Adjusting Match Weights to Partial Levels of String Agreement in Data Linkage

Dean M. Resnick[1], Lisa B. Mirel[2], Marc. I Roemer[3]
Scott R. Campbell[1]

[1] NORC at the University of Chicago, 4350 East-West Highway, Bethesda, MD 20814
[2] Centers for Disease Control, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782
[3]Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20851

Abstract

The Fellegi-Sunter record linkage paradigm in its original conception was based on the idea that for a set of comparison fields, such as first name, year of birth, and state of residence, agreement of each field between records in a pair is strictly binary: either there is complete agreement or there is not. For string comparisons, particularly for names fields, intuition tells us that having two versions of a name (e.g. 'Resnick' compared to 'Reznik') that are very similar but not identical is more indicative of a record pair being a match rather than a non-match. There are several string comparison tools such as Jaro-Winkler similarity scores and Levenshtein distances that can quantify the level of agreement as a full range of values between complete agreement and complete non-agreement. Certainly, one way of using such a metric is to establish a cutoff level above which we consider the fields essentially in agreement, but this would require a method of determining the cutoff. However, we are instead looking for a way to assess several gradations of agreement for string comparisons and assign agreement and non-agreement weights corresponding to the observed gradation. In this paper, we describe such a method that maintains and expands upon the Fellegi-Sunter approach.

Key Words: Record Linkage, Fellegi-Sunter, String Comparisons, Match Weights

1. Background

In a typical record linkage with person-level data, several of the identification fields (usually name fields) used to compare two records are text strings which may be indicative of match status even in cases that do not agree exactly. There is a wealth of literature on ways of conducting string comparisons (Chen 2012; Christen 2006; Jokinen 1996). Among methods used for string comparison are text editing and parsing, nickname conversion, phonetic encoding, and string similarity scoring. These methods all have potential value for record linkage analysis. Among these, string distance or similarity scoring is a particularly useful strategy. There are several edit distance metrics that are available for these comparisons: Levenshtein distance, Hamming distance, Damerau–Levenshtein distance, Dice coefficient, and Jaro-Winkler similarity score. These metrics evaluate the string distance over a range from complete similarity to essentially no similarity. Among these it is found that the Jaro-Winkler similarity score is an effective way of characterizing the similarity of names for use in record linkage (Budzinsky 1991). The Jaro-Winkler similarity score assigns a value between 0 and 1 for each string comparison, with 0 indicating no agreement (i.e., no characters in common) and 1 indicating complete agreement (Jaro 1989, Winkler 1990).There is a dearth of literature suggesting how to incorporate the value of the metric (be it Jaro-Winkler similarity score or some other) into a record linkage analysis. Within the Fellegi-Sunter paradigm (Fellegi and Sunter, 1969) it seems the simplest way to use the metric is to specify in some manner a fixed cutoff level for it. Values above the cutoff level would characterize a string comparison as an agreement, and scores below the cutoff value would be considered a disagreement. As with any other variable comparisons in Fellegi-Sunter, in the case of agreement, we would assign an agreement weight to the comparison, and in the case of disagreement we would assign a disagreement weight. However, several issues arise with introducing the use of the string comparison score. First, it is not clear exactly how to select the optimal cutoff level. Second, it is clear that all comparisons on the same side of the cutoff level would be assigned an equal agreement status and weight assignment even though some of those comparisons are more similar than others.

2. Methods

To assess a gradation approach for string comparators, a case study was conducted using data recently released through the National Center for Health Statistics (NCHS) Data Linkage Program (https://www.cdc.gov/nchs/data-linkage/index.htm).

2.1 Data Sources

Data from the 2016 National Hospital Care Survey (NHCS) were linked to the 2016/2017 National Death Index (NDI) (National Center for Health Statistics 2019). NHCS is an establishment survey of hospitals conducted by NCHS (https://www.cdc.gov/nchs/nhcs/index.htm). Each participating hospital provides a putatively complete set of patient encounter records within the period of analysis, a single calendar year. Patient records include personally identifying information (PII) including names, date of birth, and home addresses. Social security number (SSN) is included among the fields that are provided, but for a sizeable proportion (26.7%) of patient records, the SSN field is missing. The PII information is only available on the file for purposes of conducting data linkage and all research files are stripped of PII.

The NDI is a centralized database of death record information on file in jurisdictional vital records or statistics offices and maintained by NCHS (https://www.cdc.gov/nchs/ndi/index.htm). These data can be used to identify each person who has died in the United States and U.S. military overseas and his or her cause(s) and manner of death. NDI records include identifiers such as SSN, names, date of birth, and state of birth.

2.2 Linkage Methods Background

One salient feature of this linkage is that for a sizeable proportion of NHCS records, it is possible to use deterministically determined links based on SSN joins to estimate parameters, identify matches, and estimate error rates. After verifying the similarity of the other PII, the deterministically linked records form a truth set that can be used to estimate linkage parameters for linkage with NHCS records which were provided without an SSN. In particular, it allows the direct estimation of m probabilities, i.e. the proportion of the deterministically matched pairs in the truth source that agree on each of the common identification variables: $P(Agree|M)$, where M indicates match status.

Additionally, we have chosen to adjust weights for name comparisons based on observed frequencies for name. This means that agreement on a rare name produces a higher comparison weight than agreement on a common name, and is accomplished by computing u probabilities, $P(Agree|U)$, specific to each name, where U indicates unmatched status in the truth source: the records in the pair represent different entities—i.e., agreement on the fields is coincidental.

Our review of the literature yielded only one description of a method to transform a Jaro-Winkler similarity score to a comparison weight (Winkler, 1990). Here they use data from a truth source to compute a weight value that creates small intervals on the Jaro-Winkler similarity scale. For example, they would compute the ratio

$$R = \frac{P(\Psi \; \epsilon \; (k,l]|M)}{P(\Psi \; \epsilon \; (k,l]|U)},$$

where the numerator is the probability of the string comparator value Ψ falling in the interval from k to l for matched pairs M, and the denominator is the probability of the string comparator falling in the same interval for unmatched pairs, U.

Then the computed comparison weight would be

$W = log_2(R)$, which is consistent with the Fellegi-Sunter treatment.

By doing this for all of the intervals on the range (Winkler used from 0.6 to 1.0 by increments of 0.02), a table of weights is built. This process is repeated for several sets of similar linkages (i.e., all having the same identification variables). Next, a piecewise continuous function in the form of three line segments is fit to these tables of weights, which have been reviewed to determine they are essentially similar. The fitted line segment then forms a function, $W_i = f(\Psi_i)$, to relate the comparison weight for variable i to the string comparator value for it.

This method does have the desirable property that for any two comparisons, the one having the higher value of Jaro-Winkler similarity is assigned a higher weight. However, the method developed with this function, $W_i = f(\Psi_i)$, does not reflect name frequency; the similarity value for each name has a consistent relationship to the computed weight regardless of the name's commonness. In the current analysis, u probabilities specific to each name were computed as:

$$u_\psi(Name) = P(\Psi \geq \psi | N_A = 'Name' \cap U)$$

where N_A is a name value from file A.

These u probabilities are then incorporated into the computation of the weight. This development of name-specific weights is not possible with the direct mapping of the similarity of the string comparator value (Ψ) onto the comparison weight (W) as noted above (Winkler, 1990).

2.3 Proposed Method

We seek a method that conforms to the Fellegi-Sunter model while accounting for the frequencies of names. In the Fellegi-Sunter model, an agreement pattern is given by the vector $\gamma(a, b)$, where a is a record from file A and b is a record from file B. The components of the vector, $(\gamma_{1,} \gamma_2, \dots, \gamma_n)$ are the variable agreements, so component γ_1 could be agreement (0 = No; 1 = Yes) on first name, γ_2 on last name, γ_3 on year of birth, etc. Then assuming that the agreement patterns are independent among matched pairs (M), then then the probability that a matched pair has a particular agreement pattern is given by

$$P(\gamma(a,b)|M) = \prod_{i=1}^{n} P(\gamma_i = A_i|M), \qquad \text{(Eq. 1)}$$

where A_i is agreement status (0 or 1) for variable i.

Similarly, assuming independence, for unmatched pairs (U),

$$P(\gamma(a,b)|U) = \prod_{i=1}^{n} P(\gamma_i = A_i|U) \qquad \text{(Eq.2)}$$

Then, then number of matches with agreement pattern γ is

$$N_M(\gamma) \; = N_M \cdot \mathrm{P}(\gamma|U) = N_M \cdot \prod_{i=1}^{n} P(\gamma_i = A_i|M), \qquad \text{(Eq. 3)}$$

where N_M is the total number of matched pairs (i.e., with all possible agreement patterns) among the set being analyzed (as from all pairs developed within a single blocking pass), and the number of non-matches with agreement pattern γ is

$$N_U(\gamma) \; = N_U \cdot \mathrm{P}(\gamma|U) = N_U \cdot \prod_{i=1}^{n} P(\gamma_i = A_i|U)\,, \qquad \text{(Eq. 4)}$$

where N_U is the total number of unmatched pairs in the set.

Thus, among pairs with agreement pattern γ, the ratio of matched pairs to unmatched pairs is from

Eq. 4 and Eq. 5

$$R(\gamma) = \frac{N_M(\gamma)}{N_U(\gamma)} = \frac{N_M \cdot \; P(\gamma_1 = A_1|M) \cdot P(\gamma_2 = A_2\,|M) \cdot \dots \cdot \; P(\gamma_n = A_n|M)}{N_U \cdot \; P(\gamma_1 = A_1|U) \cdot \; P(\gamma_2 = A_2\,|U) \cdot \dots \quad P(\gamma_n = A_n|U)}$$

$$= \frac{N_M}{N_U} \cdot \left(\frac{P(\gamma_1 = A_1|M)}{P(\gamma_1 = A_1|U)}\right) \cdot \left(\frac{P(\gamma_2 = A_2|M)}{P(\gamma_2 = A_2|U)}\right) \cdot \dots \cdot \left(\frac{P(\gamma_n = A_n|M)}{P(\gamma_n = A_n|U)}\right) \qquad \text{(Eq. 5)}$$

And making a log transformation gives

$$log_2\big(R(\gamma)\big) = log_2\left(\frac{N_M}{N_U}\right) + \sum_{i=1}^{n} log_2\left(\frac{P(\gamma_1 = A_i|M)}{P(\gamma_1 = A_i|U)}\right) \qquad \text{(Eq. 6)}$$

Then, the terms from Eq. 6 of the form $log_2\left(\frac{P(\gamma_i = A_i|M)}{P(\gamma_i = A_i|U)}\right)$ are called the agreement weight, AW_i, when $A_i = 1$ and the disagreement weight, DW_i, when $A_i = 0$. Thus,

$$log_2\big(R(\gamma)\big) = log_2\left(\frac{N_M}{N_U}\right) + \sum_i \{AW_i|DW_i\} \qquad \text{(Eq. 7)}$$

This is to say that the ratio of matches to non-matches in a set of pairs is a function of the linear sum of agreement and disagreement weights for each of the identifiers with an offset of $log_2\left(\frac{N_M}{N_U}\right)$.

2.4 Exposition of Strategy

Now, we are proposing a strategy where for a given identifier i, rather than categorizing it as simply an agreement or a disagreement we have multiple agreement levels, $A_{i,0}, A_{i,1}, A_{i,2}, \dots, A_{i,J}$ such that

$$A_{i,0}\colon \Psi < \psi_0$$

$$A_{i,1}\colon \psi_0 \le \Psi < \psi_1$$

$$A_{i,2}\colon\ \psi_1 \le \Psi < \psi_2$$

$$\dots$$

$$A_{i,J}\colon\ \psi_{j-1} \le \Psi < \psi_J,$$

and $\psi_0 < \psi_1 < \psi_2 < \cdots < \psi_J$,

where Ψ is the value of string comparator and $\psi_0,\ \psi_1, \psi_2, \dots, \psi_J$ are successively greater cutoff values.

Then, using Bayes Theorem, we can compute probabilities (m_j) as

$$m_0 = P\left(A_{i,0}|M\right) = P(\Psi < \psi_0|M)$$

$$m_1 = P\left(A_{i,1}|M\right) = P(\Psi \ge \psi_0|M) \cdot P(\Psi <\ \psi_1|\Psi \ge \psi_0 \cap M)$$

$$m_2 = P\left(A_{i,2}|M\right) = P(\Psi \ge \psi_0|M) \cdot P(\Psi \ge\ \psi_1|\Psi \ge \psi_0 \cap M) \cdot P(\Psi <\ \psi_2|\Psi \ge \psi_1 \cap M)$$

$$\dots$$

$$m_{J-1} = P\left(A_{i,j-1}|M\right) = P(\Psi \ge \psi_0|M) \cdot P(\Psi \ge\ \psi_1|\Psi \ge \psi_0 \cap M) \cdot \dots \cdot P\left(\Psi \ge\ \psi_{J-1}\middle|\Psi \ge \psi_{J-2} \cap M\right) \cdot P(\Psi <\ \psi_J|\Psi \ge \psi_{J-1} \cap M)$$

$$m_J = P\left(A_{i,j}|M\right) = P(\Psi \ge \psi_0|M) \cdot P(\Psi \ge\ \psi_1|\Psi \ge \psi_0 \cap M) \cdot \dots \cdot P\left(\Psi \ge\ \psi_J\middle|\Psi \ge \psi_{J-1} \cap M\right),$$

and we can compute the probabilities (u_j) as

$$u_0 = P\left(A_{i,0}|U\right) = P(\Psi < \psi_0|U)$$

$$u_1 = P\left(A_{i,1}|U\right) = P(\Psi \ge \psi_0|U) \cdot P(\Psi <\ \psi_1|\Psi \ge \psi_0 \cap U)$$

$$u_2 = P\left(A_{i,2}|U\right) = P(\Psi \ge \psi_0|U) \cdot P(\Psi \ge\ \psi_1|\Psi \ge \psi_0 \cap U) \cdot P(\Psi <\ \psi_2|\Psi \ge \psi_1 \cap U)$$

$$\dots$$

$$u_{J-1} = P\left(A_{i,j-1}|U\right) = P(\Psi \ge \psi_0|U) \cdot P(\Psi \ge\ \psi_1|\Psi \ge \psi_0 \cap U) \cdot \dots \cdot P\left(\Psi \ge\ \psi_{J-1}\middle|\Psi \ge \psi_{J-2} \cap U\right) \cdot P(\Psi <\ \psi_J|\Psi \ge \psi_{J-1} \cap U)$$

$$u_J = P\left(A_{i,j}|U\right) = P(\Psi \geq \psi_0|U) \cdot P(\Psi \geq \ \psi_1|\Psi \geq \psi_0 \cap U) \cdot \ldots \cdot$$
$$P\left(\Psi \geq \ \psi_J\middle|\Psi \geq \psi_{J-1} \cap U\right).$$

Substituting the appropriate value of $\frac{P\left(A_{i,j}|M\right)}{P\left(A_{i,j}|U\right)}$ (i.e., depending on the value of the string comparator) into Eq. 5 yields the following set of rules for computing the comparison weight (CW) for a string comparison:

$$\Psi < \psi_0: \qquad CW = log_2\left(\frac{m_0}{u_0}\right) = DW_1$$

$$\psi_0 \leq \Psi < \psi_1: \qquad CW = log_2\left(\frac{m_1}{u_1}\right) = AW_1 + DW_2$$

$$\psi_1 \leq \Psi < \psi_2: \qquad CW = log_2\left(\frac{m_2}{u_2}\right) = AW_1 + AW_2 + DW_3$$

$$\psi_2 \leq \Psi < \psi_3: \qquad CW = log_2\left(\frac{m_3}{u_3}\right) = AW_1 + AW_2 + AW_3 + DW_4$$

…

$$\psi_{J-1} \leq \Psi: \qquad CW = log_2\left(\frac{m_J}{u_J}\right) = AW_1 + AW_2 + \cdots + AW_J$$

In effect, we have extended the Fellegi-Sunter probability model to accommodate multiple agreement levels. Note that since the overall probabilities still remain as products of conditional probabilities, the pair weights developed to rank probability ratios are still computed as sums of log-transformed probabilities ratios, called agreement and disagreement weights.

3. Application of Method to Linkage

The pairs under analysis are developed within a set of overlapping blocking passes. Within each blocking pass, among records having the same values for a blocking key, (i.e., a set of identification variables), every record from file A (i.e., NHCS) is paired with all records from file B (i.e., NDI). Included among the developed pairs are those for which the file A record and the file B record both have a non-missing value for a unique ID field (i.e., SSN) that can be compared. Since this ID field is reported with high fidelity, for a certain subset of the pairs in each blocking pass (which we will call the truth set), it is possible to infer match status with high accuracy. These pairs are then used to estimate m and u probabilities that are in turn used to score all pairs (both within and outside of the truth set) within the blocking pass.

For our analysis, m and u probabilities are estimated specific to each blocking pass from the truth set developed within it. Calculating u probabilities for first and last names, for each name seen in file A, we compute the proportion of truth set pairs with disagreeing SSN that have Jaro-Winkler similarity scores above the threshold levels of 0.85, 0.90, 0.95, and 1.00. For example, for the first name (FN) “Susan”, we would count the number of pairs (determined based on non-agreeing SSN to be non-matches) for which the similarity score is $\Psi \geq 0.85,\ \Psi \geq 0.90,\ \Psi \geq 0.95$, and $\Psi \geq 1.00$. Then

$$u_{.85}(FN = \text{‘}Susan\text{’}) = \mathrm{P}(\Psi \geq 0.85 \mid FN_A = \text{‘}Susan\text{’} \ \cap U)$$

$$u_{.90}(FN = \text{‘}Susan\text{’}) = \mathrm{P}(\Psi \geq 0.90 \mid FN_A = \text{‘}Susan\text{’} \cap U \cap \Psi \geq 0.85)$$

$$u_{.95}(FN = \text{‘}Susan\text{’}) = \mathrm{P}(\Psi \geq 0.95 \mid FN_A = \text{‘}Susan\text{’} \cap U \cap \Psi \geq 0.90)$$

$$u_{1.00}(FN = \text{‘}Susan\text{’}) = \mathrm{P}(\Psi = 1.00 \mid FN_A = \text{‘}Susan\text{’} \cap U \cap \Psi \geq 0.95).$$

Note that only $u_{.85}$ is not conditioned on Ψ being greater or equal to a lower threshold.

Because of the sparseness of truth set, we did not seek to compute m probabilities specific to each name. Instead we computed them only specific to each blocking pass (combining records with all values for name):

$$m_{.85}(All\ Names) = \mathrm{P}(\Psi \geq 0.85 \mid M)$$

$$m_{.90}(All\ Names) = \mathrm{P}(\Psi \geq 0.90 \mid M \cap \Psi \geq 0.85)$$

$$m_{.95}(All\ Names) = \mathrm{P}(\Psi \geq 0.95 \mid M \cap \Psi \geq 0.90)$$
$$m_{1.00}(All\ Names) = \mathrm{P}(\Psi = 1.00 \mid M \cap \Psi \geq 0.95).$$

That is, of all the pairs in the blocking pass which were in the truth set, we computed the proportion meeting the same cutoff levels used for the u probabilities. The m probabilities will be much closer to 1 than corresponding u probabilities since they represent name similarity among pairs of records each representing the *same* person.

Next, we computed agreement and disagreement weights according to each Jaro-Winkler similarity cutoff. For the baseline level, which in this analysis is set to 0.85, we computed (in accordance with Fellegi-Sunter model) the agreement weight, $AW_{.85} = log_2\left(\frac{m_{.85}}{u_{.85}}\right)$ and the disagreement weight as $DW_{.85} = log_2\left(\frac{1-m_{.85}}{1-u_{.85}}\right)$. For each higher level of Jaro-Winkler similarity threshold, j, we compute $m_j = P\big(\Psi \geq \psi_j \big| \Psi \geq \psi_{j-1} \cap M\big)$, $u_j = P\big(\Psi \geq \psi_j \big| \Psi \geq \psi_{j-1} \cap U\big)$ and $AW_j = log_2\left(\frac{m_j}{u_j}\right)$ and $DW_j = log_2\left(\frac{1-m_j}{1-u_j}\right)$: i.e., these probabilities are condition on $\Psi \geq \psi_{j-1}$.

Then, for any string comparison we use the following logic.

1. Comparison weight is initially set to 0.
2. Start at a baseline similarity score cutoff level of ψ_0 =.85
3. If the Jaro-Winkler similarity score is less that than cutoff level, add the disagreement weight (which has a negative value) for that level to the comparison weight. The resulting value is the final weight for this comparison.
4. If the similarity score does exceed the cutoff level,
    a. Add the agreement weight for that level to the comparison weight
    b. Proceed to the next higher similarity level (e.g. ψ_0 =.90) and go back to step 3.

4. Analysis of Results

To evaluate the results of the proposed method, we conducted two record linkage analyses of names, one with a fixed (unscaled) string comparator cutoff score and one with scaled cutoff scores. For each of these analyses, the remainder of the linkage configurations were identical.

For the fixed (unscaled) run, the comparator cutoff, $\Psi \geq \psi_0 = .85$ was used. For scoring of first and last names, if the similarity was greater than or equal to 0.85, we assigned the full agreement weight, and if it was less than 0.85, we assigned the full disagreement weight. The agreement weights and disagreement weight for the 0.85 level were computed using u probabilities computed specifically for the value of the name on the NHCS file. The m probabilities were computed using known matches among all pairs generated in the blocking pass.

For the scaled run, the comparator cutoffs were set as $\psi_0(baseline)$ =.85, ψ_1 =.90, ψ_2 =.95, and ψ_3 =1.00. As described in the earlier section, agreement and disagreement weights were computed for each level based on the corresponding m and u probabilities for those levels, and for all levels above the baseline level, they were conditional on agreeing at the next lower level. Just as for the fixed run, the u probabilities were those specifically computed for the name being analyzed, and the m probabilities for each level were computed and applied for all pairs in a given blocking pass.

For both runs, each pair that was scored was assigned an estimated value of $P(Match)$. This is the estimated probability that a given pair is a match, and was calculated as

$$P(\widehat{Match}) = \frac{R(\gamma)}{1+R(\gamma)}, \qquad \text{(Eq. 8)}$$

where $R(\gamma)$ is the ratio of matched to unmatched pairs

From Eq. 8, we have

$$\widehat{R(\gamma)} = 2^{log_2\left(\frac{N_M}{N_U}\right)+\sum_i\{AW_i|DW_i\}} \qquad \text{(Eq. 9)}$$

and so can compute $P(\widehat{Match})$.

The term $\frac{N_M}{N_U}$, the ratio of matched to unmatched pairs, is fixed for each blocking pass and can be estimated by the Expectation Maximization algorithm (Resnick and Asher, 2019). This method cycles between estimating $P(Match)$ for each pair p using an estimated value of N_M in Eq. 9 to estimate $\widehat{R(\gamma)}$ and inserting this value into Eq. 8, then re-estimating N_M as $\sum_p P(\widehat{Match})$ until convergence. We linked all pairs that had a value of $P(\widehat{Match})$ greater than fixed cutoff levels. We performed this analysis over a range of cutoffs for $P(Match)$: 0.85, 0.87, 0.89, 0.91, 0.93, and 0.95 for the scaled and unscaled approach.

The level of Type I Error was estimated using the gold standard method (Resnick and Asher, 2019), with the gold standard being the truth set: all pairs having non-missing values of SSN for both the NHCS record and the NDI record. Among probabilistically determined links, the Type I error rate is the proportion of pairs with valid SSN values for both records for which SSN values disagree. The level of Type II error is computed as the proportion of truth set records that were not linked. Sensitivity and positive predictive value were calculated from Type I and Type II estimated error rates at the different $P(\widehat{Match})$ cutoffs: sensitivity being the proportion of known matches (from the truth set) that were correctly linked and positive predictive value being the proportion of pairs with valid SSN values, where those SSN values agree.

The results of the analysis are presented in Table 1:

Table 1: Unscaled vs. Scaled: Error Comparison

P(Match) Cutoff %	*Sensitivity (Unscaled)*	*Sensitivity (Scaled)*	*Positive Predictive Value (Unscaled)*	*Positive Predictive Value (Scaled)*
	%	**%**	**%**	**%**
85	98.4	98.9	99.2	99.3
87	98.4	98.9	99.3	99.4
89	98.3	98.9	99.3	99.5
91	98.3	98.8	99.5	99.5
93	98.2	98.8	99.6	99.6
95	98.1	98.7	99.7	99.7

We can see from Table 1, that in terms of sensitivity, the scaled method produces better linkage results for capturing the number of true matches. For example, at $P(Match)$ = 0.91 the sensitivity in the unscaled approach is 98.3%, meaning we are missing about 1.7% of the true matches. The sensitivity in the scaled approach is 98.8%, meaning we are missing about 1.2% of the true matches. Thus, at this $P(Match)$ cut off, the scaled approach reduces the loss of true matches by about 30% (1.7% - 1.2% / 1.7%). In terms of positive predictive value, the scaled analysis outperformed the unscaled analysis slightly at lower values of the $P(Match)$cutoffs but was about equal for the higher $P(Match)$ cutoffs of 0.91, 0.93, and 0.95.

5. Conclusion

We have developed a method of scaling string comparison weights to reflect the level of similarity that is consistent with the record linkage paradigm described by Fellegi and Sunter (1969). Unlike other methods, ours does not assume a linear relationship of variable weights to string comparison scores and uses an approach consistent with Fellegi-Sunter theory to compute these weights based on string comparison scores. We have also developed it in a way that enables adjustment of weights to name frequencies. The results based on the 2016 NHCS linked to the 2016/2017 NDI showed that that the scaled weighting approach produced better results compared to using non-scaled string comparison scoring for different link acceptance cutoffs, across various levels of estimated P(Match).

There are some limitations based on our initial assessment. This method was only tested on a single record linkage analysis, and it would be desirable to test it on multiple record linkage analyses. It would also be desirable to compare the results from the scaled approach we present to one relying on linear transformation functions on the comparator scores.

Another limitation is the number of levels of comparator cutoffs that should be used and how to determine the optimal interval spacing; as well as fitting curves to the calculated level weights to have a continuous transformation function from comparator score to weight as suggested by Winkler (11). How best to fit these curves remains to be specified and it is not clear if these curves would produce superior linkage results. Nevertheless, our initial assessment of a scaled approach has suggested promise in making improvements for linking string comparators in entity to entity record linkages.

References

1. Budzinsky, C. D. "Automated spelling correction." Statistics Canada (1991).
2. Cohen, William W., Pradeep Ravikumar, and Stephen E. Fienberg. "A Comparison of String Distance Metrics for Name-Matching Tasks." IIWeb. Vol. 2003. 2003.
3. Christen, Peter. "A comparison of personal name matching: Techniques and practical issues." Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06). IEEE, 2006.
4. Chen, Hao. "String Metrics and Word Similarity applied to Information Retrieval." University of Eastern Finland (2012). https://epublications.uef.fi/pub/urn_nbn_fi_uef-20120382/urn_nbn_fi_uef-20120382.pdf
5. Fellegi, Ivan P., and Alan B. Sunter. "A theory for record linkage." Journal of the American Statistical Association 64.328 (1969): 1183-1210.
6. Jaro, Matthew A. "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida." Journal of the American Statistical Association 84.406 (1989): 414-420.
7. Jokinen, Petteri, Jorma Tarhio, and Esko Ukkonen. "A comparison of approximate string matching algorithms." Software: Practice and Experience 26.12 (1996): 1439-1458.
8. National Center for Health Statistics. Division of Analysis and Epidemiology. The Linkage of the 2016 National Hospital Care Survey to the 2016/2017 National Death Index: Methodology Overview and Analytic Considerations, August 2019. Hyattsville, Maryland.
9. Porter, Edward H., and William E. Winkler. "Approximate string comparison and its effect on an advanced record linkage system." Advanced record linkage system. US Bureau of the Census, Research Report. 1997.
10. Resnick, Dean, and Jana Asher (2019). "Measurement of Type I and Type II Record Linkage Error." ASA Proceedings of the Joint Statistical Meetings, American Statistical Association (Alexandria, VA).
11. Winkler, William E. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." U.S Bureau of the Census. 1990. https://www.researchgate.net/publication/243772975_///String_Comparator_Metrics_and_Enhanced_Decision_Rules_in_the_Fellegi-Sunter_Model_of_Record_Linkage