experiences due to a number of reasons: (1) patient data quality affects matching performance significantly. Data quality is often correlated to the data gathering process and the context under which data is collected. Our ground truth more accurately reflects typical clinical patient data. Previous results on cleaner data are not competitive on this real data. (2) Estimating recall is always difficult given that you do not know what you do not know. We are not aware of comparable studies utilizing large real world clinical patient records and multiple matching systems to build a high quality, pooled ground truth. Most studies use simple string similarity approaches to identify candidate pairs. As this work shows, that approach is likely to miss many relevant pairs. (3) A number of studies, such as Joffe\textsuperscript{3}, use limited blocking approaches that may bias the pairs that they consider when looking for true matches. Only blocking on one field, or only using simple blocking values such as phonetic encodings, is going to miss many relevant pairs and bias the error estimates.

In the future we want to use this rich collection of heuristics and techniques that model the semantic variance in patient duplicate records to generate much more realistic synthetic data. This will allow us to release synthetic versions of large datasets which are similar in hardness to real world data sets.

\textbf{Conclusion}

Previous evaluations have assumed simple string based approaches to measuring information similarity between demographic attributes. This work demonstrates that the rich semantic information embedded in demographic strings is worthy of additional effort to create a better matching solution.

\textbf{References}

1. Fellegi IP, Sunter AB. A Theory for Record Linkage. JASA 1969;64:1183–210.
2. Daggy J, Xu H, Hui S, Grannis S. Evaluating latent class models with conditional dependence in record linkage. Stat Med 2014;
3. Joffe E, Byrne MJ, Reeder P, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. J Am Med Inform Assoc 2014;21:97–104.
4. Joffe E, Byrne M, Reeder P. Optimized Dual Threshold Entity Resolution For Electronic Health Record Databases–Training Set Size And Active Learning. AMIA Annu Symp Proc 2013;
5. Porter EH, Winkler WE. Approximate string comparison and its effect on an advanced record linkage system. In: Advanced record linkage system. US Bureau of the Census, Research Report. Citeseer; 1997, page 190–9.
6. Tromp M, Reitsma JB, Ravelli a CJ, Méréy N, Bonsel GJ. Record linkage: making the most out of errors in linking variables. AMIA Annu Symp Proc 2006;;779–83.
7. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. J Biomed Inform 2010;43:24–30.
8. Winkler W. Overview of record linkage and current research directions. Bur Census 2006;
9. William E. Winkler and Yves Thibaudeau. An application of the Fellegi-Sunter model of record linkage to the 1990 US decennial census. 1990.
10. Churches T, Christen P, Lim K, Zhu JX. Preparation of name and address data for record linkage using hidden Markov models. BMC Med Inform Decis Mak 2002;16:1–16.
11. Schone P, Cummings C, Davey S. Comprehensive evaluation of name matching across historic and linguistic boundaries. Proc from Fam Hist Technol Work 2012;
12. Cooren Y, Clerc M, Siarry P. Performance evaluation of TRIBES, an adaptive particle swarm optimization algorithm. Swarm Intell 2009;3:149–78.
13. Voorhees EM, Harman D. Overview of the Eighth Text REtrieval Conference (TREC-8). 1999.
14. Grannis S, Overhage J. Analysis of identifier performance using a deterministic linkage algorithm. Proc AMIA 2002;;305–9.
15. Silveira D, Artmann E. Accuracy of probabilistic record linkage applied to health databases: systematic review. Rev Saude Publica 2009;43.
16. Grannis S, Overhage J, Hui S. Analysis of a probabilistic record linkage technique without human review. AMIA Annu 2003;;259–63.