Linked Records of Children with Traumatic Brain Injury*

Probabilistic Linkage without Use of Protected Health Information

T. D. Bennett[1,2]; J. M. Dean[1]; H. T. Keenan[1]; M. H. McGlincy[3]; A. M. Thomas[1]; L. J. Cook[1]

[1]Pediatric Critical Care, University of Utah School of Medicine, Salt Lake City, UT, USA;
[2]Current address: Pediatric Critical Care, University of Colorado School of Medicine, Aurora, CO, USA;
[3]Strategic Matching, Inc., Morrisonville, NY, USA

Keywords
Medical record linkage, pediatrics, brain injuries

Summary

Objective: Record linkage may create powerful datasets with which investigators can conduct comparative effectiveness studies evaluating the impact of tests or interventions on health. All linkages of health care data files to date have used protected health information (PHI) in their linkage variables. A technique to link datasets without using PHI would be advantageous both to preserve privacy and to increase the number of potential linkages.

Methods: We applied probabilistic linkage to records of injured children in the National Trauma Data Bank (NTDB, N = 156,357) and the Pediatric Health Information Systems (PHIS, N = 104,049) databases from 2007 to 2010. 49 match variables without PHI were used, many of them administrative variables and indicators for procedures recorded as International Classification of Diseases, 9th revision, Clinical Modification codes. We validated the accuracy of the linkage using identified data from a single center that submits to both databases.

Results: We accurately linked the PHIS and NTDB records for 69% of children with any injury, and 88% of those with severe traumatic brain injury eligible for a study of intervention effectiveness (positive predictive value of 98%, specificity of 99.99%). Accurate linkage was associated with longer lengths of stay, more severe injuries, and multiple injuries.

Conclusion: In populations with substantial illness or injury severity, accurate record linkage may be possible in the absence of PHI. This methodology may enable linkages and, in turn, comparative effectiveness studies that would be unlikely or impossible otherwise.

Correspondence to:
Tellen D. Bennett, MD, MS
Pediatric Critical Care, University of Colorado School of Medicine
Children's Hospital Colorado
Adult and Child Center for Outcomes Research and Delivery Science (ACCORDS)
13199 E Montview Blvd, Suite 300
Campus Mail F443
Aurora, CO 80045
USA
E-mail: tell.bennett@ucdenver.edu

Methods Inf Med 2015; 54: 328–337
http://dx.doi.org/10.3414/ME14-01-0093
received: September 15, 2014
accepted: March 15, 2015
epub ahead of print: May 29, 2015

* Supplementary material published on our website www.methods-online.com

1. Introduction

Record linkage may create powerful datasets with which investigators can conduct studies evaluating the impact of tests or interventions on health [1]. Linked data may allow studies that might otherwise be very expensive or impossible to conduct.

Databases that contain information about the same persons but do not share a unique identifier can be linked using the information that is common to the two data sources. Record linkage has roots in computer science, statistics, and epidemiology [2, 3], and many advances have come from members of governmental agencies (national census [3], vital records [4, 5], and public health organizations [6, 7]) who needed to integrate very large data sources.

The most useful variables in probabilistic linkage are highly specific, with those that uniquely identify persons (e.g. social security number [SSN]) maximally specific. Variables that have the same value for many people (e.g. gender), while potentially useful for identifying false matches, may require combination with other weakly specific variables to identify true matches. Not surprisingly, health care variables with the most discriminating power are considered protected health information (PHI) under the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [8]. To our knowledge, all linkages of health care data files to date have used PHI. Commonly used linkage variables, all of which are PHI, include names, dates such as birthdate or admission date, zip codes, and SSNs.

Linkage of clinical (e.g. registries and electronic medical records [EMRs]) and

non-clinical (e.g. billing) databases has been identified as an important advancement for patient-oriented outcomes research [9]. Linkage without using PHI would be advantageous because it would increase the number of potential linkages without increasing the risk of disclosing PHI to additional researchers.

The research question that motivated this linkage is an effectiveness study of an intervention for children with severe (Glasgow Coma Scale [GCS] 3–8) traumatic brain injury (TBI). The problem is that no existing U.S. database contains the necessary injury severity and treatment variables to perform such a study. However, two large, overlapping databases each contain a portion of the needed information: the Pediatric Health Information System (PHIS) database and the National Trauma Data Bank (NTDB).

Our group has previously demonstrated that a linkage's success is dependent on the size of the files being linked, the anticipated number of matches, and the discriminating power (or information content) of the variables that are common to the two data sources [10]. Using those three parameters, the feasibility of a proposed linkage at a given accuracy can be estimated. For a hypothetical candidate record pair from two files A and B with |A| and |B| records, respectively, assuming all V common variables agree:

$$1)\ \textit{Prior Odds} = \frac{M}{|A|\ X\ |B| - M}$$

where M is a prior estimate of the number of matched pairs, and

$$2)\ \textit{Likelihood Ratio} = \sum_{i=1}^{i=V} 2^{\textit{Match Weight}(i)}$$

where Match Weight (i) = $\log_2 \frac{m_i}{u_i}$, m_i is the conditional probability of agreement on value x of variable i given that the pair is matched, and u_i is the conditional probability of agreement on value x of variable i given that the pair is unmatched. We estimate the Likelihood Ratio (for feasibility assessment) by making simplifying assumptions: There are pairs with no data errors or omissions; probability distributions are identical for variable i in file A, file B and matched pairs, say $p_i(x)$; and value-specific match weights can be replaced with their weighted average over all values. In this case:

3) *Match Weight* $(i) \approx$ $p_i(x) \log_2 (1/p_i(x)) =$ $H(i)$, *Information Entropy/Content*

4) *Posterior Odds* $\approx$

$$\frac{M}{|A|\ X\ |B| - M} \sum_{i=1}^{i=V} 2^{H(i)} \text{ or}$$

5) *Posterior Odds* = *Prior Odds* × *Likelihood Ratio*

Consequently, any datasets where the three factors combine to produce satisfactory posterior odds (for example, odds 5.6 to 1, equivalent to probability 0.85, or odds 9 to 1, equivalent to probability 0.90) should be linkable by our method. In a similar application, Belin et al. [11] used record linkage to find duplicates in anonymous survey data without using PHI.

Because a limited number of hospitals submit data to both databases (modest file sizes), TBI is not uncommon at pediatric trauma centers (anticipated number of matches), and many variables are common to PHIS and the NTDB (discriminating power), we hypothesized that accurate record linkage of children with severe TBI would be possible without using PHI.

2. Objectives

1) Without using PHI, to link the records of children with trauma in both the NTDB and the PHIS database from 2007–2010.
2) Overall, to create a linked dataset with which to study the effectiveness of intracranial pressure (ICP) monitoring in children with severe TBI.

3. Methods

3.1 Study Design

This study was approved by the university and hospital institutional review boards and written permission was obtained from both the Children's Hospital Association (CHA, PHIS owner) and the American College of Surgeons (ACS, NTDB owner). We defined retrospective cohorts from 2007–2010 in both databases of children admitted to a hospital after trauma and created a standardized dataset from each database. The current data use agreements for the two databases do not allow identified data to be linked by a third party.

We therefore used probabilistic linkage to match patient records in the two datasets. In order to calibrate the linkage parameters and externally validate the linkage results, we compared the linked dataset to the PHIS submission, trauma registry, and EMR data from the same time period at one children's hospital that submits data to both PHIS and the NTDB. Our overall goal was to create a cohort with which to study the effectiveness of ICP monitoring in children with severe TBI.

3.2 Setting

PHIS is a benchmarking and quality improvement database containing inpatient data from 44 U.S. children's hospitals with more than 500,000 discharges per year [12]. PHIS contains rich utilization information, particularly regarding treatments such as medications and nursing interventions, but lacks important clinical variables such as injury severity. PHIS data are only available to approved researchers at member hospitals and do contain limited PHI (complete dates of birth, admission, and discharge, and a hospital identifier). PHIS contains administrative data including demographics, diagnoses, and procedures as well as utilization information for pharmacy, imaging, laboratory, supply, nursing, and therapy services [13]. The data reliability and quality monitoring processes used by the PHIS database have been reported previously [14, 15].

The NTDB contains standardized trauma registry data from more than 3 million admissions at 900 trauma centers in the United States [16]. It contains the injury and clinical variables necessary for studies of TBI, but does not contain detailed treatment information. The NTDB is de-identified and contains no PHI. The de-identified NTDB Research Data Set (RDS) contains all submitted records for a given year. In 2008, for example, the RDS contained

more than 600,000 records of hospital admissions, including more than 100,000 children. The NTDB also has a continuous data quality improvement process [16].

From 2007 to 2010, thirty children's hospitals submitted data to both PHIS and the NTDB.

3.3 Selection of Participants

We selected patients from PHIS and the NTDB who met our inclusion criteria: children < 18 years of age who were discharged from a PHIS hospital in 2007 through 2010 with an International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) discharge diagnosis code for trauma or who were included in the NTDB RDS in 2007 through 2010 (▶Figure 1). These patients represent the trauma cohort. After linking these records, we then selected for children with TBI by searching the NTDB variables for the ICD-9-CM diagnostic codes for TBI used by the Centers for Disease Control (CDC) [17]. In accordance with our overall goal of studying the effectiveness of ICP monitoring, our final cohort includes patients with severe TBI with length of stay (LOS) ≥ 24 hours and non-missing hospital disposition information (▶Figure 1).

3.4 Injury Severity and Mechanism

We calculated injury severity score (ISS) and maximum abbreviated injury scale (AIS) body region scores from ICD-9-CM diagnosis codes using ICDMAP-90 software (Johns Hopkins University and Tri-Analytics, Inc., Baltimore, MD) [18]. The NTDB contains variables for ICDMAP-derived injury scores, but to avoid bias we calculated these scores locally using the same procedures we applied to PHIS data. We categorized injury mechanism using the external cause-of-injury matrix created by the CDC (with ICD-9-CM diagnosis code 995.5 added to the child abuse/assault category) and injury type using the Barell matrix [19, 20].

3.5 Record Linkage

Probabilistic linkage was used to link the records in the de-identified PHIS and NTDB datasets. Probabilistic linkage is a well-established methodology introduced by Newcombe [6, 7] and formalized by Fellegi and Sunter [4]. The uniformly most powerful test of whether any two records are a true match is the match weight.

Match weights can be calculated for each variable and summed to get a total match weight for a candidate record pair if comparison outcomes are independent. In that case, given that the i-th variable agrees between two records, the likelihood ratio (LR) that the records are a true match = $\frac{m_i}{u_i}$, where m_i is the probability that the i-th variable agrees given that the two records refer to the same person/event and is the probability that the i-th variable agrees given that the two records do not refer to the same person/event. Given that the ith variable disagrees between the two records, the LR that the records are a true match = $\frac{1-m_i}{1-u_i}$. A match weight can be calculated for each variable = $\log_2$ (*LR of true match*). The LinkSolv linkage model incorporates more complex match weights allowing for

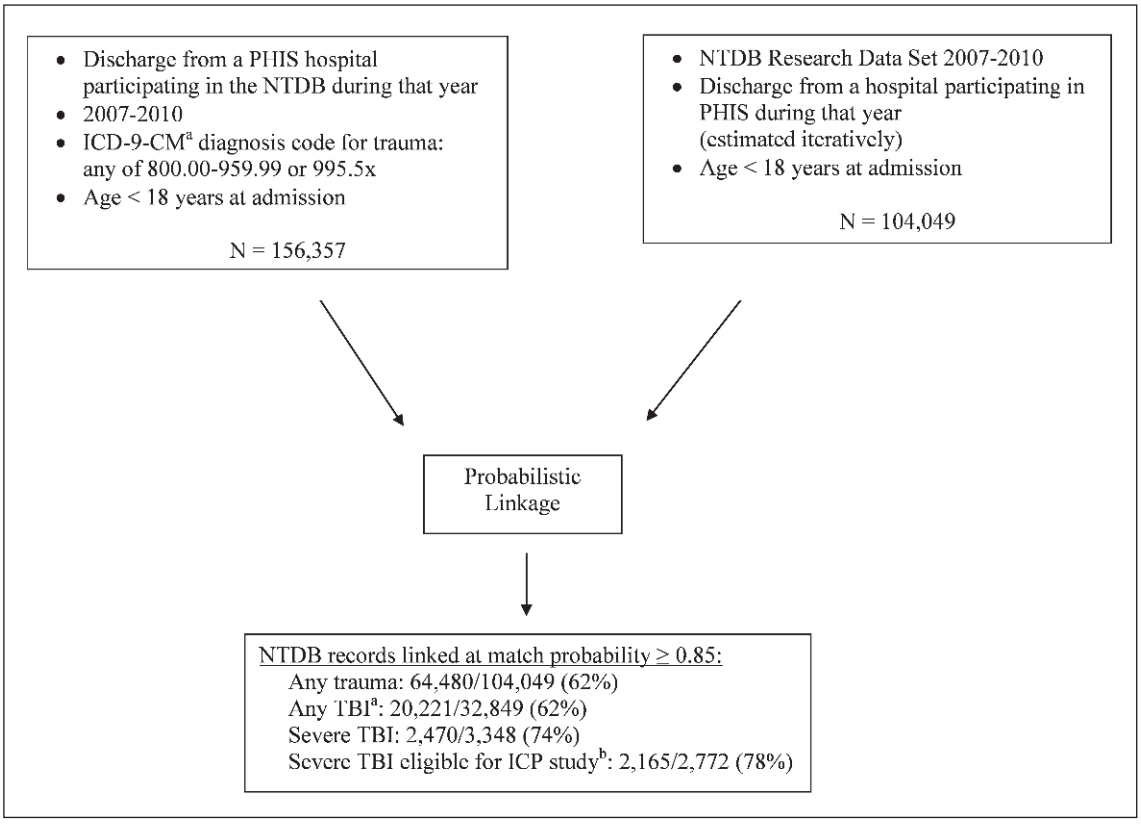

Figure 1
Patient selection method for overall linkage. ICD-9-CM, International Classification of Diseases, 9th revision, Clinical Modification; ICP, intracranial pressure; NTDB, National Trauma Data Bank; PHIS, pediatric health information systems; TBI, traumatic brain injury. [a] 2,261 linked pairs with missing Glasgow Coma Scale excluded. [b] Length of stay > 24 hours and non-missing disposition

agreements on specific values, disagreements, missing values, and dependent comparison outcomes. The probability of a candidate pair being a true match can be calculated from the overall match weight using Bayesian techniques [10, 21]. For convenience, we apply Bayes' rule in terms of odds [22].

To increase the efficiency of the computational matching, only candidate record pairs that agree on "blocking" variables are compared. For example, if age was used as a blocking variable, only the PHIS records of 8-year-old children would be compared to a NTDB record for an 8-year-old child. To account for pairs of records not compared because of disagreement or missing data in a blocking field, different groups of blocking variables are applied in subsequent matching passes.

Probabilistic linkage and information content calculation were performed using LinkSolv (Strategic Matching, Inc., Morrisonville, NY). LinkSolv is the commercial version of the linkage software used by the Crash Outcomes Data Evaluation System (CODES) network funded by the National Highway Traffic Safety Administration [21, 23].

We selected match and blocking variables present in both the NTDB and the PHIS (see Table I in the ▶Online Appendix). In order to select groups of blocking variables, we calculated the information content [5] of each variable in each dataset. Information content (entropy or uncertainty, technically, but the terminology is often substituted) for a variable is a function of the number of different values of the variable, the individual probabilities of each value, and the likelihood of that variable being missing [24, 25]. The units are bits (from $\log_2 x$), where the information content of an evenly distributed binary variable (x is the number of possibilities) without missing data would be $\log_2 2 = 1$ bit. LinkSolv uses the standard formula for entropy [24] to calculate information content $H(x) = -\sum_{i=1}^{n} p_i \log p_i$. It assumes that each variable has a multinomial distribution and estimates population proportions from sample proportions among records with non-missing values. More information content in the match variables would generally make accurate linkage more likely.

We chose blocking variables with good reliability in both databases, little or no missing data, and (when grouped) sufficient information content to make the linkage computationally efficient. LinkSolv runs in Microsoft Access, and as such the linkage database (.mdb file) must remain smaller than 2GB. In the setting of 49 match variables (see Table I in the ▶Online Appendix), blocking variables with low information content would have required evaluation of more candidate pairs than was computationally possible using our system.

Two blocking variables (admission age in years and admission year) were chosen and applied to the first matching pass and four blocking variables (admission age, gender, race/ethnicity, and intensive care unit [ICU] admission) were applied to a second pass. A unique identifier for each of the 30 hospitals in the linkage was estimated iteratively and applied to both passes of subsequent matching runs as a blocking variable.

The 49 match variables were required to match exactly (after cleaning and standardization of both datasets) with the exception of hospital length of stay, ISS, and abdomen AIS score, which were allowed tolerances of +/– 1 day or score unit. We chose these parameters over successive trial match runs by observing match performance when dif-

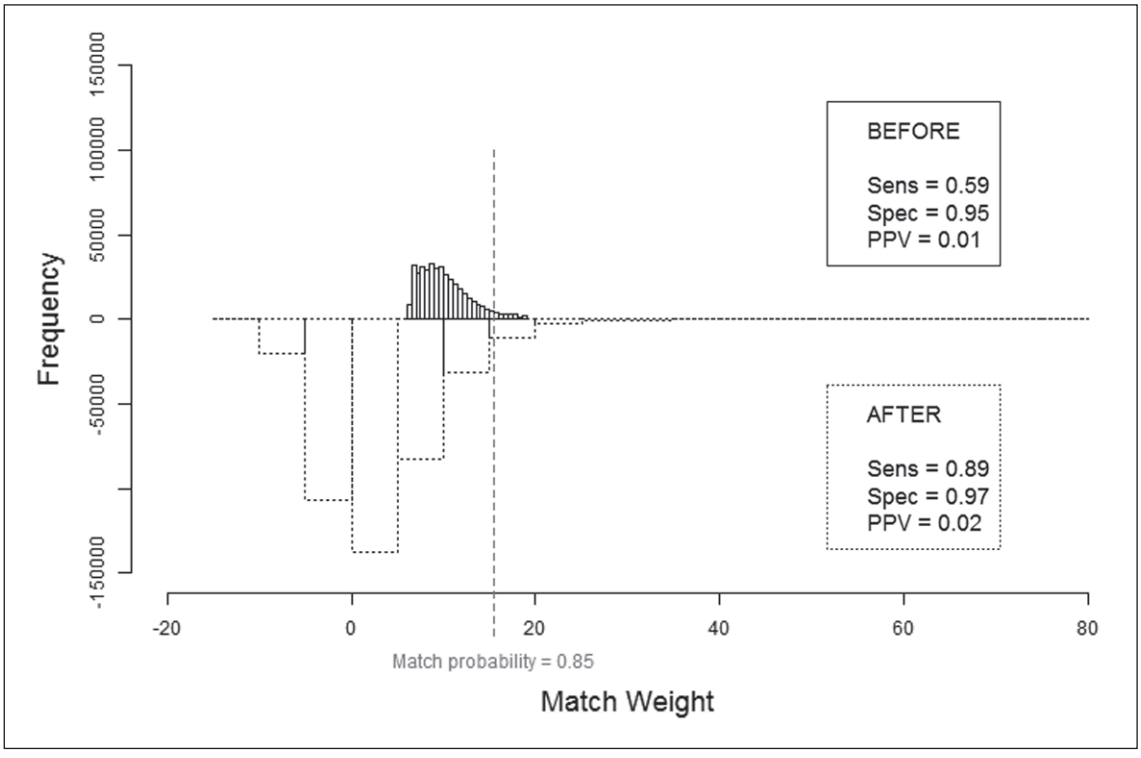

Figure 2
Candidate pairs before and after first MCMC pass, N = 394,756 each. MCMC, Markov chain Monte Carlo; PPV, positive predictive value; Sens, sensitivity; Spec, specificity

Table 1 Characteristics of children with trauma in the PHIS and NTDB datasets

	PHIS N = 156,357 n (col%)	NTDB N = 104,049 n (col%)	X^2 *P*
Age			**< 0.001**
0 to 364 days	20,040 (13)	12,931 (12)	
1 to <5 years	43,522 (28)	25,315 (24)	
5 to <13 years	57,113 (37)	42,594 (41)	
13 to <18 years	35,682 (23)	23,209 (22)	
Gender			
Male	97,314 (62)	66,682 (64)	< 0.001[a]
Missing	62 (0)	17 (0)	
Admission Year			**< 0.001[b]**
2005–2006	309 (0)	0 (0)	
2007	31,406 (20)	18,528 (18)	
2008	40,115 (26)	28,014 (27)	
2009	42,421 (27)	28,149 (27)	
2010	42,106 (27)	29,358 (28)	
Injury Mechanism			**< 0.001[a]**
Fall	42,144 (27)	45,793 (44)	
Assault/Abuse	8,630 (6)	6,876 (7)	
Motor vehicle	14,849 (10)	15,721 (15)	
Other	66,936 (43)	35,534 (34)	
No E-code (Missing)	23,798 (15)	125 (0)	
Injury Severity Score			**< 0.001**
< 15	145,443 (93)	93,140 (90)	
≥ 15	10,914 (7)	10,909 (10)	
Hospital Course			
ICU admission	22,871 (15)	14,953 (14)	0.07
Length of stay, median (IQR)	1 (1–3)	2 (1–3)	< 0.001[c]
Hospital Outcome			
Mortality	1,487 (1)	949 (1)	0.37[a]
Missing	5,439[d] (3)	11,212 (11)	

col%, column percentage; ICU, Intensive Care Unit; IQR, Interquartile range; NTDB, National Trauma Data Bank; PHIS, Pediatric Health Information Systems database
Column percentages may not add to 100% because of rounding.
[a]missing values excluded
[b]*P*-value unchanged if 2005–2006 excluded
[c]Wilcoxon rank-sum test
[d]5,254 of these (97%) are from three hospitals with known missing disposition data

ferent tolerances were allowed. Increasing a tolerance always decreases that variable's discriminating power.

After the two match passes were run, candidate pairs with a calculated match probability below 0.01 were excluded from further analysis, and the remaining candidate pairs were merged into a single table. In order to improve the likelihood of classification of each candidate pair as a true match or true non-match, a Markov chain Monte Carlo (MCMC) parameter estimation step with refinement of each pair's match probability was included. Each pairs' true match status was then imputed five times from that match probability. Several MCMC iterations were run between imputations to ensure that the five imputed sets of matched pairs were independent. LinkSolv readily incorporates the multiple imputation and MCMC steps used in this linkage.

A candidate pair identified as matched during any of the five imputations was considered for inclusion in the matched dataset [21]. In many cases, a given candidate pair matched in all five imputations. To create the linked pairs dataset, we first identified candidate pairs for which the estimated match probability was maximized for both a given NTDB record and a given PHIS record. These pairs were placed in the linked pairs dataset and other candidate pairs containing either that NTDB record or that PHIS record were discarded. From the remaining candidate pairs, we then identified pairs with the highest match probability for a given NTDB record and lower than the highest match probability for a given PHIS record. These pairs were added to the linked pairs dataset.

To determine the minimum match probability we would accept as a true link and to validate the overall linkage, we reviewed all matches from a single center that submits data to both PHIS and the NTDB. Using PHI, we performed a validation linkage that linked PHIS data, the EMR, and trauma registry data for patients in the 2007–2010 linked dataset from Primary Children's Hospital (PCH), an American College of Surgeons (ACS) Level 1 trauma center in Salt Lake City, UT (see Figure I in the ▶Online Appendix).

Most record linkage software requires users to review the distribution of match probability after their linkage and to manually define regions of match probability for linked records, possible links, and non-links [26]. Records in the region of possible links are often directly reviewed (including PHI) to determine their final match status. Because our datasets only included PHI for records from one hospital, we identified the optimal match probability cutoff for a candidate pair in the validation linkage and then applied that cutoff to the overall linkage. We chose the minimum match prob-

ability for a true link to maximize the sensitivity of the linkage process without significantly increasing the false positive rate.

3.6 Statistical Analysis

We used the chi-square test or Fisher's exact test to compare categorical data, as appropriate. Interval variables (e.g. age in years, LOS) were compared using the Wilcoxon rank-sum test. Information content per variable was tested using the Student's t-test.

Data management and validation analyses were performed using STATA™ (StataCorp LP, College Station, TX) and the R environment (version 3.0.2). Statistical significance was defined as $p < 0.05$ for group comparisons.

4. Results

4.1 All Hospitals

We identified 104,049 records in the PHIS database and 156,357 records in the NTDB at the 30 hospitals who submitted data to both databases during 2007–2010 (▶Figure 1). The patients in the NTDB file (median age 7 years, interquartile range [IQR] 3–12) were slightly older than those in the PHIS file (6 years, IQR 2–12, Wilcoxon p <0.001) (▶Table 1). Approximately two-thirds of the patients in both files were male.

Many of the patients were not severely injured: the median ISS score in both datasets was 4 (PHIS IQR 1–5, NTDB IQR 4–9). Patients in the NTDB were more likely to have a defined injury mechanism, more likely to be injured in a motor vehicle crash, and more likely to be severely injured (ISS, ▶Table 1).

The median hospital LOS was 2 days (IQR 1–3) (▶Table 1). ICU admission rates were comparable in the two files, but LOS and in-hospital mortality were greater in the NTDB file.

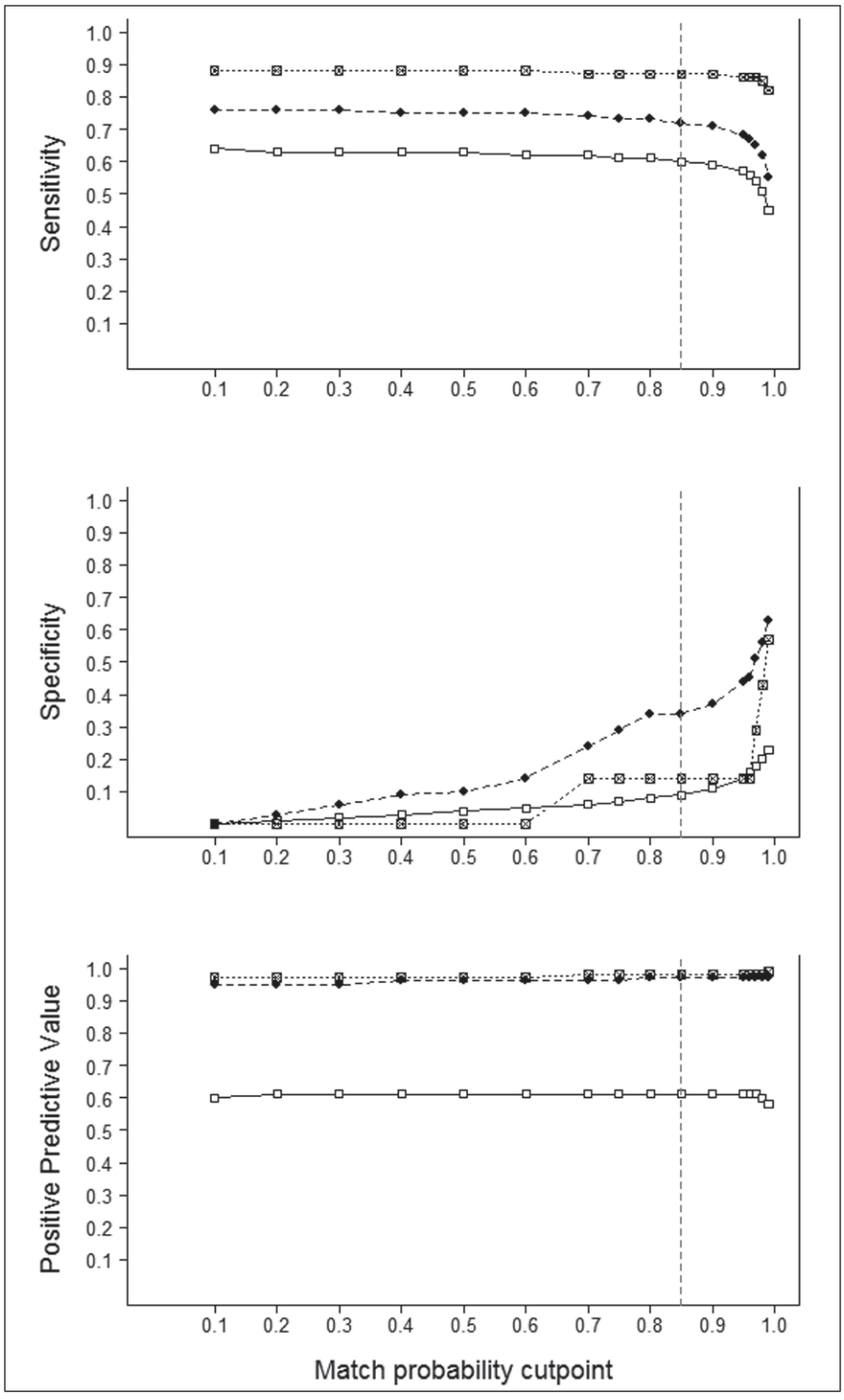

Figure 3 Selection of match probability cutpoint for overall linkage. TBI, traumatic brain injury

4.2 Match Variables and Information Content

We selected 49 match variables and the six blocking variables described in the Methods (▶Table 1). The overall information content in the PHIS file (35.0 bits) was similar to that in the NTDB file (35.6 bits) (see Table II in the ▶Online Appendix). In both files, the information was evenly split between variables that would be present in an administrative file (demographics, hospital, LOS, disposition, ICU admission) and variables derived from ICD-9-CM diagnostic and procedure codes (all others) typically present in a billing file. The information per variable was higher for admin-

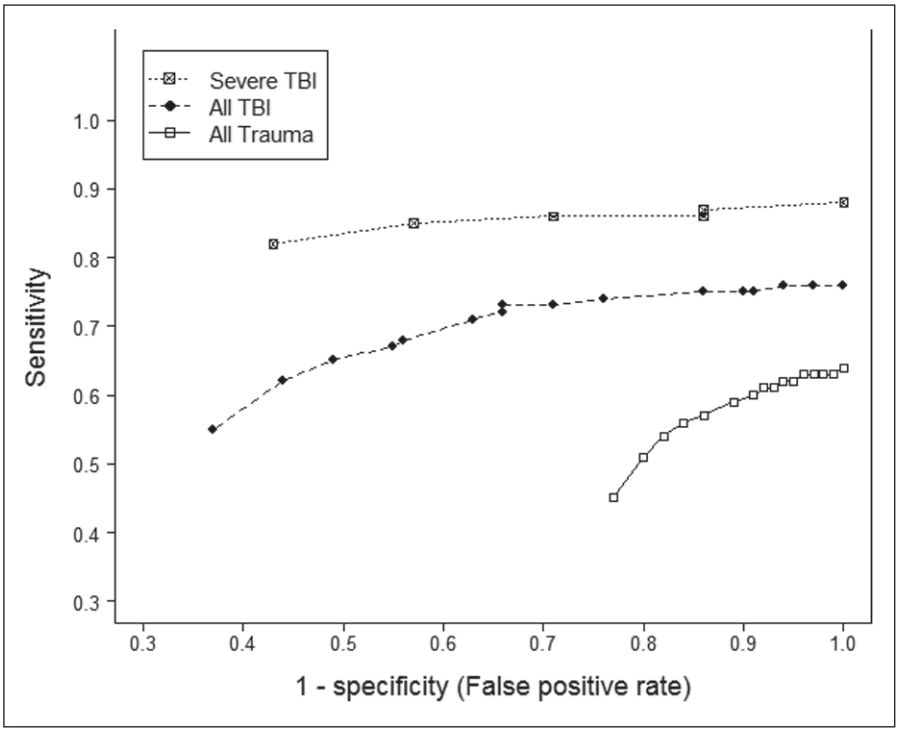

Figure 4 Linkage accuracy by severity of injury. TBI, traumatic brain injury

istrative variables (PHIS, 1.75 bits/variable and NTDB, 1.71 bits/variable) than ICD-9-CM variables (PHIS, 0.45 bits/variable and NTDB, 0.48 bits/variable), overall t-test p = 0.005.

4.3 Validation Sample (Primary Children's Hospital)

We identified 5,524 records of injured children submitted to the PHIS database and 4,306 records submitted to the NTDB from PCH during 2007–2010 (Figure I in the ▶Online Appendix). Over four years, 247 patients from PCH would be eligible for our study of ICP monitoring. Both PHIS and the NTDB gained member hospitals during the study, which is the likely explanation for the increase in patients over time in the non-PCH hospitals (Table III in the ▶Online Appendix). Differences between PCH and other hospitals in age and gender were statistically significant but unlikely to be clinically relevant. Patients from PCH were more likely to be injured in a motor vehicle crash, more likely to be more severely injured (ISS), more likely to be admitted to the ICU, and more likely to die in the hospital.

4.4 Match Parameter Refinement and Match Probability Cutpoint Selection

Match parameter refinement using the MCMC process was necessary to make the sensitivity of this linkage acceptable (▶Figure 2). Because our overall goal was to study the effectiveness of ICP monitoring in children with severe TBI, we chose a match probability cutpoint in the validation linkage to maximize positive predictive value (PPV) and sensitivity without a significant loss of specificity (▶Figure 3). PPV was robust to cutpoint selection in children with severe TBI. We selected a match probability of 0.85 to apply to the larger linkage.

4.5 Validation Linkage (PCH)

We linked PHIS and NTDB records for 69% of the patients from PCH with trauma, 72% with TBI, and 87% with severe TBI. Among those in the validation sample who would be eligible for the effectiveness study of ICP monitoring (LOS ≥ 24 hours and non-missing disposition), we found a sensitivity of 88%, a positive predictive value of 98%, and a specificity of 99.99% (see Figure I in the ▶Online Appendix) for accurate linkage of PHIS and NTDB records. The many (15,030) candidate pairs correctly identified as true negatives by the linkage process relative to the few candidate pairs correctly identified as true matches (214), false positives (4), and false negatives (28) complicates the interpretation of specificity in this analysis. Of the 28 false negatives, one child died in < 72 hours, 12 were discharged home in < 72 hours, 12 were discharged home in 3 to 7 days (median 5 days), and 3 had longer stays. All discharges home were without home nursing. A child who was discharged home alive in < 72 hours may have been classified as having severe TBI because of sedatives and other drugs administered early in their clinical course, but not have a severe injury; however, they still represent false negatives.

Patients with more severe injuries were more likely to link accurately (exact p < 0.001, severe versus non-severe TBI, ▶Figure 4). Patients who linked accurately were more likely to have longer hospital and ICU lengths of stay, higher ISS, and secondary injuries outside the head (▶Table 2). Likelihood of linkage was not related to disposition, head AIS score, or total number of procedures received.

4.6 Overall Linkage (All Hospitals)

In the overall cohort, we linked PHIS and NTDB records for 62% of the patients with trauma, 64% with TBI, and 74% with severe TBI (▶Figure 1). We linked 78% of patients eligible for an effectiveness study of ICP monitoring, totaling 2,165 patients. The denominators for each category shown include true negatives, i.e. patients that were present in the NTDB but not in PHIS and correctly did not link. Because our datasets do not contain identifiers for the overall cohort, a true linkage rate cannot be calculated directly, but is estimated to be similar to that in the validation sample. Likelihood of linkage in the overall sample was associated with age, admission and discharge year, race/ethnicity, hospital and ICU lengths of stay, discharge disposition, ISS, head AIS, secondary injuries, and procedure count (see Table IV in the ▶Online

Appendix). Some of these associations were statistically significant but do not appear to be clinically meaningful differences (admission and discharge year, disposition, ISS, head AIS).

The median proportion of NTDB records linked at each hospital was 82% (range 35–100%, IQR 76–89%). Excluding the two hospitals with proportions of NTDB records linked below 60% changed the median linked proportion to 83% (IQR 77–90%).

Table 2 Linkage status by match variable, validation linkage

	Linked[a] N = 218	Unlinked[a] N = 28	overall P	test (if not exact)
Administrative variables				
Demographics				
Age, years (median(IQR))	6 (3–12)	5 (1.5–10.5)	0.22	ranksum
Admission year, mode (%)	2008 (33)	2008 (36)	0.34	
Discharge year, mode (%)	2008 (30)	2008 (36) b	0.22	
Male, n (%)	131(60)	20(71)	0.31	
Insurance type, mode (%)	"other", (43)	"insurance", (50)	0.11	
Race/ethnicity, mode (%)	"white", (61)	"white", (71)	0.24	
Missing, n (%)	52(24)	6(21)		
Hospital Course				
ICU admission, n (%)	217 (100)	27 (96)	0.22	
ICU days, median (IQR)	3 (1–8)	1 (1–2)	< 0.001	ranksum
Missing, n (%)	1 (0)	1 (4)		
Hospital LOS, days(median (IQR))	9 (4–17)	4 (2.5–5)	< 0.001	ranksum
Discharge disposition, mode (%)	"home", (66)	"home", (93)	0.04	
ICD-9-CM variables				
Injury Severity				
Injury Mechanism, mode (%)	"MVT", (38)	"other" (39)	0.04	
Injury Severity Score (ISS)	17 (10–27)	9 (9–16)	< 0.001	ranksum
Body region AIS scores				
Head, mode (%)	3 (39)	3 (43)	0.31	
Number Non-Head AIS ≠ 0, median (IQR)	1 (0–2)	0 (0–1)	< 0.001	ranksum
Procedures[c]				
Total count, median (IQR)	1 (0–2)	1 (0–1)	0.06	ranksum

5. Discussion

We found that the medical records of children with severe TBI in the PHIS and NTDB databases can be accurately linked without using PHI. This linked dataset can be used to study the effectiveness of ICP monitoring in this population.

MCMC match parameter refinement appears to expand the range of Fellegi-Sunter probabilistic linkage. Accurate linkage may be possible in some scenarios where probabilistic linkage was thought to have significant limitations: datasets containing only variables with relatively low information content, and substantial dependence between variables. Winkler has also reported that good linkage decision rules can be developed if conditional independence between dataset variables is violated [2, 27, 28].

In part because of data security and privacy concerns, privacy-preserving record linkage is an active and robust area of research [29–33]. Our method of linkage without using PHI could be considered a member of that family of methods. When one or more of the datasets that investigators are attempting to link lack PHI, our method may be very useful, assuming sufficient common information content is present. When common PHI is present in the two datasets but the governance challenges of sharing it are prohibitively difficult, our method and, for example, that of Weber et al. [33] might be considered.

One strength of this study is that validation of linkage accuracy was conducted using identifiable data from a single center that submits data to both PHIS and the NTDB. Because the linkage variables are common administrative and billing data from two standardized national databases, validation at a single center should be representative. Other linkages between two standardized databases have been validated in a similar fashion [34, 35]. We matched the records of 88% of our intended effectiveness study population in the validation linkage. Even when PHI is available, accurate probabilistic linkage of approximately 90% of records is common when linking PHIS to other national databases [34, 36]. Of the 10% of children in the validation linkage whose PHIS and NTDB records were not matched, approximately half were early deaths or early home discharges unlikely to benefit from an intervention such as ICP monitoring. Children with altered mental status from sedatives and/or neuromuscular blockade given during the prehospital phase of their care can be inappropriately classified as severe TBI, and the unlinked early home discharges likely are examples of that scenario. GCS has known limitations as a measure of TBI severity [37], but it is the current gold standard.

Our study had several advantages that represent limitations to the generalizability of this technique. First, and likely most importantly, children with severe TBI often require intensive care, long hospital stays, and procedures, all of which generate database information content. We found a direct relationship between severity of illness and match likelihood, and that relationship is likely mediated by information content. Second, ICD-9-CM diagnosis codes for

trauma are multidimensional. Injury type and mechanism, both of which have important information content, can be derived from these codes without additional database variables. Third, children's hospitals common to the two databases are not particularly common (30, in this case). Iteratively, we could estimate a hospital identifier accurately. These advantages are not ubiquitous, but we doubt that they are unique to this linkage.

The process by which users of a Bayesian record linkage method choose 1:1 pairs requires careful consideration. Linked record pairs found by LinkSolv are not 1:1. This is a general concern for hierarchical Bayesian record linkage techniques because imposing a 1:1 constraint after fitting the linkage model can lead to statistical inconsistencies [38]. Jaro's [39] use of a Linear Sum Assignment algorithm is an example of this practice. LinkSolv uses a quick but greedy algorithm to find 1:1 pairs during each MCMC iteration while fitting the linkage model. Because our overall goal was to create a dataset with which to study children with severe TBI, we used a method to maximize the positive predictive value of identified links.

Our group has previously reported a method to estimate the necessary agreement weight (analogous to information content) for a given linkage [10]. That estimation method does not take into account MCMC-augmentation of probabilistic linkage, but it may assist potential users in determining if a linkage is possible. Proposed linkages of relatively common records in the datasets (relative to file size), overall modest file sizes, and rich information content are more likely to succeed.

In conclusion, using multiple imputation and MCMC methods, accurate medical record linkage is possible in the absence of PHI. The success of such linkages is more likely when the population of interest has substantial illness or injury severity requiring prolonged hospital stays and procedures that generate database information content. When investigators or health personnel are attempting a linkage and any one of the datasets lacks PHI, our method may be very useful, assuming sufficient common information content is present. Our method may enable linkages and, in turn, comparative effectiveness studies that would be unlikely or impossible otherwise [1]. The linked dataset of more than 2,000 patients with severe TBI we generated in this analysis can be used to study the effectiveness of ICP monitoring.

Acknowledgments

This work was supported by the *Eunice Kennedy Shriver* National Institute for Child Health and Human Development at the National Institutes of Health (Grant K23HD074620 to TB). We are indebted to David Bertoch and Matthew Hall at the Children's Hospital Association and to Melanie Neal and her team at the American College of Surgeons.

References

1. Weiss NS. The new world of data linkages in clinical epidemiology: are we being brave or foolhardy? Epidemiology 2011; 22 (3): 292–294.
2. Herzog TN, Sheuren FJ, Winkler WE. Data quality and record linkage techniques: Springer; 2007.
3. Winkler WE. Overview of Record Linkage and Current Research Directions. Washington, DC: Statistical Research Division, U.S. Census Bureau, 2006.
4. Fellegi IP, Sunter AB. A Theory for Record Linkage. Journal of the American Statistical Association 1969; 64 (328): 1183–1210.
5. Roos LL, Wajda A. Record linkage strategies. Part I: Estimating information and evaluating approaches. Methods Inf Med 1991; 30 (2): 117–123.
6. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science 1959; 130 (3381): 954–959.
7. Newcombe HB, and Kennedy JM. Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information. Communications of the Association of Computing Machinery 1962; 5 (11): 563–566.
8. United States Department of Health and Human Services. Understanding Health Information Privacy 2014 [cited May 16, 2014]. Available from: http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html.
9. Navathe AS, Clancy C, Glied S. Advancing research data infrastructure for patient-centered outcomes research. JAMA 2011; 306 (11): 1254–1255.
10. Cook LJ, Olson LM, Dean JM. Probabilistic record linkage: relationships between file sizes, identifiers and match weights. Methods Inf Med 2001; 40 (3): 196–203.
11. Belin TR, Ishwaran H, Duan N, Berry S, Kanouse D. Identifying likely duplicates by record linkage in a survey of prostitutes. In: Gelman A, Meng X, editors. Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. Wiley; 2004.
12. Gerber JS, Newland JG, Coffin SE, Hall M, Thurm C, Prasad PA, et al. Variability in antibiotic use at children's hospitals. Pediatrics 2010; 126 (6): 1067–1073.
13. Weiss PF, Klink AJ, Hexem K, Burnham JM, Leonard MB, Keren R, et al. Variation in inpatient therapy and diagnostic evaluation of children with Henoch Schoenlein purpura. J Pediatr 2009; 155 (6): 812–8 e1.
14. Slonim AD, Khandelwal S, He J, Hall M, Stockwell DC, Turenne WM, et al. Characteristics associated with pediatric inpatient death. Pediatrics 2010; 125 (6): 1208–1216.
15. Conway PH, Keren R. Factors associated with variability in outcomes for children hospitalized with urinary tract infection. J Pediatr 2009; 154 (6): 789–796.
16. American College of Surgeons Committee on Trauma. National Trauma Data Bank Research Data Set User Manual, Admission Year 2009. Chicago, IL: December 2010.
17. Langlois JA, Rutland-Brown W, Thomas KE. Traumatic Brain Injury in the United States: Emergency Department Visits, Hospitalizations, and Deaths. Division of Injury Response, Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, 2006.
18. Tri-Analytics Inc. and The Johns Hopkins University. ICDMAP-90 Software User's Guide 1997.
19. Centers for Disease Control and Prevention. Recommended framework for presenting injury mortality data. MMWR 1997; 46 (RR-14).
20. Barell V, Aharonson-Daniel L, Fingerhut LA, Mackenzie EJ, Ziv A, Boyko V, et al. An introduction to the Barell body region by nature of injury diagnosis matrix. Inj Prev 2002; 8 (2): 91–96.
21. McGlincy MH. A Bayesian Record Linkage Methodology for Multiple Imputation of Missing Links. Section on Survey Research Methods, American Statistical Association; 2004. pp 4001–4008.
22. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC; 2004.
23. McGlincy MH. Using Test Databases to Evaluate Record Linkage Models and Train Linkage Practitioners. Section on Survey Research Methods, American Statistical Association; 2006. pp 3404–3410.
24. Shannon CE. A Mathematical Theory of Communication. The Bell System Technical Journal 1948; 27 (3): 379–423, 623–656.
25. Schneider TD. Information Theory Primer 2013 [May 16, 2014]. Available from: http://schneider.ncifcrf.gov/papers/primer/.
26. Mason CA, Tu S. Data linkage using probabilistic decision rules: a primer. Birth defects research, Part A: Clinical and molecular teratology. 2008; 82 (11): 812–821.
27. Winkler WE. Advanced Methods for Record Linkage. Washington, DC: Statistical Division, United States Bureau of the Census; 1994.
28. Winkler WE. Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. Washington, DC: Statistical Division, United States Bureau of the Census; 1993.
29. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC medical informatics and decision making 2009; 9: 41.

30. Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. Journal of Biomedical Informatics 2013.
31. Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. Methods Inf Med 1998; 37 (3): 271–277.
32. Kuzu M, Kantarcioglu M, Durham EA, Toth C, Malin B. A practical approach to achieve private medical record linkage in light of public resources. J Am Med Inform Assoc 2013; 20 (2): 285–292.
33. Weber SC, Lowe H, Das A, Ferris T. A simple heuristic for blindfolded record linkage. J Am Med Inform Assoc 2012; 19 (e1): e157–161.
34. Deans KJ, Cooper JN, Rangel SJ, Raval MV, Minneci PC, Moss RL. Enhancing NSQIP-Pediatric through integration with the Pediatric Health Information System. J Pediatr Surg 2014; 49 (1): 207–212; discussion 12.
35. Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J 2009; 157 (6): 995–1000.
36. Pasquali SK, Jacobs JP, Shook GJ, O'Brien SM, Hall M, Jacobs ML, et al. Linking clinical registry data with administrative data using indirect identifiers: implementation and validation in the congenital heart surgery population. Am Heart J 2010; 160 (6): 1099–1104.
37. Saatman KE, Duhaime AC, Bullock R, Maas AI, Valadka A, Manley GT. Classification of traumatic brain injury for targeted therapies. J Neurotrauma 2008; 25 (7): 719 –738.
38. Larsen MD. Comments on hierarchical Bayesian record linkage. Joint Statistical Meeting, Proceedings of the Survey Methods Section 2002. pp 1995–2000.
39. Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995; 14 (5 –7): 491– 498.