HHS Public Access

Author manuscript

Am J Phys Med Rehabil. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

Am J Phys Med Rehabil. 2017 January ; 96(1): 17–24. doi:10.1097/PHM.0000000000000513.

A Probabilistic Matching Approach to Link De-identified Data from a Trauma Registry and a Traumatic Brain Injury Model System Center

M. Kesinger[1], **RG. Kumar**[2,3], **AC. Ritter**[2,3], **JL. Sperry**[1,*], and **AK. Wagner**[2,4,5,*]

[1]Department of Surgery

[2]Department of Physical Medicine and Rehabilitation

[3]Department of Epidemiology

[4]Department of Neuroscience

[5]Safar Center for Resuscitation Research

Abstract

Objective—There is no civilian TBI database that captures patients in all settings of the care-continuum. The linkage of such databases would yield valuable insight into possible care interventions. Thus, the objective of this article is to describe the creation of an algorithm used to link the Traumatic Brain Injury Model Systems (TBIMS) to trauma data in state and national trauma databases.

Design—The TBIMS data from a single center was randomly divided into two sets. One subset was used to generate a probabilistic linking algorithm to link the TBIMS data to the center's trauma registry. The other subset was used to validate the algorithm. Medical record numbers were obtained and used as unique identifiers to measure the quality of the linkage. Novel methods were used to maximize the positive predictive value (PPV).

Results—The algorithm generation subset had 121 patients. It had a sensitivity of 88% and a PPV of 99%. The validation subset consisted of 120 patients, and had a sensitivity of 83% and a PPV of 99%.

Conclusions—The probabilistic linkage algorithm can accurately link TBIMS data across systems of trauma care. Future studies can utilize this database to answer meaningful research questions regarding the long-term impact of acute trauma complex on healthcare utilization and recovery across the care-continuum in TBI populations.

Keywords

Probabilistic Linkage; TBI; Trauma; Rehabilitation; Trauma Registry; TBI Model System

Corresponding Authors. Amy K. Wagner, MD, Associate Professor Physical Medicine & Rehabilitation and Neuroscience, Vice-chair Research and Academic Development, Physical Medicine and Rehabilitation, Associate Director Rehabilitation Research, Safar Center for Resuscitation Research, University of Pittsburgh, 3471 Fifth Avenue, Suite 202, Pittsburgh PA, 15237, wagnerak@upmc.edu, Jason Sperry, MD MPH, Associate Professor of Surgery and Critical Care, Associate Director of Acute Care Surgery Fellowship, Suite F1268 PUH, University of Pittsburgh Medical Center, 200 Lothrop St., Pittsburgh, Pennsylvania 15213, office 412-802-8270, Fax 412-648-6872, sperryjl@upmc.edu.

1. Background

Record linkage is a powerful tool in the field of public health.[1,2] Through computational means, two large independent datasets can be combined to increase data sharing and provide opportunities to answer research questions not possible with either single dataset alone. The two forms of record linkage are: 1) deterministic and 2) probabilistic linkage. Crucially, the decision on the type of record linkage to utilize is based on the presence, or absence, of a unique identifier common between the two datasets. In instances where a unique identifier exists between two datasets, like first and last name, or social security number, subjects with exact matches on the linking variables are defined as matches. Using this exact matching criterion of a unique identifier is known as deterministic record linkage. However, in instances where a common unique identifier is not available, it is possible that datasets may still be linked through *probabilistic* means. In this case, common data elements in both datasets can be compared to assess the *likelihood* that two patients are the same, given equal values on a number of variables.

Probabilistic matching has been used historically in a number of settings of public health and health services research. One of the most common applications is the linkage of infant birth records and administrative, public health, and mortality databases.[1, 3–6] With this kind of research, it is possible to address etiologic questions using a “life course epidemiology” approach. Further, in the field of injury epidemiology, there have been several examples of probabilistic linkage, including matching of medical, police, and traffic crash databases.[7–9] Furthermore, as Sayers and colleagues point out, with the recent “big data” movement, there is a push towards undertaking research that links data from multiple sources.[2] Though the method of probabilistic matching is not a novel methodology specifically, its application in the field of Traumatic Brain Injury (TBI) is novel and can provide an immense resource to the field with the ability to study injury and recovery over the continuum of care.

Individuals sustaining a TBI have a continuum of care that begins with pre-hospital emergency medical services and emergency room care, and their care continues with acute hospitalization and treatment from a multidisciplinary team of trauma surgeons, neurosurgeons, and intensivists. Beyond acute care, many individuals require long term rehabilitation with specialists in physical medicine and rehabilitation, along with a multidisciplinary therapy team providing care. Clinical investigations in TBI rarely study questions that bridge this continuum of care, and for many reasons, clinical databases are similarly limited, creating many gaps in knowledge about how early care can influence long-term survival and outcome.

To date, the data available on long-term outcomes in the trauma literature is sparse.[10, 11] Similarly, the rehabilitation literature lacks substantial information from the acute hospitalization or pre-hospital characteristics.[12, 13] With these issues in mind, there is a great public health need for collaboration between the fields of neuro-trauma and neuro-rehabilitation in order to answer clinically important questions regarding the long-term consequences of acute care trauma issues. The ability to link databases across fields that share the same patients can address this need.

The Traumatic Brain Injury Model System (TBIMS) includes a longitudinal database involving data collected at 21 acute rehabilitation centers over the last 25 years, historically funded by the Department of Education.[14] The database includes information on individuals with TBI from multiple sites across the country, and prospectively follows these individuals with severe, moderate, or complicated mild TBI who survive to initial hospital discharge and are subsequently admitted to a TBIMS rehabilitation center. Due to the nature of their injuries, individuals enrolled in the TBIMS are also included in national, state, and local trauma registries. For both databases, individual identifier information is removed, rendering the individuals in each of these databases "de-identified". Thus, we propose that a probabilistic matching algorithm can be a useful and effective tool from which to match data from both databases, creating a large, integrated, multi-center dataset to address relevant research questions linking early patient care to long term outcome for individuals with TBI.

This report includes: (1) an overview of the methods for probabilistic linkage; (2) a novel method used to create a matching algorithm designed to link TBIMS data to trauma registries that feed state and national trauma registries; (3) the results of that algorithm on the linkage of a single TBIMS center to a trauma registry; and (4) a discussion of implications and future directions for large scale study development using a matched dataset.

2. Methods

2.1 Probabilistic Linkage—Weight Generation

Probabilistic linkage is a computationally intensive method that creates "comparisons" between 2 cases, *a* and *b*, across two relatively large datasets, A and B.[15] Using linking variables (data elements common to both datasets), a linkage algorithm assigns estimation weights to each comparison based on the *similarity of the data* shared by cases *a* and *b*. The core concept with probabilistic matching is that, the greater the estimation weight for each comparison of *a* and *b*, the more likely it is that the information for these cases belong to the same person (i.e. a true match).

Though there are subtle differences in probabilistic linkage methods, the technique always uses two main criteria that contribute to the calculation of the matching weights for each comparison: 1) the quality of the data, and 2) the probability of random agreement.[16]

The quality of the data can be described by *m*, that is the probability that the values across a given linking variable, *i*, are the same for a case comparison, given that the cases being compared are actually the same person (i.e. a true match). Or,

$$m_i = \Pr(a_i = b_i | a \equiv b)$$

where *a* and *b* each describes the same person, *i* is the linking variable common across datasets, and a_i and b_i are the values of *i* for cases *a* and *b*, respectively. For example, if *m* = 0.98 for the variable *hospital length of stay*, 98% of true matching cases (i.e. cases in the two datasets that are actually the same person) will have the same value for hospital length of stay (e.g. a_i=10 days and b_i=10 days). Likewise, the complement to this statement is that 2% of the time the value for *hospital length of stay* is different between the two datasets when in

actuality it is the same person (e.g. a_i=9 days and b_i=10 days). Since datasets are most often processed by human operators, data are subject to errors in spelling or coding. The value of *m* accounts for this issue and represents the likelihood that any given variable has inherent error in coding. The specific values of *m* are usually derived by expert judgment or from a reference dataset.[16]

The second fundamental value in probabilistic linkage is the probability that any two cases will randomly have the same value for a given variable, which is defined mathematically as *u*. Unlike *m*, which is more stable across variables, *u* is determined by the frequency distributions of each variable. For example, in a purely representative population, the chance that any two people will randomly share the same sex is 50%; however, the probability that they randomly share the same month of birth is only 8.3% (1/12). The equation for *u* is:

$$u_i = \Pr(a_i = b_i)$$

where *i* is the variable being compared across datasets, and a_i and b_i are the values of variable *i* for subjects *a* and *b*.

Trauma patients are predominantly men, so the probability that two cases are both men will be higher than the probability that two cases are both women. The distributions that determine the values of *u* are chosen from the relatively larger or more comprehensive of the datasets being compared because of the perceived greater variance and variable distribution.[16]

Furthermore, once the values of *u* and *m* are determined for each matching variable, *i*, the weights can be computed by taking the log of the ratio. This ratio varies based on whether or not a given comparison either agrees or disagrees, according to the following formulas:

For comparisons across *i* that agree,

$$w_i = \log(m_i / u_i)$$

Whereas, for comparisons across *i* that differ,

$$w_i = \log(1 - m_i)(1 - u_i)$$

By taking the log, w_i will be positive for comparisons that have the same value and negative for those that have differing values. The total weight (w_t) for any comparison across datasets is:

$$w_t = \sum w_i$$

The more variables that 2 cases across the datasets share, the greater the total weight and the greater the likelihood that the two cases being compared are the same subject. Therefore, more common variables with high quality data across datasets will produce a linkage with a higher sensitivity. The frequencies of w_t can then be plotted on a histogram, which will

theoretically create a bimodal distribution, consisting of a larger distribution of smaller (or more negative) weights and a smaller distribution of larger (or more positive) weights (Supplementary Figure 1). That is, for a given patient in dataset A, there will be many more participants in dataset B with *dissimilar* values for the matching variables, then there are with similar values in dataset B.

Furthermore, a value of w_t is set as the decision point, above which comparisons are considered to be matches and below which comparisons are considered to be non-matches. This value is set at or near the point where the two distributions meet. A relatively large cut-off point will have fewer false positives but will also capture fewer true matches. Likewise, a relatively low cut-off point will capture more true matches but will also have more false positives.

2.2 Probabilistic Linkage—Blocking

When attempting to use a probabilistic matching algorithm for two large datasets, the comparison matrix is a large Cartesian product, which can be thought of as the comparison of values between subject *a* in dataset A, compared to values for subject *a*, *b*, *c*, etc. in dataset B. This series of comparisons is not only extremely computationally intensive, but also is quite inefficient. For example, when using an example with one dataset of 10,000 subjects and another of 500, there would be 5 million comparisons. Not surprisingly, there will be several comparisons between datasets in this scenario that are highly unlikely to be true matches. In order to reduce the number of comparisons and improve efficiency, a method known as ‘blocking’ allows comparisons only among cases that have the same values in a subset of variables. For example, if age and sex are chosen as blocking variables, then only cases that have the same age and sex will be compared across datasets. Blocking can be conceptualized as a funnel or filtering tool before completing the linkage. Ideal blocking variables will have a very high *m*, therefore unlikely to be subject to input error, in order to minimize the exclusion of true matched comparisons.[1,2] Of note, there is no canonical minimum number of variables to use for blocking or matching, but the more high quality variables that can be compared across datasets, the greater will be the sensitivity and specificity of the linked product.

2.3 Probabilistic Linkage—Training Sets

The quality of a linkage can be thought of as a 2×2 table commonly used to describe the quality of a clinical test (Supplementary Table 1). Every comparison has a total weight, w_t. If a comparison has a w_t greater than the cut-off value, that comparison is considered to be a positive test, or match (whether or not the cases compared are actually the same person). Likewise, a w_t less than the cutoff value will be a negative test result—the comparison is considered to be from two different people. As with any clinical test, the quality of a probabilistic linkage algorithm can be measured using sensitivity, specificity, and the positive predictive value (PPV). But most often, these values can only be roughly estimated because the true status of the comparisons is never known.[17]

A training set can be derived and explored when a subset of cases from both datasets shares a unique identifier. Training sets allow for the generation of high-fidelity linking algorithms

that are used to link the remainder of the datasets that do not share a unique identifier. The advantages of a training set are that the values of m_i, the sensitivity, specificity, and positive predictive value (PPV) of the algorithm can be precisely calculated and then used as a precise estimate of quality of the linkage for the portion of the datasets that do not share unique identifiers. Using the most conservative criteria, extra value is placed on high PPV (95% or greater), limiting false positives to 5% or less. The remainder of this report provides an exemplar of probabilistic linkage using data from the trauma registry of a single institution and the TBIMS data from that same institution. The goal of this exemplar is to use a training set from a single institution to create a sensitive and specific algorithm to be used to link TBIMS data from many centers to the trauma databases whose patients feed into these centers for their post-acute rehabilitation care.

2.4 Data

Using medical record numbers (MRN) as unique identifiers, de-identified data from the trauma registry and the TBIMS database were obtained from a single institution that participated in the TBIMS program between 2003 and 2007. The institutional trauma registry recorded data from every admitted trauma patient during the same time period. As in most institutions, the portion of this trauma care information that meets requisite inclusion criteria is subsequently submitted to the Pennsylvania trauma database and the National Trauma Data Bank (NTDB). The trauma registry data were first reduced to include only head injuries as defined by the Barell Matrix, which uses ICD-9 codes to classify injuries.[18] We randomly selected 50 percent of the data to be used to generate a linking algorithm. The remaining data was used to validate the algorithm. The two halves of the dataset were compared to determine if there were significant differences in values and distributions of variables of interest.

2.5 Blocking Variable Selection

The blocking variables selected were age, sex, and year of injury, ensuring that all comparisons between datasets agreed on these three variables. These variables were selected because of likely high specificity and unlikely input error, and these types of variables are commonly used blocking variables.[1,2]

2.6 Linking Variables and Weight Generation

The variables that existed in both datasets and served as linking variables were the following: acute care length of stay, initial systolic blood pressure and respiratory rate in the emergency department, race, initial Glasgow Coma Scale (GCS), presence of alcohol, cranial surgery, cause of injury, intubation status, and head injury pattern (fracture of base of skull or fracture of calvarium). Given that true match status was known through MRNs, we generated m by determining the probability that values agreed across datasets for the true matches. The u value was determined for each match based on frequencies of values in the trauma registry. A cut point was estimated based on the weight distribution of the algorithm generation set. We validated the method by using the weight generation algorithm to predict true matches in the validation half of the dataset. The m values from the algorithm generation set were assigned to the validation set, though true match status was known for both. Linking was conducted using Stata 13th Edition (College Station, TX).

2.7 Clustering

If variables with low specificity are used for blocking in a probabilistic linkage, each case will be compared to many cases in the opposing dataset. We call these groups of comparisons that share the same case from a dataset 'clusters'. For example, if subject A in the TBIMS dataset is compared to subjects 1, 2, and 3 in the trauma registry, then we call these three comparisons a cluster (A-1, A-2, and A-3). In this situation, the comparison with the highest weight within a cluster is assumed to be a match, unless more than 2 comparisons within a cluster share the highest weight.[17] Two similar weights within the same cluster, therefore, may only differ slightly, based on the difference of only a single binary variable. For example, the comparison of subject A from the TBIMS dataset to subjects 1 and 2 of the trauma registry (A-1 vs. A-2), could have very similar weights, by having the same values for every linking variable except one: presence of alcohol at the time of injury. In this case, this single variable would be the deciding factor in whether A-1 or A-2 is considered a match. If we have missing data from either subject 1 or 2 in a dataset it makes it difficult to determine the true match within a cluster. If, however, there are relatively many non-specific linking variables, and the datasets are sufficiently large, the comparisons in the clusters will more likely have similar rather than identical weights, and more ambiguity regarding what is the true match status. For example, subject 2 from the trauma registry may be compared to subjects A, L, D, and W from the TBIMS dataset. This four comparison cluster (A-2, L-2, D-2, and W-2) has its own set of weights. We considered comparisons false matches unless they shared the highest weight in both clusters from which they stemmed. In our example, comparison A-2 would only be considered a match if it had the greatest weight of the cluster that contained subject A (A-1, A-2, and A-3) and the greatest weight of the cluster that contained subject 2 (A-2, L-2, D-2, and W-2).

To further minimize the risk of false positives, we investigated the difference in weights within clusters using MRNs. We subtracted w_t of true matches from w_t of the next highest weight within the cluster, and then compared this value to the similar subtraction from the false matches of the next highest weight. We call this the cluster weight difference (CWD). We chose a CWD value that would yield a conservative algorithm that would allow us to be highly confident that the matches are true matches. In order for a case comparison to be considered a match, we required that the CWD be greater than the chosen value. If the two highest-weighted comparisons within a cluster differed by a factor less than the chosen value, we rejected all comparisons and did not consider there to be any true matches for that case. To follow with the example, if A-1 and A-2 differed only slightly, both comparisons would be rejected and neither considered a match.

Though we had unique identifiers in this study, the goal was to use a training set to create a linkage algorithm that could be used to link TBIMS datasets to the NTDB and state trauma registries, which do not share identifiable patient health information or a unique identifier.

2.8 Validation Set Comparisons

We compared the weight distributions of the algorithm created and of the validation sets in regards to sensitivity, specificity, PPV, and the relation of clusters, weights, and match status.

3. Results

3.1 Initial Datasets

We identified 241 cases in our local TBIMS dataset between 2003 and 2007. One-hundred-twenty-one randomly selected cases were included in the set used to generate the linking algorithm, and the 120 remaining cases in the validation set. There were no significant differences in demographic or injury characteristics between the two. Our local trauma registry had a total of 14,389 cases recorded from this same timeframe. We narrowed this case number to 5,338 by excluding those that did not have a head injury according to the ICD-9 codes of the Barell Matrix.[18] We next blocked on age, sex, and year of injury. Missingness of data for datasets is presented in Supplementary Table 2.

3.2 Algorithm Generation Set

The resulting set used for algorithm generation had 1,281 comparisons consisting of 1,091 cases from the trauma database joined to the 121 cases from the TBIMS dataset, providing a ratio of trauma to rehabilitation cases 9:1 (Table 1). The size of clusters ranged between 1 to 25 comparisons for each rehabilitation case and 1 to 3 for each trauma registry case. The mean (SD) CWD for true matches was 26.28 (15.5) whereas for false matches it was 2.24 (3.14) (Table 2). We used the 90th percentile of the CWD for false matches (5.34) as the threshold margin of error for matching because it was considered advantageous since it was only marginally smaller than the 5th percentile of CWD for true matches (5.55). In the algorithm generation set, 7/121 rehabilitation cases (5.8%) had no comparisons to their true matches in the trauma registry due to differences in blocking variables (N=1) and because there were ICD-9 codes that did not include head injury (N=6). True matches had the highest weight of their respective clusters 109/121 times (90.1%).

The distribution of weights in the algorithm generation set had the characteristic bimodal distribution. (Figure 1A) A value of 5 was chosen as the cut-off based on the distribution. The comparison with the highest weight of each cluster was considered a true positive if the weight was greater than 5; under these conditions, the algorithm had a sensitivity of 0.89 and a PPV of 0.98. (Table 3A) With a CWD of less than 5.34 as another criterion of exclusion, the linking algorithm further improved and had a sensitivity of 0.88 and a PPV of 0.99. (Table 3B) Figure 2A shows the distribution of CWD over weights for true and false matches. As expected, true matches tended to have greater weights. Interestingly, there is a sharp inflection in the distribution curve where the true matches begin. This inflection point of the curve signifies a greater difference between the true match and the next closest comparison within a cluster compared to a false match and its next closest within-cluster neighbor.

3.3 Validation Set

The validation set had 1,347 comparisons consisting of 1,147 cases from the trauma database joined to 120 cases from TBIMS, resulting in a ratio of 9.6:1. (Table 1) Sizes of clusters ranged from 1 to 26 for each rehabilitation case and 1 to 3 for each trauma registry case. The distribution of weights among all cases in the validation set was similar to that of the algorithm generation set (Figure 1B). Further, the weights among true matches were

higher than false matches in both the algorithm generation and validation set (Figure 3A/3B).

The validation set had a similar, but slightly higher, CWD than the training set (Table 2B and Figure 2B). Taking true matches in the validation set to have weights >5, and the greatest weight within a cluster, the sensitivity was 0.87 and PPV was 0.97 (Table 4A). Further excluding comparisons with a CWD not greater than 5.34 resulted in a sensitivity of 0.83 and a PPV of 0.99 in the validation set (Table 4B).

4. Discussion

We generated a probabilistic matching algorithm based on a training set to link patients from a hospital trauma registry with a single TBIMS center. We were able to verify the quality of this algorithm because we used MRNs as unique identifiers across datasets. This yielded a merged dataset that has acute variables not normally present in long-term follow-up databases. Though it is a small dataset, the linked dataset has yielded potentially high impact findings already.[19] With this algorithm, we obtained good sensitivity with a very high PPV. This method can be employed to link patients from other TBIMS centers with de-identified data to their corresponding trauma registries whose de-identified data are fed into the NTDB. We expect a similar sensitivity with centers that have similar levels of missingness in their data. Centers that differ significantly in their patterns of missingness will have lower sensitivity, but we estimate that the PPV will still be high because of the conservative measures taken to generate the weights, which included using low CWD as an exclusion criteria and requiring that any comparison taken to be a true match have the greatest weight in clusters from both datasets.

The greatest value in linking institutional trauma registries to the TBIMS is that within these trauma registries, there is at least as much data as is required to participate in state or national trauma registries. Though institutional trauma registries may not be available to the majority of researchers, this algorithm could be used to link TBIMS data to state and national databases that report the same information as the institutional trauma registries. For instance, our institution contributes to both the Pennsylvania Trauma Outcomes Study and to the NTDB, and the majority of TBIMS centers also have trauma centers that contribute to the NTDB.[20] The data contained in our registry, and the variables used in our linkage algorithm, are the same data that is contributed to these larger trauma databases and therefore the method should be generalizable to those databases. The vast majority of states have registries, though the inclusion/exclusion criteria and the data contained vary. Severe injuries are more likely to be captured than non-severe, and the diagnostic data we used in this study through the abstraction of ICD-9 codes are almost universal in state trauma registries.[21]

The most significant direct implication of this study is the potential for a significant increase in our knowledge and understanding of TBI from the ability to investigate the long-term effects of acute treatments, complications, and injury patterns. Potential further implications include the ability to conduct similar linkage protocols with the Spinal Cord Injury Model Systems and the Burn Model Systems Databases to acute trauma databases.

Normally, probabilistic linkage is conducted on datasets with highly specific variables like date of birth, date of presentation, and even names or initials.[22, 23] Issues of data privacy, especially in the United States, have limited the kinds of questions that researchers can ask, and in doing so, have likely limited the research insights generated from these datasets that would lead to improved outcomes and reduced healthcare costs. When linking on variables that have low specificity, there are likely to be many comparisons with highly positive, identical weights within clusters.[5] However, our weights were generated using relatively many variables, which resulted in no identical positive weights. Nevertheless, comparisons within clusters with positive weights occasionally had only small differences, rendering it difficult to determine the difference between a true positive and a false positive among some cases. In this situation, the difference between a true match and a false positive may be as minor as a difference in recorded race, or intubation status, or the presence of alcohol at time of injury. This issue led us to investigate the characteristics of weights within clusters and to use an additional exclusion criteria of "CWD outside of our margin of error" for matching.

This use of CWD may have wide-ranging implications for probabilistic linkage in that it may broaden the ability to link datasets that only share variables with relatively low specificity and yet still achieve a high PPV. It may also have implications in the ability to link datasets of dramatically different sizes. Further, when graphing CWD over weight, a very large upward inflection in the distribution curve is apparent where true matches begin. This finding may be significant because though true matches will always tend to have greater weights than false positives, it is not usually clear at what point the weights of true matches start and false matches stop just by visually examining the bimodal distribution. In linkages with more specific variables, this difficulty is somewhat mitigated through manual review.[9, 24] However, in many datasets like de-identified trauma registries that have very few if any variables with high specificity, this is impossible. Graphing the CWD values could be used as another piece of information to determine a more accurate cutoff weight, and it could be a valuable accuracy check of the traditional bimodal weight distribution.

Taking true matches to be cases with weights greater than the bimodal intersection, and that were the highest in their given cluster, yielded good sensitivity and PPV. The use of a high CWD as the margin of error for matching criteria added only a slight increase in PPV. If the datasets were larger, and the average number of comparisons per cluster was greater, the CWD exclusion may then produce more dramatic results. This point may also be true if the missingness was higher, but the data from this TBIMS center had relatively little missingness (Supplementary Table 2). The cause of missingness—whether at random or not at random—was not determined. Missingness leads to false negatives. Because the rate of false negatives was low in this study, it is likely that the level of missingness did not play a significant role in our ability to match subject records from each database accurately. Further work is needed to determine the generalizability of the use of CWD with probabilistic linkage.

The use of probabilistic matching approaches to generate research datasets to examine care for individuals with rehabilitation relevant diagnoses, and receiving care across the acute care and rehabilitation continuum, has significant potential to inform and influence the care provided. For example, there is potential beyond the TBIMS national database for applying

this methodology, using matching and blocking variables appropriate to these datasets, is to link trauma data from the NTDB[25] to research information collected through both the Burn and the SCI Model Systems.[26,27] Also, while blocking and matching variables might differ from those presented here, the methodology also might be useful in linking up acute care data generated for activity duty service members to corresponding VA medical records for research questions that span the continuum of medical care among military populations.

Currently, the matched records from this study have been used to generate an initial report about the effects of hospital acquired pneumonia on long term outcomes after moderate-to-severe TBI.[19] Also, the methodology and matching criteria have been used to link data from all TBIMS centers with acute care data from the NTDB, and we generated a database that includes matched records for over 3500 individuals in the TBIMS National Database. We have used this matched dataset to examine the influence of extracranial injury on mental health outcomes,[28] and we will explore how other aspects of the acute trauma complex, comorbidities, as well as acute trauma complications and treatments, influence health care utilization and long term outcomes. The long-term goal of this work in TBI is to generate a greater insight into how heterogeneity with individual factors like comorbid burden, as well as how heterogeneity with the polytrauma complex and acute care practices, can influence recovery trajectories for individuals with moderate-to-severity TBI that receive acute rehabilitation services.

This study is not without limitations. The values of m could be substantially different at different centers due to different registrars and different methods of inputting data. Nonetheless, using a training set from a single center is likely to generate a better estimate than would otherwise be possible. This limitation is potentially mitigated using the CWD distribution from this study. A similar distribution at other centers at which we would apply this algorithm would lend credence to the accuracy of the matching algorithm. Using only patients who had ICD-9 codes indicating TBI in the trauma database reduced the size of the trauma database by 63%. However, it excluded 19 of the TBIMS patients (7.8%) from trauma dataset and made it impossible to find true matches for those cases. Correcting the miscoding the blocking variables, on the other hand, excluded only 1 TBIMS case (0.5%).

5. Conclusions

A training set was used to create an algorithm to probabilistically match the TBIMS data set with a trauma database using many non-specific variables. The algorithm had good sensitivity and a high PPV. We will use the algorithm described here to link the additional TBIMS centers to local, state, and national registries to evaluate relevant research questions about TBI care, healthcare utilization, and outcomes across the recovery continuum.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Disclosures

This work was supported in part by NIDILRR 90DP0041-02-01 (Wagner PI), TL1TR000005, NIH K23GM093032 and NIH 2R25 MH054318.

References

1. Jaro MA. Probabilistic Linkage of large public health data files. Stat Med. 1995; 15(5–7):491–498. 14.
2. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic Record Linkage. Int J Epidemiol. 2015 Dec 20. pii. Dyv322. [Epub ahead of print].
3. Fair M, Cyr M, Allen AC, Wen SW, Guyon G, MacDonald RC. An assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada. Chronic Dis Can. 2000; 21(1):8–13. [PubMed: 10813688]
4. Sauver JLS, Grossardt BR, Yawn BP, Melton LJ, Rocca WA. Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. Am J Epidemiol. 2011; 173(9):1059–1068. [PubMed: 21430193]
5. Bentley JP, Ford JB, Taylor LK, Irvine KA, Roberts CL. Investigating linkage rates among probabilistically linked birth and hospitalization records. Med Res Methodol. 2012; 12:149.
6. Zhang Y, Cohen B, Macaluso M, Zhang Z, Durant T, Nannini A. Probabilistic linkage of assisted reproductive technology information with vital records, Massachusetts 1997–2000. Matern Child Health J. 2012; 16:1703–1708. [PubMed: 21909704]
7. Ferrante AM, Rosman DL, Knuiman MW. The construction of a road injury database. Accid Anal Prev. 1993; 25(6):659–665. [PubMed: 8297434]
8. Clark DE. Development of a statewide trauma registry using multiple linked sources of data. Proc Annu Symp Comput Appl Med Care. 1993; 1993:654–658.
9. Patterson L, Weiss H, Schano P. Combining multiple data bases for outcomes assessment. Am J Med. 1996; 11(1):S73–S77.
10. Sigurdardottir S, Andelic N, Wehling E, Roe C, Anke A, Skandsen T, et al. Neuropsychological functioning in a national cohort of severe traumatic brain injury: demographic and acute injury-related predictors. J Head Trauma Rehabil. 2015; 30:E1–E12.
11. Heltemes KJ, Holbrook TL, Macgregor AJ, Galarneau MR. Blast-related mild traumatic brain injury is associated with a decline in self-rated health amongst US military personnel. Injury. 2012; 43:1990–1995. [PubMed: 21855064]
12. Grauwmeijer E, Heijenbrok-Kal MH, Ribbers GM. Health-related quality of life 3 years after moderate to severe traumatic brain injury: a prospective cohort study. Arch Phys. Med. Rehabil. 2014; 95:1268–1276. [PubMed: 24561059]
13. Hammond FM, Grattan KD, Sasser H, Corrigan JD, Rosenthal M, Bushnik T, et al. Five years after traumatic brain injury: a study of individual outcomes and predictors of change in function. NeuroRehabilitation. 2004; 19:25–35. [PubMed: 14988585]
14. TBI Model System. [Accessed 3/15/2015] Available from: http://www.msktc.org/tbi/model-system-centers.
15. Meray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. J Clin Epidemiol. 2007; 60:883–891. [PubMed: 17689804]
16. Mason CA, Tu S. Data linkage using probabilistic decision rules: a primer. Birth Defects Res A Clin Mol Teratol. 2008; 82:812–821. [PubMed: 18988225]
17. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol. 2002; 31:1246–1252. [PubMed: 12540730]
18. [Accessed 4/12/15] Barell Matrix. Available from: http://www.cdc.gov/nchs/injury/ice/barell_matrix.htm
19. Kesinger MR, Kumar RG, Wagner AK, Puyana JC, Peitzman AP, Billiar TR, et al. Hospital-acquired pneumonia is an independent predictor of poor global outcome in severe traumatic brain injury up to 5 years after discharge. J Trauma Acute Care Surg. 2015; 78:396–402. [PubMed: 25757128]

20. FACS Certified Trauma Centers. [1/18/2015] Available from: http://www.facs.org/accrediated-trauma-centers.

21. Mann NC, Guice K, Cassidy L, Wright D, Koury J. Are Statewide Trauma Registries Comparable? Reaching for a National Trauma Dataset. Acad. Emg. Med. 2006; 13:946–953.

22. Dean JM, Vernon DD, Cook L, Nechodom P, Reading J, Suruda A. Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: a potential tool for evaluation of emergency medical services. Ann Emerg Med. 2001; 37:616–626. [PubMed: 11385330]

23. Herman AA, McCarthy BJ, Bakewell JM, Ward RH, Mueller BA, Maconochie NE, et al. Data linkage methods used in maternally-linked birth and infant death surveillance data sets from the United States (Georgia, Missouri, Utah and Washington State), Israel, Norway, Scotland and Western Australia. Paediatr Pewrinat Epidemiol. 1997; 11(Suppl 1):523.

24. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. AMIA Annu Symp Proc. 2003:259–263.

25. National Trauma Databank. [12-26-15] Available from: https://www.facs.org/quality%20programs/trauma/ntdb.

26. Burn Model Systems. [12-26-15] Available from http://www.msktc.org/burn/model-system-centers.

27. Spinal Cord Injury Model Systems. [12-26-15] Available fromhttp://www.msktc.org/sci/model-system-centers.

28. Kesinger MR, Juengst SB, Bertisch H, Niemeier JP, Kumar RG, Krellman J, Pugh MJ, Sperry J, Arenth PM, Fann J, Wagner AK. Acute trauma Factor Associations with Suicidality across the First 5 Years after Moderate-to-Severe with Traumatic Brain Injury. Phys Med Rehabil. Revision 2016 Arch.

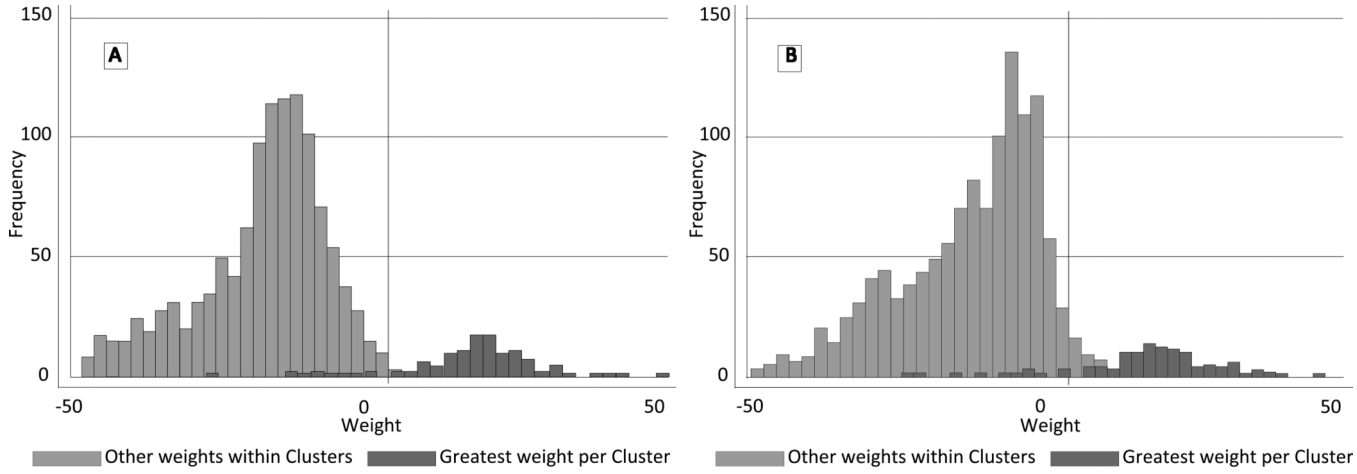

Figure 1.

Frequency distributions of the weights for each comparison assessed. A cluster is defined here as the comparison between one TBI-MS patient with all potential trauma registry patient matches, after blocking for age, sex, and year of injury. The grey bars represent the cases with the highest weight in each cluster. Panel A is the distribution from the algorithm generation subset, and B is the distribution from the validation subset. The vertical line signifies our chosen cutoff value(s).

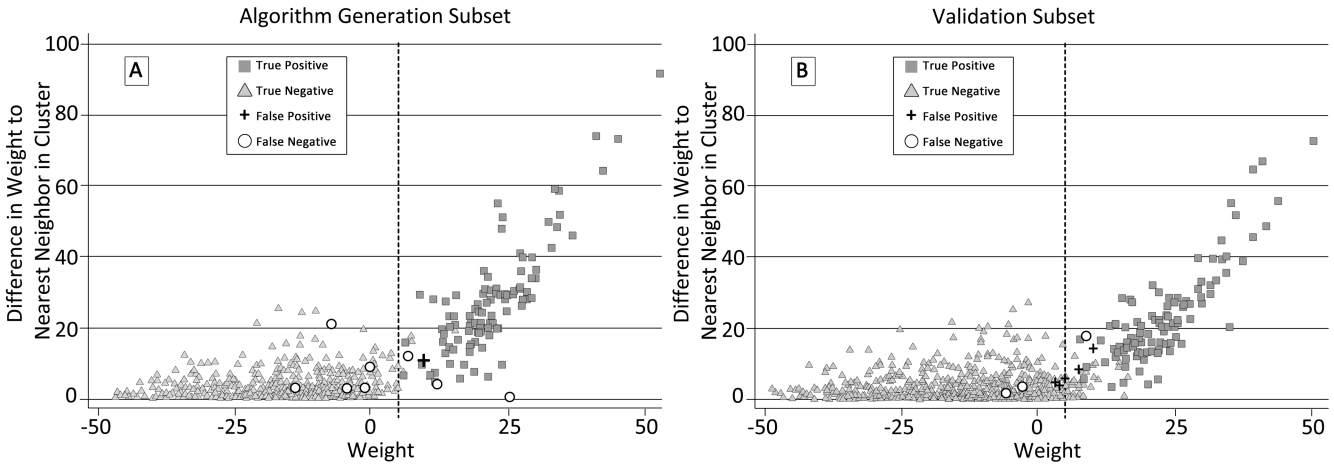

Figure 2.

Cluster weight difference by weights. A is the algorithm generation set and B is the validation set. *Only 8 false negatives appear, though there were 15 false negatives. Discrepancies in ICD-9 codes from the trauma database resulted in 6 false negative cases being excluded, and discrepancies in blocking variables prevented 1 TBI-MS case from being compared to its true match in the trauma database. †Only 7 false negatives appear though there were 20 false negatives. Discrepancies in ICD-9 codes from the trauma database resulted in 13 false negative cases to be excluded.

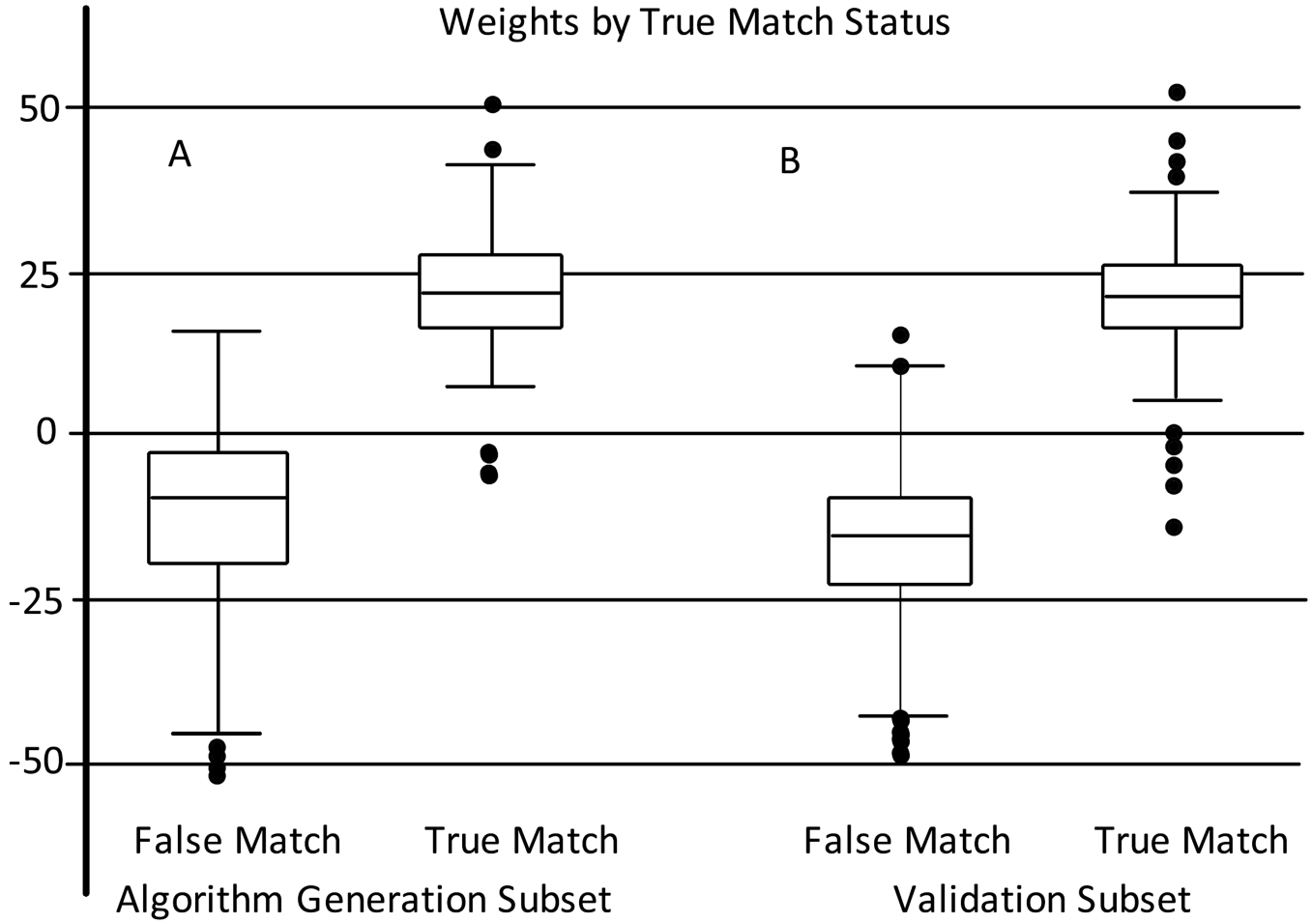

Figure 3.

Box plot of weights for algorithm generation and validation sets based on match status. Panel A includes weights for false and true matches from the algorithm generation subset, and panel B includes weights for false and true matches from the validation subset.

Table 1

Post-blocking descriptive characteristics of weight generation and validation subset

	Trauma	Rehab	Total Comparisons	Mean cases per cluster	# of rehab cases that did not block to true match.	True Match with top weight in cluster (%)
Algorithm generation	1091	121	1281	7.4	7	110 (90.9)
Algorithm Validation	1147	120	1347	7.5	13	106 (88.3)

Table 2

Nearest-neighbor differences for true and false matches in the algorithm generation (A) and validation sets (B)

A. Algorithm Generation Set		
	False Match	True Match
Mean (SD)	2.2 (3.1)	26.3 (15.5)
Median (IQR)	1.1 (0.31—2.83)	24.4 (17.75—30.25)
5th percentile	0.045	5.55
90th percentile	5.3	47.9
B. Validation Set		
Mean (SD)	2.5 (3.5)	22.8 (13.5)
Median (Range)	1.20 (0.42—2.98)	20.39 (14.29—27.86)
5th percentile	0.066	4.71
90th percentile	6.3	39.7

Table 3

Algorithm Generation Set

A) Cases with highest weight in cluster greater than 5

B) Adding as an exclusion criteria a CWD > 5.34

Link Status A	True Match Status		Total
	True	False	
Link	108	2	110
Non-Link	13	1,158	1,171
	121	1,160	1,281

Link Status B	True Match Status		Total
	True	False	
Link	106	1	107
Non-Link	15	1,159	1,174
	121	1,160	1,281

True match: Two cases are the same person in both datasets and is identified as the same person through the linkage

False match: Two cases are different people in the datasets, but they are falsely identified as the same person through the linkage

Table 4

Validation Subset:
A) Cases with highest weight in cluster greater than 5
B) Adding as an exclusion criteria a CWD > 5.34

Link Status A	True Match Status		Total
	True	False	
Link	104	3	107
Non-Link	16	1,224	1,240
	120	1,227	1,347

Link Status B	True Match Status		Total
	True	False	
Link	100	2	102
Non-Link	20	1,225	1,245
	120	1,227	1,347

True match: Two cases are the same person in both datasets and is identified as the same person through the linkage
False match Two cases are different people in the datasets, but they are falsely identified as the same person through the linkage