Published in final edited form as:
Am J Phys Med Rehabil. 2018 April ; 97(4): 236–241. doi:10.1097/PHM.0000000000000838.

Probabilistic Matching of Deidentified Data from a Trauma Registry and a Traumatic Brain Injury Model System Center: A Follow-up Validation Study

RG Kumar[*,1,2], **Z Wang**[*,1,2], **MR Kesinger**[5], **MA Newman**[3], **TT. Huynh**, **JP. Niemeier**[3], **JL Sperry**[5], and **AK Wagner**[1,6,7,8]

[1]University of Pittsburgh Department of Physical Medicine and Rehabilitation, Pittsburgh, PA

[2]University of Pittsburgh Department of Epidemiology, Pittsburgh, PA

[3]Department of Physical Medicine and Rehabilitation, Carolinas Rehabilitation, Charlotte, NC

[4]Acute Care Trauma Surgery, Carolinas Healthcare, Charlotte, NC

[5]University of Pittsburgh Department of Trauma, Pittsburgh, PA

[6]University of Pittsburgh Center for Neuroscience, Pittsburgh, PA

[7]University of Pittsburgh Safar Center for Resuscitation Research, Pittsburgh, PA

[8]University of Pittsburgh Department of Neuroscience, Pittsburgh, PA

Abstract

In a previous study, individuals from a single Traumatic Brain Injury Model Systems (TBIMS) and trauma center were matched using a novel probabilistic matching algorithm. The TBIMS is a multicenter prospective cohort study containing >14,000 participants with TBI, following them from inpatient rehabilitation to the community over the remainder of their lifetime. The National Trauma Databank (NTDB) is the largest aggregation of trauma data in the United States, including over 6 million records. Linking these two databases offers a broad range of opportunities to explore research questions not otherwise possible. Our objective was to refine and validate the previous protocol at another independent center. An algorithm generation and validation dataset were created, and potential matches were blocked by age, sex, and year of injury; total probabilistic weight was calculated based on 12 common data fields. Validity metrics were calculated using a minimum probabilistic weight of 3. The positive predictive value was 98.2% and 97.4% and sensitivity was 74.1% and 76.3%, in the algorithm generation and validation set, respectively. These metrics were similar to the previous study. Future work will apply the refined probabilistic matching algorithm to the TBIMS and NTDB to generate a merged dataset for clinical TBI research utilization.

Corresponding Author: Amy K. Wagner, MD, University of Pittsburgh, Department of Physical Medicine & Rehabilitation, 3471 Fifth Ave, Suite 202, Pittsburgh, PA, USA 15213, T: 412-648-6666, F: 412-692-4354, wagnerak@upmc.edu.

[*]: co-first authors

INTRODUCTION

Database linkage is a powerful statistical methodology that can be leveraged to answer important questions in the field of medicine that are not possible in either dataset alone. In instances where unique identifiers (e.g. medical record numbers) are available, deterministic linkage offers a quick and efficient way to link records between databases. However, many publicly available datasets are de-identified for privacy reasons, making record linkage a more computationally challenging endeavor. Probabilistic linkage, which relies on the matching values of common data elements between databases, can be implemented in such instances without the need for unique identifiers.[1]

Clinical care of patients with moderate to severe TBI occurs along a continuum, beginning with emergency room care and acute inpatient hospitalization at a trauma center. After discharge from the acute hospital, many patients also require comprehensive inpatient rehabilitative services prior to integration into the community. Most of the research conducted to date in TBI has been divided, either: 1) exploring the effect of acute care trauma factors on hospital-based outcomes, or 2) examining long-term recovery in the chronic stages of TBI, beginning during inpatient rehabilitation and extending months to years after TBI. Very few research studies have bridged these two avenues of research to explore the long-term effects of acute care trauma factors, largely because of a lack of available data across these two fields to address these cross-disciplinary types of research questions. Linking the NTDB and TBIMS offers the unique opportunity of simultaneously access both the largest acute trauma care database in the world (NTDB) and the largest longitudinal TBI outcomes national database, the TBIMS, which follows patients for the entirety of life post-injury.

In a previous report, we developed a novel probabilistic matching algorithm at a single medical center to link two databases, the Traumatic Brain Injury Model Systems (TBIMS) single site to trauma registry data records submitted to the National Trauma Databank (NTDB).[2] A parallel deterministic linkage was possible due to available medical record numbers, allowing for us to derive a true match status. Thus, validity metrics were calculated based on concordance/discordance between linked matches from the probabilistic matching algorithm and true match status from the deterministic linkage. Correspondingly, an increased emphasis was placed on two specific metrics in the probabilistic matching algorithm: positive predictive value (PPV) and sensitivity. In this context, PPV is defined as the proportion of individuals linked in our probabilistic algorithm between databases that, in reality, are the same individual. Sensitivity is the proportion of individuals that are true matches between the two datasets that are linked using the probabilistic algorithm. In a previous single site study applying our algorithm, we achieved a PPV of 99% in both an algorithm generation and validation subset; and a sensitivity of 88% and 83% in these algorithm generation and validation subsets, respectively.[2] This initial result is important in that it indicates the accuracy and validity of the proposed probabilistic matching algorithm, in which more than 80% of target cases were matched, and almost 99% of matched cases were the same individual.

As a next step of evaluating the veracity of this probabilistic matching protocol, the purpose of the present study is to apply and validate this novel probabilistic algorithm in another, independent single medical center TBIMS dataset and trauma registry records. This validation is technically possible because of the availability of true match status between the two datasets for all patients. Therefore, in this follow-up study, we conducted a parallel deterministic linkage to allow for calculations of algorithm validity metrics. Having the algorithm validated in an independent center will add a greater level of veracity and confidence to the protocol, with a long-term goal of this project to have a refined and validated probabilistic algorithm that can be applied to the TBIMS National Database and NTDB on a national scale.

METHODS AND MATERIALS

Probabilistic linkage

This study was approved by the local institutional review board. Background of the mechanics of the probabilistic linkage method applied to TBIMS and trauma dataset in previous single-site study has been described in great detail elsewhere.[2] Briefly, for each matched pair, agreement for each linking variable was evaluated in the algorithm by assigning a weight for each corresponding variable. The total weight was summed over all matching variables. The higher the total weight, the greater the probability that the matched pair in reality belongs to the same person. When deciding whether or not cases are considered linked between the two datasets, three tiers of criteria were examined and checked for validity metrics, with each increasing tier having more stringent criteria for matching.

To estimate matching weight, we applied two commonly used criteria: the quality of the data and the probability of random agreement. The quality of data metric is described by *m*, or the probability of matched pair agreement on a given linking variable within each value of the variable in the trauma dataset, given the pair is a true match. For example, if 90% of the matched pairs agree on systolic blood pressure (SBP) when SBP is 140 in trauma, then m=0.9. For a matched pair in this example that does not agree on SBP at 140, then m=0.1. The probability of random agreement is defined by *u*, which estimates the probability that a matched pair will randomly have the same value for a given linking variable. *U* is determined by the frequency distribution of each linking variable. For instance, while the probability of a matching pair randomly matching on sex is 50%, the probability of randomly matching on same birthday will be 0.27% (1/365).[1]

Matching blocking

To increase the efficiency of matching, blocking was employed using the variables: age, sex and year of injury. Only individuals in each database with exact value matches for these three variables were included in the probabilistic match. Blocking can be regarded as filter process to remove matching pairs that are highly unlikely to be the true match.[3] This step is crucial in reducing the computational load of the matching procedure. Age, sex and injury year were applied in the previous study[2] and we observed a low likelihood of human data entry error, resulting in a high specificity.

Linking variables and weight estimation

After blocking procedures were complete, the following variables were selected in the probabilistic matching: acute care length of stay, initial Glasgow Coma Scale (GCS) motor, verbal, eye movement, total (sum of the previous three GCS sub-scores), race, respiratory rate and initial systolic blood pressure in the emergency department, head injury pattern (fracture of base of skull or fracture of calvarium), cause of injury and acute care health insurance payer information. When compared with the probabilistic linkage algorithm we used in the previous study[2], we excluded four binary matching variables: intubation status, sedated status and spinal injury status (SCI) in the current study due to: 1) poor data quality (m<0.7), 2) low differentiation between deterministic true and false matches, and 3) very little appreciable improvement in overall sensitivity or PPV. Of note, the binary variables with high m values (>0.7) were included (cranial surgery and skull based fracture) as they were deemed very high quality data to use for the purposes of matching.

Since the true match status was known through medical record numbers, m was calculated from the probability of agreement for true matches. The value of u was estimated from the frequency distribution for each linking variable in the trauma registry, the larger of the two datasets. The weight for each matched pair on each linking variable (w_{ij}) was assigned if the pair **agreed** on the matching variable:

$$W_{\mathrm{ij}}=\log\ (m_{\mathrm{ij}}/u_{\mathrm{ij}})$$

where i was the ith linking variable and j was the jth matching pair. Also, the following weight was assigned if the pair **disagreed** on the matching variable by:

$$W_{\mathrm{ij}}=log\ [(1-m_{\mathrm{ij}})/(1-u_{\mathrm{ij}})]$$

where i was the ith linking variable and j was the jth matching pair. Total weight was the sum of the weight for each matching variable. In probabilistic linkage, there is a characteristic bimodal distribution of weights: one large distribution reflecting weights of comparisons that are primarily disagreeing negative weights (left distribution), and another, smaller distribution, reflecting comparisons that primarily agree and have mostly positive weights (right distribution). Of note, it is common to have some small overlap between the left and right distributions.

Clustering and cluster weight difference

For each case in the TBIMS dataset, multiple cases within the trauma registry are “potential matches” contingent on sharing the same age, sex and injury year as the TBIMS case. This group of “potential matches” is called a cluster. Within each cluster, the matched pair with the highest total weight is regarded as the most probable match. Occasionally, however, the total weights between two independent potential matches can be very similar. For example, a matched pair theoretically could differ with each other by only one or two matching variables. The cluster weight difference (CWD) was introduced as a quantitative measure of this issue. CWD was computed as the difference of the highest total weight to the second highest total weight within each cluster. If CWD was less than the chosen threshold value, all matched pairs within that cluster were rejected because of the difficulty in distinguishing within a certain margin of error which pair is the true match. Similar to our previous probabilistic matching algorithm[2], we applied threshold values for CWD that corresponded to the 90th percentile of CWD for false matching.

For this validation study, validity metrics were calculated and assigned to one of three "tiers", which designate from *more liberal to more stringent criteria* (hereafter refer to Tier I–III) for considering cases to be linked between datasets (see detailed schematic representation in Figure 3).

> ***Tier I***: the greatest weight in each cluster is considered the linked match;
>
> ***Tier II***: met criteria for Tier I, and the total weight value that corresponds to the right tail of the overlapping distribution of weights;
>
> ***Tier III***: met criteria for Tier II, and CWD greater than 90th percentile CWD for false matches

We considered the Tier III criteria to be the most stringent and most conservative criteria because of the added consideration of a margin of error. It is possible that two cases in the trauma database have similarly large weights. That is, there is a strong agreement in values of several matching variables, and in such a case, the CWD is small, making it harder to correctly identify the true match. A scatter plot was generated of weight by CWD, stratified by true and false match status, with lines overlaying the Tier II and III cut points. We expected that individuals meeting both Tier II and III criteria (top right quadrant of scatterplot) will be mostly true positives (Figures 2a/b).

RESULTS

To generate the probabilistic matching algorithm and validate it, a random number was generated from the normal distribution on the interval between 0 and 1 for each subject in the TBIMS set. A threshold of 0.5 was applied to randomly divide the dataset into training and validation set. The final datasets contained 497 and 544 cases in the training and validation set, respectively. After blocking individuals in each database on age, sex and injury year, we obtained 4,428 matched pairs for the training set and 4,743 pairs for validation set.

With 440 TBIMS rehabilitation cases in the matched pair training set, a total of 4,429 comparisons were obtained from a trauma dataset that contained 12,942 trauma cases. Using Tier I criteria, the sensitivity was 82.3% (Table 1). Based on a visual inspection of the frequency distribution in the training set stratified by greatest weight per cluster vs. all other weights in the cluster, the weight threshold was set to 3 (Figure 1a). Using this Tier II criteria threshold, sensitivity was at 74.1% and positive predictive value (PPV) was 98.2% if the highest weight of each cluster was considered a positive match. The 90th percentile of CWD (7.0) for false matches was used as our threshold for CWD in both the training and validation datasets (Supplemental Table S1). Using the added Tier III criteria of CWD of 7.0, sensitivity and PPV were 66.6% and 99.3%, respectively (Table 2).

For the validation set, a total of 485 TBIMS rehabilitation cases and 12,942 trauma cases were used to form a 4,744 matched pair validation set using the same blocking procedure. Using Tier I criteria, sensitivity was 84.1% (Table 1). Applying the same Tier II threshold cutoff for weight of 3 as the training set, sensitivity was 76.3% and PPV was 97.4%. When a further Tier III criteria of CWD greater than 7 subsequently was applied (as derived from the training set), sensitivity and PPV were 70.7% and 98.0%, respectively (Table 2).

For the training and validation dataset, a scatterplot was generated of the total weight by CWD, with a vertical and horizontal line overlaid to depict the Tier II (weight=3) and Tier III (CWD=7) cut points, respectively (Figure 2a/b). The true and false matching status is shown, with the top right quadrant representing individuals meeting both Tier II and III criteria. As expected, a majority of individuals in the top right quadrant are true positives in the training and validation datasets.

To assess for potential selection bias in the demographics of matched vs. unmatched individuals, selected blocking and matching fields were examined by Tier II criteria (Supplemental Table S2). Our data indicated that blocking and matching fields largely did not significantly differ between matched and unmatched cases except for age and LOS in the training set and SBP in the validation, suggesting a low likelihood of selection bias.

DISCUSSION

The aim of the present study was to refine and validate a probabilistic matching algorithm to link data from the TBIMS to the trauma records from a single clinical site that submits data to the NTDB. In this study, we executed a similar probabilistic procedure in an independent health system where true match status is known, allowing for the calculations of algorithm validity metrics. Importantly, any given dataset and patient population in a single site may differ from another single site in another region, the metrics and threshold values are subject to some degree of fluctuation. Specifically, the derivation of the *m* and *u* values are a function of the data quality and frequency distribution of values in a specific dataset. Validation in an independent site thus is a crucial step to refine our novel probabilistic algorithm before full implementation in a scenario where true match status is unknown.

The probabilistic algorithm used in the original study was modified by omitting three binary variables, including intubation status, sedated status and spinal cord injury (SCI) status. In instances where the quality of the data is determined to be poor ($m<0.7$), the probability that the value for a binary variable will match between datasets by chance alone will be increased. Therefore, we made the determination to set the *u* value to 0.5 to correct for uneven distributions on the likelihood of 0 or 1, and base the score of the weight of binary variables on the data quality. For binary variables with a moderate to high data quality (M value is at least greater than 0.7 for both levels), we still retained binary variables in the matching algorithm such as cranial surgery and skull base fracture. To compensate for inflation of U due to uneven distribution, we set any U above 0.5 to 0.5, and thus, made U irrelevant in the total weight computation.

Though binary variables were removed from the prior probabilistic algorithm, it is important to note that no new variables were added into the algorithm. Given the fact that we refined the algorithm, we derived a training and validation set in the present study. In the training set, using tier II criteria of a weight greater than 3, we achieved a PPV of 98.2% and sensitivity of 74.1%; and in the validation set a PPV of 97.4% and sensitivity of 76.3%. These results are roughly in line with the metrics obtained in the prior study.[2] In probabilistic matching having utmost confidence that cases that are claimed to be linked by the algorithm are in fact true matches, the definition of PPV, is the most important validity metric. In lay terms, since we know it is impossible to have two trauma cases matching to the same TBIMS case, if two weights are reasonably close, then it is better to throw out that TBIMS case, then to risk incorrectly choosing the true match. Of note, when applying Tier III criteria (CWD>7) in this study, we noticed a reduction in sensitivity with only small improvements in PPV, which suggests that this criterion may be too stringent, and Tier II may be sufficient for practical applications.

The data quality, m, is also another important consideration when conducting a probabilistic match. We observed that a majority of m values were comparable (within 20% percent difference) between the current study and the prior study[2] (data not shown). In moving forward to a national merge, we plan to use the m values derived from the current study because of its larger sample size relative to the prior study.

Our study has limitations that should be considered. First, our deterministic linkage was based on cases from a limited time period (1999–2012). Availability of validated matching variables in the algorithm can change over time. For example, systolic blood pressure and respiratory rate are no longer collected in TBIMS after 2013. Therefore, a regular reevaluation and adaptation of this algorithm likely will be needed at later points in time. Also, there could be other unmeasured or unidentified variables which may have higher data quality and lower random agreement rate than current matching fields in our matching algorithm. Based on a probabilistic algorithm developed from a single site[2], the present study refined and validated this algorithm to match patients in the TBIMS to the NTDB in another independent single medical center. Due to the availability of true match status for these patients, we could calculate validity metrics to assess the sensitivity and PPV of our algorithm.

Implications of the Project and Future Directions

Our future directions are to apply this refined protocol to the multi-site TBIMS and NTDB using only probabilistic matching. With the advent of the Federal Interagency Traumatic Brain Injury Research (FITBIR) network, there is a push by the United States federal government to share data across the entire TBI research field. The merger of the TBIMS and NTDB adds to his growing movement of data linking, combining these two datasets are of immense interest in answering a wealth of previously unexplored research questions on the relationships between acute care variables and hospital course on long-term outcomes among individuals with TBI. The NTDB contains a wealth of data on the acute hospitalization, including procedure codes, complication codes, and extensive injury information (cerebral and extracerebral injury severity). However, a major limitation of the

NTDB is that there is only follow-up information until hospital discharge, which restricts the scope of research questions that can be answered. In the TBIMS national database, there is a wealth of follow-up information years after the injury, until a patient is deceased, allowing researchers to assess chronic recovery from moderate to severe TBI. The TBIMS national database has only limited data collection for acute variables; therefore, the long-term effects of acute factors, such as procedures and complications, immediately after TBI cannot be assessed fully without full access to trauma care data. For instance, one such application of initial single site trauma-rehab merged dataset[4] was the examination of the long-term effects of hospital-acquired pneumonia on global outcomes after TBI. Examining hospital-acquired pneumonia effects on long-term recovery for thousands of individuals with data captured in the TBIMS national dataset may have immense implications for the field of TBI, as there is still equivocation in clinical care guidelines with respect to the administration of antibiotic prophylaxis for ventilated patients with TBI.

This initial finding serves as an exemplar for the tremendous potential of our merged database to serve as a platform to address previously unanswerable research questions that have the potential to impact clinical care and future research priorities. It is also important to consider that our methods are not confined to TBI alone, and could have a lasting impact on other rehabilitation disciplines. That is, other model systems injury databases, specifically spinal cord injury and burn injury, also may be well suited for probabilistic matching with the NTDB in future studies. Our research group plans to disseminate our probabilistic matching algorithm code linking the TBIMS data records with the NTDB through an open source on www.rehabilomics.pitt.edu. Along with the code, we will include detailed notes and a standard operating procedure (SOP) regarding utilization of the algorithm. Importantly, the code we will present online is highly specific to the TBIMS and NTDB; however, generalization to other datasets is possible, pending adjustment of the algorithm to include the common data elements that are specific to each of the new datasets. This adjustment will likely require some substantial data cleaning to mirror the coding of variables between datasets, which necessitates clear understanding of the nature of the data collected and how it is coded. The anticipated release of the open source code and SOP will be January 2018. Importantly, the merged dataset generated from the TBIMS and the NTDB will provide multiple opportunities for collaborative projects with interested investigators, and we welcome future collaborations and inquiries.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

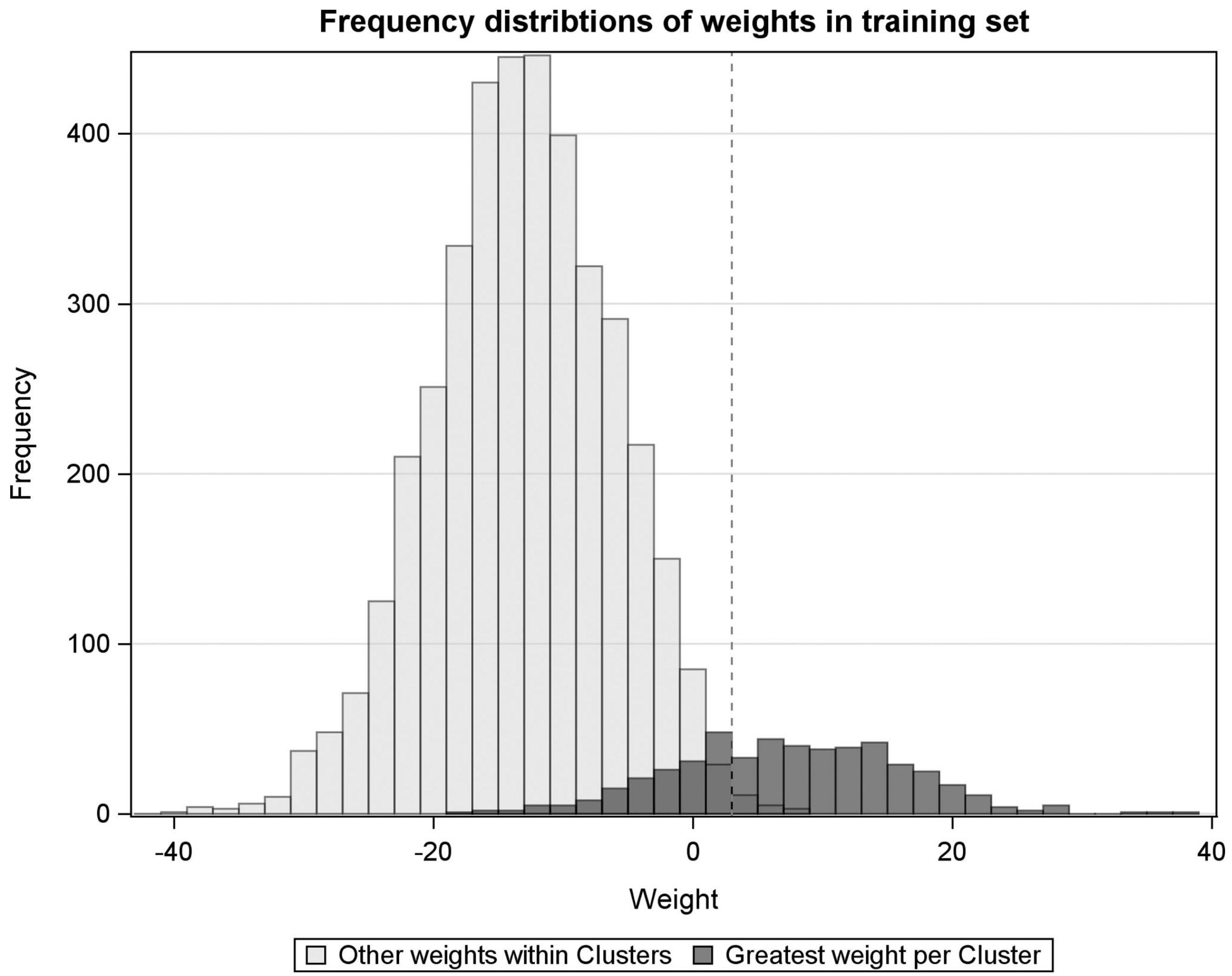
References

1. Mason CA, Tu S. Data linkage using probabilistic decision rules: A primer. Birt Defects Res A Clin Mol Teratol. 2008; 82(11):812–821.

2. Kesinger, MR., Kumar, RG., Ritter, AC., Sperry, JL., Wagner, AK. [Accessed September 28, 2016] Probabilistic Matching Approach to Link Deidentified Data from a Trauma Registry and a Traumatic Brain Injury Model System Center. Am J Phys Med Rehabil Acad Physiatr. 2016. http://europepmc.org/abstract/med/27088479

3. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol. Dec. 2015 doi: 10.1093/ije/dyv322

4. Kesinger MR, Kumar RG, Wagner AK, 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(2):396–402. DOI: 10.1097/TA.0000000000000526 [PubMed: 25757128]


Frequency distribtions of weights in training set
Frequency
0
100
200
300
400
-40
-20
0
20
40
Weight
□ Other weights within Clusters
■ Greatest weight per Cluster

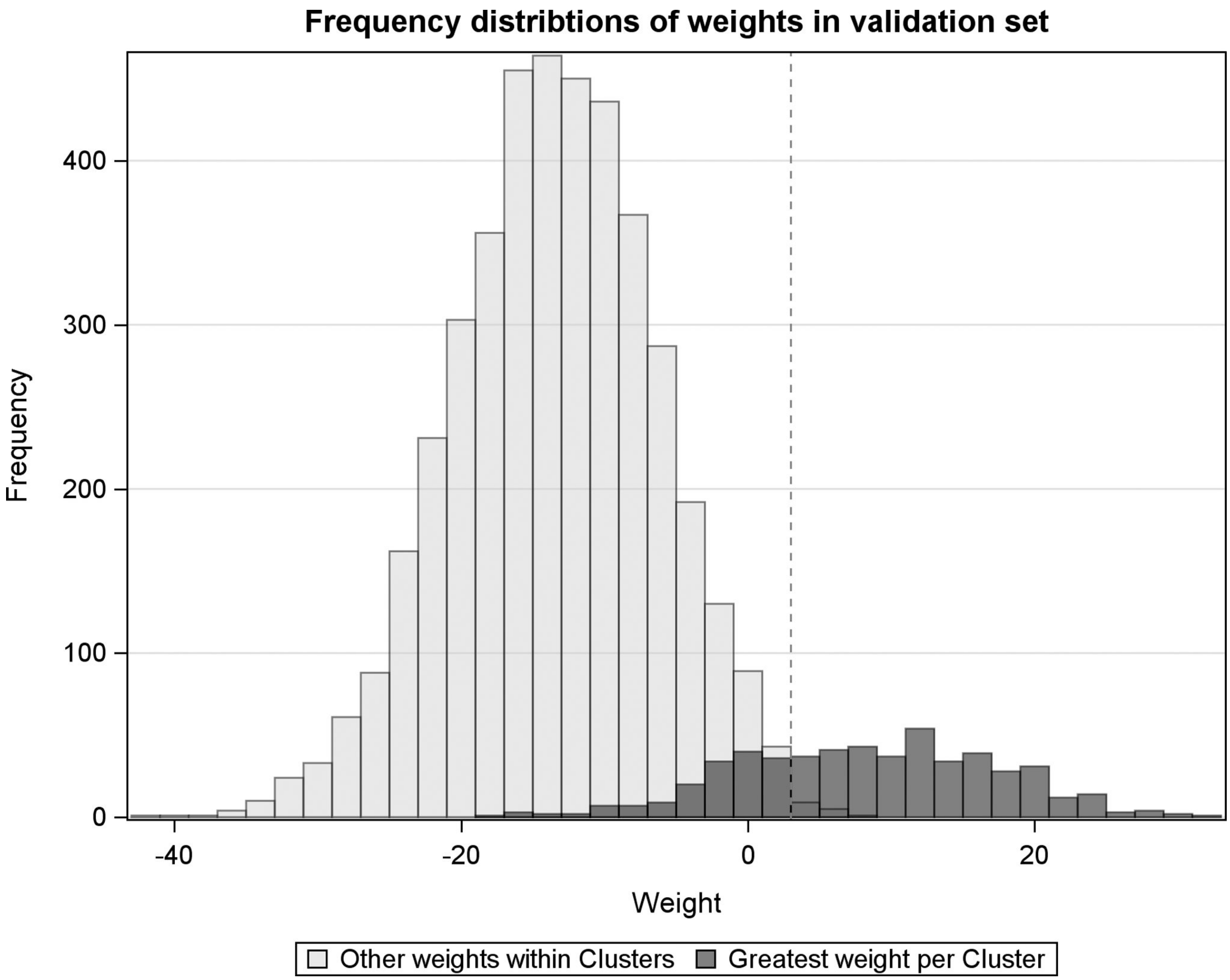
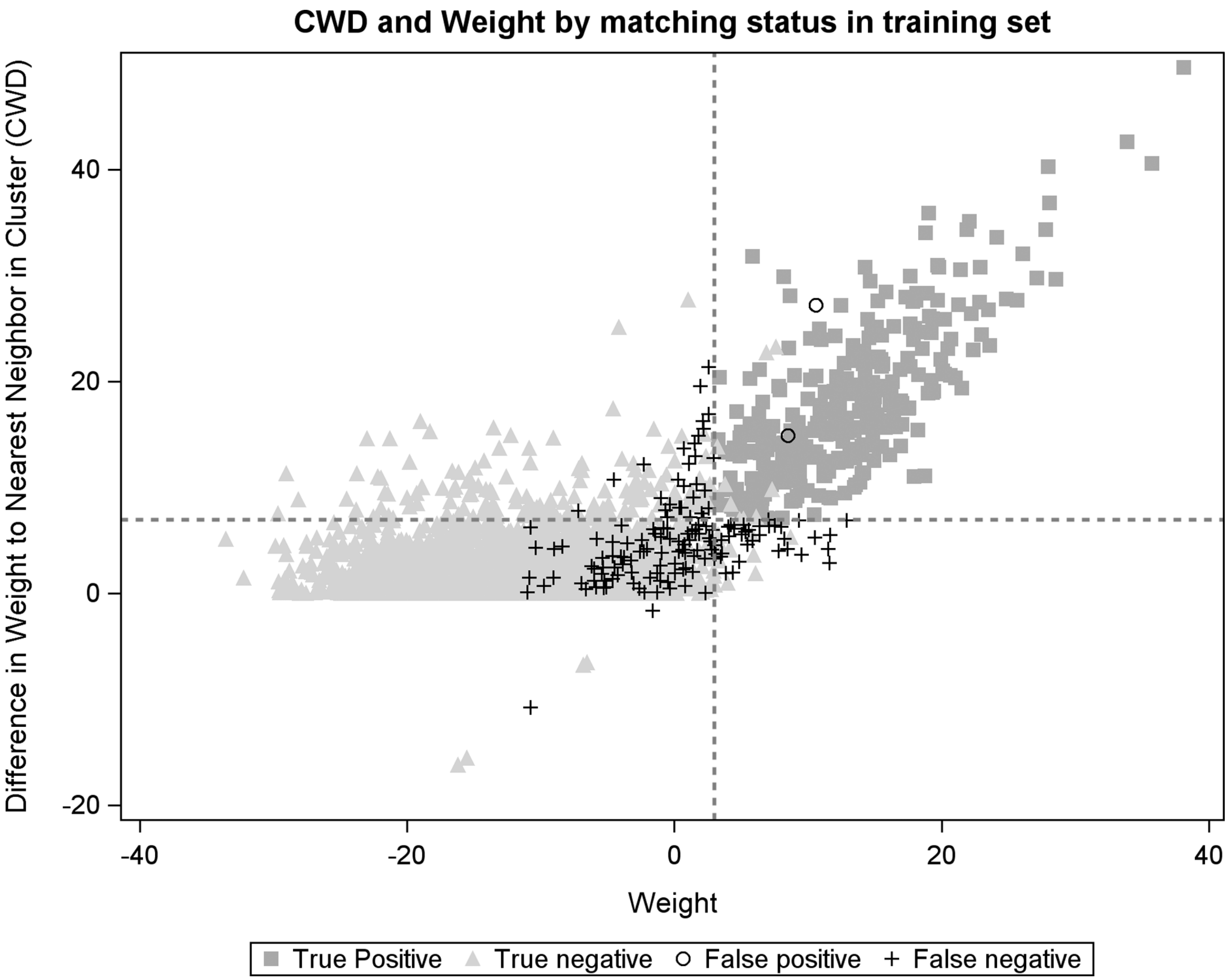


Figure 1.

Panel (A) is the training set, and panel (B) represents the validation set. The frequency distribution of weights among those with the greatest weight in the cluster (dark gray), compared to other weights within cluster (light gray). A cluster is defined as all the trauma cases that are compared to a single TBI-MS case, after blocking for age, sex, and year of injury. The vertical line represents the Tier II criteria of weights greater than 3.


CWD and Weight by matching status in training set
Difference in Weight to Nearest Neighbor in Cluster (CWD)
40
20
0
-20
-40
-20
0
20
40
Weight
■ True Positive
▲ True negative
○ False positive
+ False negative

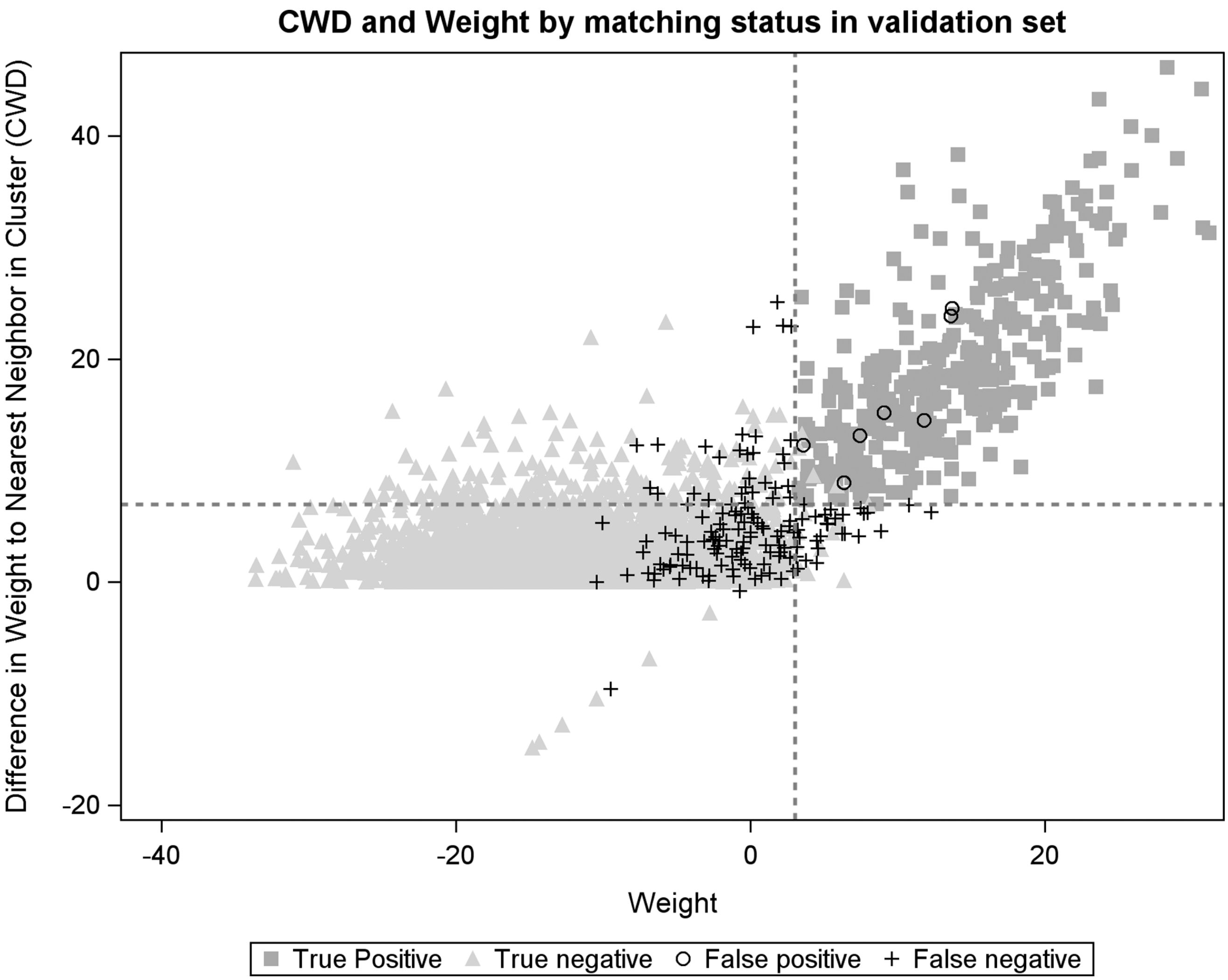


Figure 2.

Panel (A) is the training set, and panel (B) represents the validation set. The case weight difference (CWD) by weight. The vertical line represents the Tier II criteria of weights greater than 3. The four symbols: square, triangle, plus, and circle, correspond to true positive, true negative, false positive, and false negatives when comparing the results of the probabilistic linkage algorithm to the gold standard deterministic linkage.

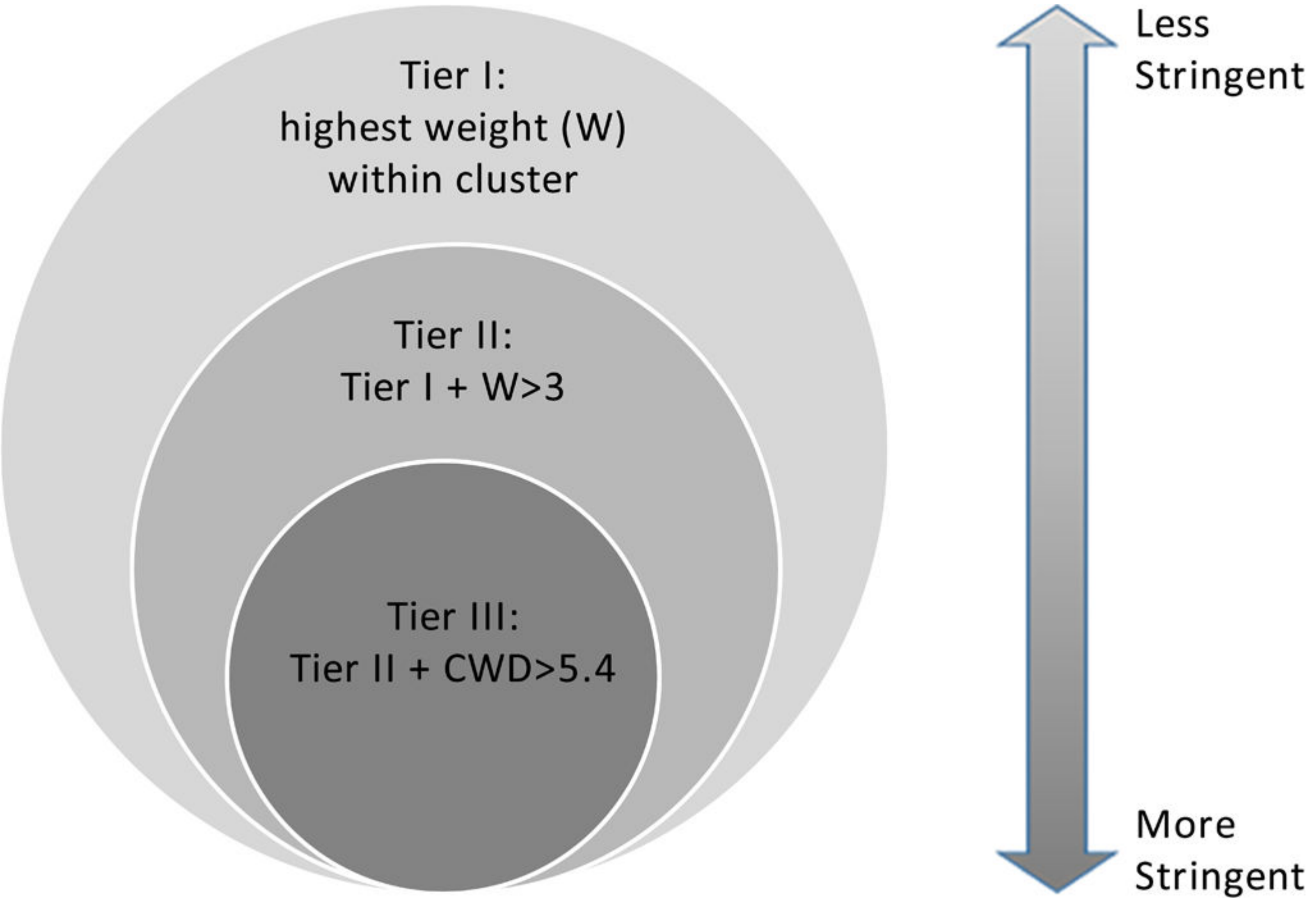


Figure 3.

Schematic representation of determining probabilistic linkage by using progressively stringent criteria (Tier I–III). Validity metrics were calculated in three “tiers”, which designate from *more liberal to more stringent criteria* (hereafter refer to Tier I–III) for considering cases to be linked between datasets. ***Tier I***: the greatest weight in each cluster is considered the linked match; ***Tier II***: met criteria for Tier I, and the total weight value that corresponds the right tail of the overlapping distribution of weights; ***Tier III***: met criteria for Tier II, and cluster weight difference (CWD) greater than 90th percentile CWD for false matches

Table 1

Post-blocking descriptive characteristics of algorithm generation and validation subset

	Trauma	Rehab	Total comparisons	Mean cases per cluster	No. of rehab cases did not block to true match	True match with top weight in cluster (%) (TIER I criteria)
Training set	12,942	440	4,429	8.9	57	408 (82.3)
Validation set	12,942	485	4,744	8.8	59	455 (84.1)

Table 2

True match status by probabilistic linkage status in training set

	Training set			Validation set		
	True match status		Total	True match status		Total
Link status	True	False		True	False	
A						
Link	326	6	332	370	10	380
Nonlink	114	3,983	4,097	115	4,249	4,364
Total	440	3,989	4,429	485	4,259	4,744
Sensitivity	74.1%			76.3%		
PPV	98.2%			97.4%		
B						
Link	293	2	295	343	7	350
Nonlink	147	3,987	4,134	142	4,252	4,394
Total	440	3,989	4,429	485	4,259	4,744
Sensitivity	66.6%			70.7%		
PPV	99.3%			98.0%		

A. Cases with highest weight in cluster greater than 3 (TIER II criteria)

B. Adding as an exclusion criteria a CWD>7 in addition to A (TIER III criteria)