The Utility of Imputed Matched Sets

Analyzing Probabilistically Linked Databases in a Low Information Setting

A. M. Thomas; L. J. Cook; J. M. Dean; L. M. Olson

The Intermountain Injury Control Research Center, University of Utah School of Medicine, Department of Pediatrics, Division of Critical Care, Salt Lake City, Utah, USA

Keywords
Probabilistic linkage, imputed, high probability, matched sets

Summary
Objective: To compare results from high probability matched sets versus imputed matched sets across differing levels of linkage information.

Methods: A series of linkages with varying amounts of available information were performed on two simulated datasets derived from multiyear motor vehicle crash (MVC) and hospital databases, where true matches were known. Distributions of high probability and imputed matched sets were compared against the true match population for occupant age, MVC county, and MVC hour. Regression models were fit to simulated log hospital charges and hospitalization status between the two methods. High probability and imputed matched sets were significantly different from occupant age and MVC county (p < 0.002), but imputed matched sets were not (p > 0.493). High information settings saw no significant differences in inference of simulated log hospital charges and hospitalization status between the two methods.

Results: High probability and imputed matched sets were not significantly different from occupant age, MVC county, and MVC hour in high information settings (p > 0.999). In low information settings, high probability matched sets were significantly different from occupant age and MVC county (p < 0.002), but imputed matched sets were not (p > 0.493). High information settings saw no significant differences in inference of simulated log hospital charges and hospitalization status between the two methods. High probability and imputed matched sets were significantly different from the outcomes in low information settings; however, imputed matched sets were more robust.

Conclusions: The level of information available to a linkage is an important consideration. High probability matched sets are suitable for high to moderate information settings and for situations involving case-specific analysis. Conversely, imputed matched sets are preferable for low information settings when conducting population-based analyses.

1. Introduction

As large databases have become more available, computers more powerful, and probabilistic linkage software more prevalent, studies using probabilistic linkage have become more widespread. Probabilistic linkage is a method for combining information from different databases into a single dataset for analysis [1–3]. Desired information about study subjects is often contained in two or more databases [1, 4, 5]. If a unique key, such as social security number or personal health number, does not exist between databases, it is not possible to combine the information from each database directly. Rather than relying on a unique key, probabilistic linkage utilizes fields that are common to each database [1–6]. Comparisons of multiple fields lead to the determination of the probability that two records refer to the same person and event and should therefore be linked. High probabilities assigned to pairs can be achieved by using specific and accurate common fields [2]. Algorithms exist for determining the amount of information contained in linkage fields and how much information is needed to achieve a desired match probability [1–6].

Probabilistic linkage has been used for studying many health outcomes and processes, including tracking patients after hospital discharge, updating disease registries, and studying patient outcomes following traumatic events [7–11]. Databases used for probabilistic linkage are usually administrative in nature and are collected for purposes other than linkage [3, 4, 12–14]. As a result, the quality or completeness of these databases is variable. If linkage fields are frequently missing or erroneous, then the linkage may miss true matches. If missing or erroneous data are related to some mechanism, such as injury severity, biases may unintentionally be introduced into the linked dataset [15].

Probabilistic linkages with high probability matched sets consist of focusing analysis on pairs of records that achieved a match probability higher than a predefined cutoff, commonly 0.90 [6, 16, 17]. A threshold can be calculated to determine if the majority of true matches will achieve match probabilities above the predefined cutoff [6]. This threshold, referred to as the minimum potential match probability, is the theoretical lowest match probability a pair of records can receive using a set of linkage variables. High minimum potential match probabilities indicate the set of linkage variables as a whole contains high in-
formation while low minimum potential match probabilities indicate the set of linkage variables as a whole contains low information. If the minimum potential match probability is greater than the predefined cutoff, then the majority of true matches will have match probabilities greater than the predefined cutoff. Otherwise, there is a risk that true matches will not achieve the predefined cutoff and the linkage should not be performed.

Linkage variable information is high when variables contain many levels, such as name and birth date. Low information linkage variables have fewer, broader levels that make it difficult to distinguish people or events from one another [2, 4, 18]. Rare values also convey more information than common values. In a motor vehicle crash (MVC) database, matching on an age of 21 years provides less information than matching on a rare age of 89 years. High information linkage variables may be unavailable in administrative databases, making it difficult to achieve minimum potential match probabilities that are greater than the predefined cutoff required by high probability linkages.

Statistical imputation has been used to augment databases with missing values in survey research, and studies of MVC and trauma [19–22]. By considering true match status as a missing value to be imputed, a method for extending multiple imputation techniques to probabilistic linkage can be applied [23]. Rather than fixing a lower bound that a pair of records must achieve to be considered a true match, this method takes several weighted samples of all possible matches. Analyses are run on each sample and the results are combined across samples [24]. Multiple imputation methods have been used to fill in missing data post-linkage in other studies [25, 26]. While post-linkage imputation allows researchers to account for missing data in the analysis fields, it fails to address situations where the missing data are in the linkage fields themselves, which has the potential to influence linkage results.

In an effort to eliminate false matches, many researchers go to great lengths to ensure that their linkage models produce high quality matches; however, the influence of excluding true matches from the resulting linkage is rarely addressed. While high probability matched sets guarantee that the resulting matches will most likely be correct, a large number of true matches may fail to achieve the predefined cutoff and will be missed by the linkage. It is necessary to understand the consequences of using high probability matched sets when the minimum potential match probability is low.

2. Objective

The goal of this study is to compare results obtained from high probability matched sets versus imputed matched sets across differing levels of minimum potential match probabilities.

3. Methods

We performed a series of linkages and analyses on two simulated datasets in which we could identify true matches. The University of Utah Institutional Review Board approved this study.

3.1 Data Source

We used the Utah motor vehicle crash (MVC) database obtained from the Utah Department of Transportation, Division of Traffic and Safety. This database contains information on all reported MVCs in Utah. A MVC is reportable if it occurs on public roadways and results in at least one injury or fatality or at least $1,500 in property damage. The data are collected by the responding police officer at the scene of the MVC and include identifying information on persons involved (i.e., name, birth date, sex) as well as MVC details that pertain to them individually (i.e., injury status, seating position, restraint use); detailed information about the time, location, and type of MVC; and vehicles involved.

From the Utah MVC database, we created two simulated datasets (referred to as File A and File B) with known true matches between the datasets. This was accomplished by first selecting 10,000 records from the MVC database and placing these records in both File A and File B. We selected an additional 180,000 records from the MVC database that were different from the first selection and inserted 90,000 records in File A and the remaining 90,000 records in File B. The result was two simulated datasets, each with 100,000 records, 10,000 of which should match exactly to a single record and 90,000 that should not match to any records in the other dataset.

Probabilistic linkage is often conducted to combine information between MVC and hospital databases to better understand the medical outcomes of occupants [9, 11, 18, 27–29]. To replicate a typical study of linked MVC and hospital databases, we generated two hospital outcomes for File B: simulated log hospital charges and hospitalization status. Cases that linked from File A to File B were considered hospitalized cases; otherwise, cases were considered non-hospitalized. Simulated log hospital charges were obtained from a linear regression model with errors normally distributed. The values of the parameters and variance of the error term used to simulate log hospital charges were derived from a linear regression model of observed log hospital charges from previously linked Utah MVC and hospital data. The model is summarized as can be seen in Figure 1.

The following covariates were binary: occupant sex (female vs. male), MVC location (urban vs. rural) and police suspicion of alcohol or drug use (suspected vs. not suspected). Occupant age was a continuous covariate.

3.2 Linkages

To understand the impact of including or excluding low probability matched pairs on analyses, we performed five linkages with File A and File B, varying the amount of available information in each linkage. High information variables, such as name and birth date, may be unavailable for the linkage process for a variety of reasons. Therefore, we used some variable combinations that virtually guaranteed a perfect linkage and other combinations that made use of very little information, making identification of correct matched pairs less certain. We calculated the minimum potential match probability to quantify the quality of
linkage variable combinations using methods outlined elsewhere [6]. The linkage variable combinations and the corresponding minimum potential match probabilities are summarized in Table 1. Minimum potential match probability is defined as the theoretical lowest match probability that a pair of records can receive using a set of linkage variables. High minimum potential match probabilities indicate the set of linkage variables as a whole contains low information while low minimum potential match probabilities indicate the set of linkage variables as a whole contains low information. For example, in the high information linkage setting as in linkage A, if all available seven linkage variables match, the minimum probability of a matched pair would be 0.999. In the low information setting as in linkage E, if all five available variables match, the minimum probability of a matched pair would be 0.027. All linkages were conducted using Strategic Matching LinkSolv version 8.1.9077 [30].

We derived high probability and imputed matched sets from each linkage. For high probability matched sets, we only retained matched pairs that achieved a probability of 0.90 or greater. Imputed matched sets are designed to include high and low probability matched pairs [23]. All matched pairs that achieved a probability of 0.01 or greater were used to generate a distribution of candidate matched pairs. Imputed matched sets were generated by taking probability samples of this distribution weighted according to match probability, i.e. matched pairs with higher match probabilities had a greater chance of being selected for the imputed matched sets and matched pairs with lower match probabilities had less of a chance of being selected [23, 31]. For this analysis, we created five imputed matched sets. Analyses were conducted on each imputed matched set and the resulting statistics were combined using the SAS MIANALYZE procedure [24, 32].

### Table 1 Summary of linkage variable combinations and corresponding minimum potential match probability for each linkage performed

<table>
<thead>
<tr>
<th>Linkages</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum potential match probability</td>
<td>0.999</td>
<td>0.913</td>
<td>0.470</td>
<td>0.189</td>
<td>0.027</td>
</tr>
<tr>
<td>Occupant first name</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupant first initial</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupant last name</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupant last initial</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupant birth date</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupant birth month and day</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupant age</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>MVC date</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>MVC time</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MVC hour</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Occupant sex</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>MVC county</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

### 3.3 Analysis

Distributions of high probability and imputed matched sets were compared against the distribution of correct matched pairs for each of the five linkages across three linkage variables: occupant age, MVC county, and MVC hour. The variables selected for the analysis (age, location, and time) are commonly used to describe the matched population following a probabilistic linkage regardless of the type of health data. Kolmogorov-Smirnov tests were used to test for differences between the known distributions of these variables among the true matches compared to distributions obtained following probabilistic linkage.

Regression models were fit to both a continuous and binary outcome to study differences in statistical inference between high probability and imputed matched sets. We used simulated log hospital charges as the continuous outcome and hospitalization status as the binary outcome. For both outcomes, a model with the same covariates was applied to each linkage for high probability and imputed matched sets. Covariates used to model the outcomes were occupant sex, occupant age, MVC location, and police suspicion of alcohol or other drug use. Coefficients and corresponding 95% confidence intervals from the continuous outcome were compared to the coefficients used to generate simulated log hospital charges (Figure 1). Because hospitalization status was not simulated but derived from linkage status, we fit a model to hospitalization status using the 10,000 true matched pairs as the hospitalized cases. This result would have been achieved if the linkage perfectly identified the 10,000 true matched pairs without including any false matches. The coefficients derived from this model are considered the true value of the parameter when comparing the coefficients and 95% confidence intervals.

\[
simulated \log \text{hospital charges} = 6.2993 + 0.0379 \times \text{occupant sex} + 0.0057 \times \text{occupant age} + 0.0875 \times \text{MVC location} + 0.4795 \times \text{police suspicion of alcohol or drug use} + \varepsilon \sim N(0,0.9527)
\]
confidence intervals generated from a model fit to the results of each of the five linkages.

4. Results

There is wide variation between the total matched pairs identified across linkages A (high information) through E (low information), especially among high probability matched sets. The total matched pairs identified by high probability and imputed matched sets are summarized in Table 2. Nearly 10,000 matched pairs are identified in linkage A for both high probability and imputed matched sets. The number of matched pairs identified by high probability matched sets drops to 5,627 and 601 for linkages D and E, respectively. A drop of the same magnitude is not observed in the imputed matched sets, with 8,197 total matched pairs identified in linkage E.

4.1 Linkage Variables

Figure 2 compares the distribution of high probability and imputed matched sets from linkages A through E with the distribution of correct matched pairs for occupant age, MVC county, and MVC hour. Across the three linkage variables, high probability and imputed matched sets are nearly identical to the distribution of correct matched pairs for linkages A, B, and C (all: p > 0.999). These distributions show that in high information settings, there is no difference between high probability and imputed matched sets. High probability and imputed matched sets from linkage D begin to show deviations from the distribution of correct matched pairs across all linkage variables; however, these deviations are not significant (all: p > 0.1). Under linkage E, the distribution of high probability matched sets is significantly different from the distribution of correct matched pairs for occupant age (p = 0.001) and MVC county (p < 0.001); however, it is not significantly different for MVC hour (p = 0.278). The distributions of imputed matched sets is not significantly different from the distributions of correct matched pairs across these three linkage variables (all: p > 0.493).

The differences between the distribution of high probability matched sets and the distribution of correct matched pairs for occupant age and MVC county, as observed in linkage E, are apparent in Figure 2. Occupants aged 80 to 89 years account for 1.63% (n = 163) of correct matched pairs in the true match population. High probability matched sets estimate the proportion of occupants aged 80 to 89 years at over five times (8.49%, n = 51) the true proportion. An overestimate of occupants aged 80 to 89 years exists in imputed matched sets, but on a smaller scale (1.83%, n = 150). Additionally, nearly 50% (n = 4,753) of all correct matched pairs are from county 35; however, high probability matched sets estimate that only 2.83% (n = 17) of matched pairs are from county 35. The next three most populous counties, 11, 49, and 57, are also underestimated by high probability matched sets, with the least populous counties being overestimated. Imputed matched sets do not show this same pattern and are within less than 2% of the correct matched pairs for each county. High probability matched sets from linkage E overstated the contribution of rare match pairs, as seen by the large percent of matched pairs that have older ages and occur in less populous counties. Imputed matched sets follow the distribution of correct matched pairs more closely.

4.2 Estimating Simulated Log Hospital Charges

We modeled simulated log hospital charges to understand the impact of high probability and imputed matched sets. The coefficient estimates of high probability and imputed matched sets from linkages A, B, and C are not significantly different from the coefficients used to generate simulated log hospital charges, as assessed through 95% confidence intervals. The coefficient for age from imputed matched sets is significantly different from the coefficient used to generate simulated log hospital charges for linkage E. The same is seen with police suspicion of alcohol or other drug use but for linkages D and E. The coefficients from high probability matched sets are not significantly different from the coefficients used to generate simulated log hospital charges for linkages D and E; however, the confidence intervals associated with high probability matched sets in linkage E are nearly twice as wide (>1.6 times) for occupant sex and police suspicion of alcohol or other drug use and over twice as wide (>2.2 times) for occupant age and MVC location compared to the same confidence intervals from imputed matched sets.

4.3 Estimating Hospitalization Status

To understand the impact of high probability and imputed matched sets on an analysis of a binary outcome, we modeled hospitalization status. The coefficients for police suspicion of alcohol or other drug use from high probability matched sets are not significantly different from the true coefficient across all linkages. Similarly, the coefficients for occupant sex, occupant age, and MVC location are not significantly different from the true coefficients across linkages A, B, and C. These three coefficients are significantly different from the true coefficients for linkages D and E. This differ-

Table 2 Total matched pairs identified by each linkage

<table>
<thead>
<tr>
<th>Linkage</th>
<th>A (high information)</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E (low information)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High probability matched sets</td>
<td>9,966</td>
<td>10,022</td>
<td>10,049</td>
<td>5,627</td>
<td>601</td>
</tr>
<tr>
<td>Imputed matched sets</td>
<td>9,997</td>
<td>10,235</td>
<td>10,439</td>
<td>10,078</td>
<td>8,197</td>
</tr>
</tbody>
</table>
Figure 2  Distribution of high probability and imputed matched sets from linkages A through E for the linkage variables occupant age, MVC county, and MVC hour
ence is especially evident for MVC location in linkage E. The true odds ratio for MVC location is 0.84; however, high probability matched sets from linkage E estimate this same odds ratio at 7.04, a difference of 6.20 in the odds ratio. The coefficients from imputed matched sets are not significantly different from the true coefficients across all linkages with one exception: the coefficient for occupant age in linkage E is significantly different from the true coefficient. ▶ Figure 4 contains the coefficients and corresponding 95% confidence intervals of high probability and imputed matched sets from linkages A through E.

5. Discussion

Our study compared results from high probability matched sets versus imputed matched sets across differing levels of minimum potential match probability. We have three main findings. First, when minimum potential match probability is low, the resulting linked dataset may not be representative of the true match population. Second, the accuracy of statistical inference of the resulting linked dataset may be jeopardized by low minimum potential match probability. Finally, the level of information available to a probabilistic linkage is an important consideration before undertaking a linkage.

We found that when minimum potential match probability was lowest (0.027), the distribution of matched pairs from high probability matched sets was significantly different from the distribution of matched pairs from the true match population. High probability matched sets overestimated the percentage of rare values and underestimated the percentage of common values in the linked results. For example, the high probability matched sets in our simulation greatly overestimated the proportions of the linked population related to older persons and MVCs occurring in rural counties. Conversely, more common ages and more urban counties were underestimated. Probabilistic linkage assigns rare values more weight than common values because
rare values convey more information [2–4, 18]. In a low information setting, such as in linkage E, matched pairs with rare values are able to achieve the predefined cutoff probability required by high probability matched sets, while true matches that contain common values are unable to do so, excluding them from the linked dataset [6]. As a result, the linked dataset only identified individuals and events with unusual or rare attributes. A previous study suggested that a linkage should be performed only if the minimum potential match probability was greater than the predefined cutoff required by high probability matched sets.[6] Lack of control over available linkage variables in databases can make achieving this cutoff improbable. It may be unsatisfying or impossible to avoid a linkage in such a situation. Imputed matched sets may be a solution. Even in linkage E, the lowest information setting, the distribution of matched pairs for imputed matched sets did not significantly deviate from the true match population. This suggests that even in low information settings imputed matched sets can still be representative of the true match population. However, it is important to note that the same is not true of high probability matched sets.

We also found that statistical inference of the resulting linked datasets varied due to the amount of information available to the linkage. Models derived from high probability and imputed matched sets were not significantly different from the true models when information was high. Even at a minimum potential match probability of 0.470, as in linkage C, the models were not significantly different from the true models. When minimum potential match probability fell below 0.470, problems with inference emerged for both high probability and imputed matched sets. For example, the small number of matched pairs led to wider confidence intervals from high probability matched sets, which were more than twice as wide as the corresponding intervals computed from imputed matched sets for simulated log hospital charges. This becomes problematic when examining associations and reporting results since wide confidence intervals are less informative and indicate reduced power, making it dif-
difficult to identify significant associations between covariates and outcomes. Confidence intervals from imputed matched sets were narrower than high probability matched sets; however, half of the coefficient estimates were significantly different from the true model, thus modeling simulated log hospital charges incorrectly. For hospitalization status, coefficient estimates from high probability matched sets were almost always significantly different from the true model when minimum potential match probability was low. This is because the resulting linked datasets over represented rare values [2–4, 6, 18]. In one case, an odds ratio derived from high probability matched sets was over eight times higher than the true odds ratio, illustrating that inference from high probability matched sets can be vastly different from the truth. Imputed matched sets were not significantly different from the true coefficients in all but one instance. Caution should be taken when analyzing linked datasets with low minimum potential match probability. Imputed matched sets were more robust than high probability matched sets, especially when the outcome was based only on linkage status; however, both are susceptible to errors in low information settings.

The level of information available to a probabilistic linkage is an important consideration. We found that when the minimum potential match probability was high, both high probability and imputed matched sets performed equally well. Even at a minimum potential match probability of 0.470, high probability and imputed matched sets were not significantly different from the true match population. This suggests that only performing linkages when the minimum potential match probability is greater than the predefined high probability cutoff may be too conservative [6]. Our simulation showed that results from probabilistic linkage are also informative in low information settings, such as linkages D and E. We found that imputed matched sets offset many of the biases introduced from having low minimum potential match probabilities by including more true matched pairs. Unlike high probability matched sets, imputed matched sets were generally not significantly different from the true match population. High probability matched sets tended to over represent rare values and exclude many true matched pairs.

5. Limitations

Our simulated databases were built from MVC data. While we are uncertain if our results generalize to other public health-related databases, our choice of variables to study including age, sex, location, and time of onset are available and used for studies in most health-related datasets increasing the generalizability of our findings. While we did not study the effect of missing or erroneous data on the linkages, missing or inaccurate data in linkage fields leads to lower match probabilities. The results for Linkages D and E show that focusing on high probability matched pairs when overall match probabilities are low can lead to inaccurate conclusions. While lowering the overall distribution of match probabilities would also likely affect the imputed results as well, the methodology for obtaining imputed matched sets provides all pairs the opportunity to be selected as a matched pair, thus reducing the impact of missing and erroneous data. The relationship between different database sizes, the percent of true matched pairs, and the resulting linked dataset is not presented and remains uncertain. To confirm our results hold in more than the example presented in this study, we used the same linkage methodology to analyze linked results from databases as large as 500,000 records and with as few as 1% expected match rates. While there was some variation in the results we found similar patterns between the proportions of matched pairs identified and significance (or lack thereof) between high probability and imputed matched sets regardless of the file sizes. Simulated log hospital charges may not be representative of real log hospital charges; however, the risk of this low because simulated log hospital charges were based on a model derived from actual data.

6. Conclusions

We compared results from high probability matched sets versus imputed matched sets by conducting five linkages on simulated data with varying amounts of information and comparing the resulting distributions and inference between high probability and imputed matched sets and the true match population. Our study shows that probabilistic linkage is a method that should be carefully evaluated in the context of available information. Before beginning a linkage project, it is important to evaluate the amount of available information for the linkage and consider how the information may influence the linkage results. High probability matched sets, which are designed to specifically limit the probability of finding false matches, are suitable for high to moderate information settings and for situations involving case-specific analysis. On the other hand, imputed matched sets, which are more representative of the true match population, are preferable for low information settings when conducting population-based analyses.

Acknowledgments

This work was partially funded by NHTSA cooperative agreement number DTNH22-08-H-00304.

References