Impact of Record-Linkage Methodology on Performance Indicators and Multivariate Relationships

Kevin M. Campbell, DrPH

Washington State Division of Alcohol and Substance Abuse

Box 45330

Olympia, WA 98504-5330

360-725-3711 (voice)

360-407-1044 (fax)

campbkm@dshs.wa.gov

Abstract

Program evaluation often requires the linkage of records from independently maintained data systems (e.g., substance abuse treatment and criminal justice). Data entry errors (e.g., misspelled names, transposed digits) complicate the linkage task. In this investigation, three record-linkage algorithms (match-merge, common patient identifier, and probabilistic) are used to link recipients of publicly funded outpatient substance abuse treatment to statewide arrest and death data. The impact of record-linkage algorithm on performance indicators, prevalence indicators (i.e., arrest rates, and death rates) and Hazard Ratios derived from a multivariate survival analysis predicting arrests following admission to outpatient substance abuse treatment is evaluated. Choice of algorithm substantially impacted estimates of arrest rates (range: year prior to admission: 39.8%-53.4%, year following admission: 24.7%-33.1). The Hazard Ratio associated with "prior arrest" as a predictor of arrest following admission to outpatient substance abuse treatment (HR range: .20-.37, $p<.05$) was also influenced by algorithm choice.

The Impact of Record-Linkage Methodology on Performance Indicators, Prevalence Indicators and Multivariate Relationships

1. Introduction

A growing body of literature documents the importance of probabilistic record linkage software in the analysis of administrative data, illustrates the increased accuracy of probabilistic algorithms (relative to deterministic or match-merge algorithms), and identifies important considerations in the selection of software to perform such linkage (Campbell, Deck, & Krupski, 2008 ; Christen & Goiser, 2006; Clark, 2004; Contiero et al., 2005; Dal Maso, Braga, & Franceschi, 2001; Gill, Goldacre, Simmons, Bettley, & Griffith, 1993; Gomatam & Carter, 1999; Gomatam, Carter, Ariet, & Mitchell, 2002; Grannis, Overhage, & McDonald, 2002; Jaro, 1995; Jones & Sujansky, 2004; Kendrick, Douglas, Gardner, & Hucker, 1998; Newcombe, Kennedy, Axford, & James, 1959; Wajda, Layefsky, & Singleton, 1991; Weiner, Stump, Callahan, Lewis, & McDonald, 2003; Whalen D, 2001). Probabilistic record linkage software promises to identify unique individuals within data systems and accurately link them to service records across data systems. Unlike match-merge linking (which, for example, requires an exact match on first name, last name, and birth date to link records), probabilistic record linkage software use phonetic equivalence, spelling distance, and probabilistic linkage algorithms to deal with "reasonable" variation in client identifying information. For example, a match-merge of the following data records would fail to link them:

Dean E Bailey 6/30/1995 9549211032 (SSN)

Deane E Baley 6/30/1995 9549211023 (SSN)

The first name and last name are spelled differently and the SSN is different (due to an apparent transposition of the last 2 digits). Given the overwhelming similarity of the information present, however, the probability of these two records belonging to the same person is extremely high. A spectrum of linkage options are available for linking client records but their ability to accommodate the type of reasonable variation described above is highly variable. Match-merge linking, the most restrictive, requires an exact match on all data elements used to establish links. Probabilistic linking forgives reasonable variation in the form of spelling errors, digit transpositions etc. thereby producing a more comprehensive linkage solution.

A particularly clear and comprehensive description of the probabilistic linkage process is provided by Whalen et al. (Whalen D, 2001). In sum, probabilistic linking is accomplished through statistical analysis of the global similarity between data elements for a "record-pair" (i.e., a pair of records containing client names, birthdates, etc.). For each data element common to a record-pair, a "similarity score" (ranging from "no similarity" to "exact match") is calculated. A variety of algorithms are available to generate similarity scores for string variables (e.g., "approximate string matching" (Whalen D, 2001) and the Jaro-Winkler algorithm (Yancey, 2005)). Ultimately, a formula (similar to a regression equation) is derived which generates a "linkage score" for each record-pair and cut-points to identify "definite" matches, "possible" matches, and "non matches". In computing the linkage score for a record-pair, the similarity score for each element is multiplied by a weighting factor (reflecting the relative importance of specific data elements in predicting a match) and a scaling factor. Scaling factors adjust the weights based on the "rarity" of the data value. For example, a common last name (e.g., Smith) would adjust the last name weight downward to that "record-

pair" while an uncommon last name would adjust the last name weight upward to the associated record-pair.

A recent study has established the ability of 2 public domain applications for record-linkage and unduplication (The Link King and Link Plus) to distinguish between "reasonable" variations of this nature from "unreasonable" variation. Sensitivity and positive predictive values in the mid-90s were found for both applications (Campbell et al., 2008). Sensitivity is defined as the proportion of "true" links that were captured by the linkage algorithm. Positive predictive value is defined as the proportion of linked records that, in fact, represent the same person. For example, a linkage algorithm with 90% sensitivity has failed to link 10% of the records that should have been linked. A positive predictive value of 90% means that 10% of the links established by the algorithm are invalid.

Deterministic algorithms fill the gap between the strict match-merge and the more forgiving probabilistic linking. Deterministic algorithms can range from the simple to complex. A simple deterministic algorithm is found in the State of California's Family Outcomes Project (FOP). The FOP has adopted a Common Patient Identifier (CPI) constructed from the following data elements: gender, birth date, birthplace, first 3 characters of first name, and first 3 characters of last name. More complex deterministic algorithms are discussed in the literature (Campbell et al., 2008 ; Gomatam et al., 2002; Grannis et al., 2002).

Probabilistic record-linkage software's ability to maximize sensitivity and positive predictive value in the linkage of client records has been established. However, the potential for choice of

linkage algorithm to impact results of data analysis has been overlooked in current record-linkage literature.

1.1 Research Question

This investigation examines the impact of three record linkage options on 1) performance measures for publicly funded substance abuse treatment programs, 2) arrest rates and mortality rates among participants in outpatient (OP) or intensive outpatient (IOP) substance abuse treatment programs, and 3) multivariate relationships in a survival analysis where the dependent variable is time to arrest following admission to OP/IOP substance abuse treatment. The relative accuracy of these linkages has been established in a prior study (Campbell, 2007). Thus, the primary purpose of this inquiry is not to establish the relative accuracy of the linkages created. Rather, recognizing the diversity of linkage algorithms currently utilized by academic, governmental, and research institutions, this inquiry seeks to identify the impact of a range of reasonable algorithms on a variety of performance and outcome indicators.

2 Materials and Methods

2.1 Linkage Methods

The three linkage methods employed in this investigation include: match-merge linking, Common Patient Identifier (CPI) linking using a modified version of the CPI employed by the State of California's Family Outcomes Project (detailed below), and probabilistic linking by The Link King (a public domain SAS application available from www.the-link-king.com). Essentially, the three linkage methods represent a continuum with match-merge being the most restrictive, The Link King's algorithm being the most flexible, and the CPI somewhere in the

middle. Previous research suggests that the positive predictive value of all three linkage algorithms used in this research would in the low to high 90s (Campbell et al., 2008 ; Gomatam & Carter, 1999; Gomatam et al., 2002; Grannis et al., 2002) and that sensitivity is likely to range from the mid-80s (match-merge) to the mid-90s (probabilistic).

2.1.1 Match-Merge Linkage:

In this investigation, match-merge linkage used the following data elements to link service records: first name, last name, middle name, birth date, and SSN. Middle name (or initial) was missing for 19% of records and SSN was missing for 10% of records. A match-merge link required all data elements with non-missing values to match exactly in order to establish a link. In situations where one member of a record-pair contains a full middle name while the other member contains only a middle initial (30% of records), a match on middle initial was acceptable. Current literature suggests that this match-merge algorithm would have moderate sensitivity (in the neighborhood of 80%) with very high positive predictive value (97% or higher).

2.1.2 Common Patient Identifier (CPI) Linkage:

The CPI used in this evaluation is a modification of the CPI used by the State of California's Family Outcomes Project and requires an exact match on the following data elements: gender, birth date, 1st 3 characters of first name, and 1st 3 characters of last name. Birthplace – included in the Family Outcomes Project's CPI - is excluded because it is not collected in any of the datasets used in this analysis. There were no missing values for first name, last name, birth date, or gender. The CPI is, essentially, a simple deterministic algorithm with decision rules slightly

more relaxed than the match-merge algorithm but considerably less flexible than a probabilistic algorithm. As a result, compared to the match-merge algorithm, one could expect slightly better sensitivity (perhaps in the low 90s) and slightly worse positive predictive value (likely in the low 90s).

2.1.3 Probabilistic Algorithm:

Probabilistic linkages were established using The Link King (www.the-link-king.com), a public domain SAS application for record linkage and unduplication developed by the author of this report. Full details of The Link King's algorithms are available in The Link King Use Manual (Campbell, Deck, Cox, & Broderick, 2006), available at www.the-link-king.com. A recent study estimates The Link King's sensitivity at 96.6% and positive predictive value at 96.1% (Campbell et al., 2008).

2.2 Datasets and Associated Measures

2.2.1 Performance Measures from Washington State's Division of Alcohol and Substance Abuse (DASA)

TARGET, the administrative dataset for Washington State's Division of Alcohol and Substance Abuse, contains service records for over 500,000 individuals who received publicly funded substance abuse assessment, detoxification services, or treatment services from the mid-1990s to the present. This inquiry utilizes data for services received by adults (i.e., 18 years of age or older) in calendar year (CY) 2005.

Upon admission to publicly funded substance abuse treatment or receipt of a publicly funded substance abuse assessment from Washington State's Division of Alcohol and Substance Abuse, an encrypted "Client Identifier" is assigned to each client. Within a given provider's agency, providers have the ability to look-up a client's treatment history at that agency. Additionally, for client's receiving treatment under Washington State's Alcohol and Drug Abuse Treatment and Support Act (ADATSA), providers have the ability to look-up the entire history of ADATSA funded treatment. ADATSA is a legislative enactment providing state-financed treatment and support to chemically dependent indigent persons. ADATSA provides eligible people with substance abuse treatment as well as a program of shelter services if the chemical dependency has resulted in incapacitating physiological or cognitive impairments.

In these two situations – treatment within a provider's agency and ADATSA funded treatment – treatment histories are easily tracked through the linking of services by the encrypted "Client Identifier". In all other situations, treatment histories can only be constructed by using record-linkage software to identify clients who have received services from multiple providers and link their substance abuse related services. This process of linking clients within an administrative data system is referred to as "unduplication". Administrative data systems are particularly likely to benefit from unduplication when clients' may enter the system through receipt of services from multiple unrelated providers. Failure to unduplicate administrative data will result in an inflated client count (i.e., a given individual may be counted multiple times).

Using the administrative dataset for DASA, each of the three linkage algorithms was used to independently unduplicate the listing of clients receiving services in CY2005. Subsequently, for

each of the 3 resulting "client master listings", WC substance abuse treatment performance measures were calculated for CY2005.

The Washington Circle (WC) group is comprised of national experts in substance abuse policy, research and performance management (including representatives from DASA) seeking to improve the quality substance abuse treatment services through the development and use of performance measurement systems(Garnick et al., 2002). In 2007, the WC group developed and pilot tested the following nine performance measures for substance abuse treatment and prevention for use in publicly-funded and commercially-insured systems of care (Garnick, Lee, Horgan, & Acevedo, 2008):

1. Initiation of outpatient treatment (OP)
2. Engagement in outpatient treatment
3. Initiation of intensive outpatient treatment (IOP)
4. Engagement in outpatient treatment
5. Continuity of care after assessment service
6. Continuity of care after detoxification
7. Continuity of care after Short-term Residential
7. Continuity of care after Long-term Residential
8. Continuity of care after Intensive Inpatient

The "initiation" measure for OP and IOP services reflects the percent of individuals admitted to OP or IOP service with no other Alcohol or Other Drug (AOD) services in the previous 60 days

who receive a second AOD service (other than detoxification or crisis care) within 14 days after the index service. The "engagement" measure reflects the percent of individuals admitted to OP or IOP service who both "initiated" outpatient treatment and received two additional services within 30 days after initiation.

"Continuity of care" refers to the percent of individuals who receive an AOD service within 14 days after being discharged from a detoxification, residential, or inpatient stay, or after an assessment that results in a diagnosis of an AOD disorder.

2.2.3 Arrest Rates from the Washington State Patrol

Washington State Patrol (WSP) maintains a dataset of misdemeanor, gross misdemeanor, and felony arrests in Washington State. Each of the three linkage algorithms was used to link the associated "client master listing" (i.e. the unduplicated list of DASA clients created by the algorithm in question) of clients receiving OP/IOP substance abuse services in CY05 to a dataset containing WSP arrests. Subsequently, using results from each linkage algorithm, arrest rates were calculated for the 12-months before and after an individual's the first OP/IOP episode of care in CY05. Arrest rate estimates include arrests for misdemeanor, gross misdemeanor, and felony crimes.

2.2.4 Death Rates from Washington State's Department of Health

Washington State's Department of Health, Center for Health Statistics maintains a listing of all known deaths occurring in the State of Washington. Each of the three linkage algorithms was used to link the associated "client master listing" of clients receiving OP/IOP substance abuse

services in CY05 to a listing of deaths from the Center for Health Statistics. Subsequently, using results from each linkage algorithm, death rates were calculated for the 12-months following an individual's first OP/IOP episode of care in CY05.

2.3 Multivariate Analysis

Cox regression survival analysis was used to identify the relationship between participation in OP or IOP substance abuse treatment and risk of arrest in the year following admission to treatment. Model covariates include age (<18 years, 18-25 years, 25-50 years, and 50+ years), gender, race/ethnicity (Caucasian, African American, Native American, Asian/Pacific Islander, Multi-racial, Other), recent history of substance abuse treatment (any vs. none in the year prior to OP/IOP admission), primary substance of abuse (alcohol, cocaine, marijuana, opiates, other), outpatient modality (OP, IOP), and WC performance indicator status (Engaged, Initiation Only, Neither Initiated or Engaged).

The survival analysis model operationalizes the WC performance indicator as a time dependent variable, allowing an individual's "initiation" and "engagement" status to vary during the 1st 45 days. This model uses the day of admission as the start date and right censors at the earliest of a) date of offence for an alleged crime resulting in an arrest, b) date of death, c) residential substance abuse treatment admission, or d) 365 days after the start date.

3. Results

3.1 Washington Circle Performance Measures:

OP/IOP services following an OP/IOP admission are automatically linked to the admission in DASA's administrative dataset; therefore, in all but a handful of "special case" scenarios there is little opportunity for choice of linkage algorithm to impact performance measures for OP/IOP services. As a result, "initiation" and "engagement" indicators are not included in the comparisons listed in Table 1. Regardless of linkage algorithm, initiation rates for OP and IOP were 73% and 82%, respectively. Engagement rates were 63% (OP) and 76% (IOP)

Overall, the probabilistic algorithm generated the highest values for Washington Circle Continuity of Care (COC) performance indicators (Table 1). However, the difference between performance estimates derived from probabilistic linkages and those derived from the other two linkage algorithms was fairly small. The greatest discrepancy was found in the COC rate for detoxification services where match-merge and CPI continuity-of-care estimates were 21.8% and 23.1% (respectively) while the probabilistic continuity-of-care estimate was 24.9%.

****** TABLE 1 ******

3.2 Arrest Rates and Death Rates:

Overall, the probabilistic algorithm generated the highest estimates for mortality and arrest rates (Table 2). The difference in arrest rates derived from probabilistic and CPI linkages were considerably larger than those found for performance indicators. For example, arrest rates during the year prior to a clients' first OP/IOP episode in CY05 ranged from 40% (match-merge linkage) to 53% (probabilistic linkage).

****** TABLE 2 ******

3.3 Multivariate Relationships:

The risk of arrest among OP/IOP clients "engaged" in treatment was 19-23% less than clients "not initiated or engaged" in treatment (HR=.77 - .81, p<.05) (Table 3). There was no statistically significant difference in the risk of arrest among OP/IOP clients "initiated" (but not "engaged') in treatment and clients "not initiated or engaged". Linkage methodology did not have a statistical or meaningful impact on Hazard ratios.

****** TABLE 3 ******

For most covariates (age, gender, ethnicity, history of substance abuse treatment in 12-months prior to OP/IOP admission, primary substance of abuse, outpatient modality), linkage algorithm did not have a significant or meaningful impact on Hazard Ratio. However, linkage algorithm had a statistically significant and substantial impact on the relationship between prior arrest history and arrest following admission to OP/IOP. Results from all linkage algorithms indicate a reduction in the risk of arrest for clients with no recent arrest history (compared to clients with a recent history of arrest). However, Hazard Ratios for prior arrests incrementally increased from 0.20 for match-merge linkage to 0.37 for probabilistic linkage (p<.05) (Table 4).

****** TABLE 4 ******

4. Discussion

As theory and empirical research would predict, probabilistic linkage consistently produced the highest estimates for WC performance indicators. However, the magnitude of the difference between performance measures based on probabilistic linkage and those based on CPI or match-merge linkage was often fairly small. Unduplication by the probabilistic algorithm yielded a 'master listing' remarkably similar to the match-merge results: match-merge estimated the unduplicated count at 16,044 while the probabilistic estimate was 15,562 (a difference of only 482 or about 3%). This suggests that – during the time period covered by this investigation - admission records for individuals admitted multiple times contained identifying information that was consistent from one admission to the next. A longer study period may produce a greater discrepancy. To some extent, the quality of DASA's data (in terms of consistent identifying information across a client's multiple admissions) may result from the training provided to treatment providers: training and help-desk staff routinely advise providers to enter client identifiers based on review of an official document (e.g., state identification card, drivers license).

On the other hand, match-merge estimates of arrest rates (both in the year prior to OP/IOP admission and the year following OP/IOP admissions) were substantially lower than estimates based on CPI or probabilistic algorithms. Obviously, whenever possible, law enforcement officers use official documentation to identify arrestees; however, some arrestees may be unwilling or unable to provide such documentation. As a consequence of the reduced validity of identifying information in arrest records, both the probabilistic and CPI algorithms linked at least 27% more clients to WSP than the match-merge algorithm.

Observed variation in WC performance indicators and prevalence rates are assumed to result primarily from increased levels of sensitivity in the probabilistic algorithm (compared to match-merge and CPI) and the CPI algorithm (compared to match-merge). The impact of inappropriately linked records is assumed to be negligible given the high levels of positive predictive value associated with the algorithms used in this investigation.

Choice of linkage algorithm did not significantly impact the multivariate relationship between most covariates (including WC performance measures for OP/IOP services) and arrest following admission to OP/IOP. However, estimates of the strength of the relationship between "prior arrests" and arrests following OP/IOP admission was significantly stronger in the dataset created from match-merge linkage (compared to datasets created from CPI or probabilistic linkage).

If failure to link records is a random error then the mechanism responsible for these observed variations in Hazard Ratios (or lack thereof) is fairly straightforward (see Appendix a for elaboration). However, one cannot assume that the error is evenly distributed, and therefore bias in the failure to link records is a distinct possibility. For example, simplistic linkage algorithms (e.g., match merge and CPI) may be less likely to link women than men due to a greater probability of last-name mismatch among women. Even sophisticated linkage algorithms may be biased to some extent. For example, some phonetic equivalence algorithms used by probabilistic linkage routines to determine the degree of similarity between names may be specific to the name's language of origin. The Double Metaphone algorithm (one of the phonetic equivalence algorithms employed by The Link King) incorporates alternative pronunciations

(i.e., the U.S. pronunciation and the native pronunciation) for names from Italian, Spanish, French, and various Germanic and Slavic languages (Phillips, 2000).

To assess the degree of bias in links missed by the match-merge or CPI algorithms but not the probabilistic algorithm, results of probabilistic linking of DASA substance abuse treatment data to WSP arrest records were compared to linkage results from the match-merge and CPI algorithms. Specifically, characteristics of clients linked to WSP arrests by the match-merge algorithm were compared to those of clients linked by the probabilistic algorithm but not by match-merge. Similarly, characteristics of clients linked to WSP arrests by the CPI algorithm were compared to those of clients linked by the probabilistic algorithm but not by CPI. Comparisons were made for the following variables: age, gender, ethnicity, and primary substance of abuse.

In sum, a disproportionate number of links missed by match-merge (but captured by the probabilistic algorithm) were female or belonged to a minority racial/ethnic group. Specifically, 33.2% of clients linked to WSP data by match-merge were female and 35.5% belonged to a minority racial/ethnic group. There were 2,389 arrestees not linked by match-merge but linked by the probabilistic algorithm. Of these, 38.1% were female (vs. 33.2%, $p<.05$) and 39.8% belonged to minority racial/ethnic group (vs. 33.2%, $p<.05$). A similar gender bias was found for the CPI algorithm (compared to the probabilistic algorithm): 34.1% of clients linked to WSP data by the CPI algorithm were female while 41.1% of the 521 clients not linked by the CPI but linked by the probabilistic algorithm were female ($p<.05$).

The important point to recognize is that choice of linkage algorithm has the potential to significantly impact estimates of interrelationships among variables even if the failure to link records is assumed to be random. Further, simple linkage algorithms may be more likely to overlook linkages for women and minority populations (compared to probabilistic algorithms), increasing the potential for error in estimates of interrelationships between variables. As a result, research and evaluation studies with female and minority populations may be particularly susceptible to the impact of record-linkage algorithm on data analyses.

In sum, even in this limited investigation, the impact of record-linkage algorithms on prevalence indicators and multivariate relationships may range from negligible to substantial. As illustrated in Appendix A, the generous assumption that missed links occur at random cannot be used to justify the use of simplistic linkage algorithms. Complex research and evaluation projects involving the creation of indicators whose values are reliant on data linkage from multiple datasets may be particularly susceptible to errors arising from missed links. Accurate results from research and evaluation projects requiring the linkage of administrative datasets can only be obtained through the use of record-linkage software with high levels of positive predictive value **and** sensitivity.

Table 1:

Washington Circle Performance Measures for Continuity of Care in 2005

	Record Linkage Algorithm		
Performance Measure	**Match-Merge**	**CPI[a]**	**Probabilistic**
Assessment	37.8%	38.1%	38.5%
Detoxification	21.8%	23.1%	24.9%
Intensive Inpatient	48.0%	48.85	50.3%
Long Term Residential	35.0%	35.8%	37.7%
Recovery House	42.5%	43.1%	45.1%

[a]Common Patient Identifier

Table 2:

Arrest Rates and Death Rates

Among Adults with a Outpatient (OP) or Intensive Outpatient (IOP)

Substance Abuse Service Episode Beginning in CY05

Prevalence Measure	Record Linkage Algorithm		
	Match-Merge (n=16,044)	**CPI (n=15,871)**	**Probabilistic (n=15,562)**
Arrest Rates			
Year Prior to OP/IOP Admission	39.8%	50.4%	53.4%
Year Following OP/IOP Admission	24.7%	31.4%	33.1%
Death Rate			
Year Following OP/IOP Admission	0.41%	0.56%	0.62%

Table 3:

Hazard Ratios[a] for WC Performance Measure as a Predictor of Arrest following Admission to Outpatient (OP) or Intensive Outpatient (IOP) Substance Abuse Treatment in 2005

WC OP Performance Measure	Record Linkage Algorithm		
	Match-Merge (n=16,044)	**CPI[b] (n=15,871)**	**Probabilistic (n=15,562)**
Not Initiated or Engaged	Reference	Reference	reference
Initiation Only	1.02	1.05	1.05
	(0.91, 1.14)	(0.95, 1.17)	(0.95, 1.16)
Initiation and Engagement	0.77**	0.81**	0.80**
	(0.71, 0.83)	(0.76, 0.87)	(0.75, 0.86)

** p<.05, *** p<.001

[a] Controlling for age, gender, race/ethnicity, ethnicity, recent history of arrest (prior year), recent history of substance abuse treatment (prior year), primary substance of abuse, outpatient modality (OP/IOP). 95% CI in parentheses.

[b]Common Patient Identifier

Table 4

Hazard Ratios[a] for Prior Arrest History as a Predictor of Arrest following Admission to Outpatient (OP) or Intensive Outpatient (IOP) Substance Abuse Treatment in 2005

	Record Linkage Algorithm		
Arrested History in 12-months prior to OP/IOP admission	**Match-Merge (n=16,044)**	**CPI[b] (n=15,871)**	**Probabilistic (n=15,562)**
Arrested	reference	reference	reference
Not Arrested	0.20**	0.32**	0.37**
	(0.19, 0.22)	(0.30, 0.34)	(0.35, 0.40)

** p<.05

[a] Controlling for age, gender, race/ethnicity, recent history of substance abuse treatment (prior year), primary substance of abuse, outpatient modality (OP/IOP), and WC performance indicator status. 95% CI in parentheses.

[b]Common Patient Identifier

Appendix A

The following examples illustrate the potential for record linkage methodology to impact estimates of interrelationships between variables even when it is assumed that failure to link records is a random process.

The first example identifies the impact of linkage algorithm on the relationship between two variables similar in nature (e.g., "prior arrests" and "arrests in the year following admission") where the value of both variables are determined by same linkage task. The potential for choice of linkage algorithm to significantly impact Odds Ratio (OR) estimates is illustrated and explained.

The second example considers the impact of linkage algorithm on the relationship between two qualitatively different variables (e.g., "treatment engagement" and "arrests in the year following admission") where choice of linkage algorithm substantially impacts the distribution of the dependent variable ("arrests in year following admission") but not the independent variable ("treatment engagement"). In comparison to the first example, choice of linkage algorithm is shown to minimally impact OR estimates.

Distributions in the examples given were constructed to reflect those found in the empirical data from this study.

Example #1: A sample of 1,000 clients has been linked to arrest data resulting in a relationship between "prior arrests" and "arrests in year following admission" as detailed in Table A1 (Scenario #1).

****** TABLE A1 ******

Note that 450 records were linked to arrest data (i.e., sum of cells A, B, and D). Of the 450 linked records:

- 50.0% (225/450) had a prior arrest but were not arrested in the year following admission (cell A)
- 33.3% (150/450) had both a prior arrest and were arrested in the year following admission (cell B)
- 16.7% (75/450) had no prior arrest but were arrested in the year following admission (cell D)

Now, assume use of an alternative algorithm appropriately links an additional 150 clients to arrest data (i.e., linked a total of 600 records). If failure to initially link these records was a random error, then the 150 newly linked records would be distributed into cells A, B, and D following the distribution shown above (i.e. Cell A=225+(.5*150)=300, Cell B=150+(.333*150)=200, Cell D=75+(.167*150=100)). The count in Cell C (records not linked to WSP) will be reduced by 150 (the number of newly linked records).

Among clients with a prior arrest, there is no change in the odds of arrest following admission (because the counts in cells A and B are proportionally increased). However, among clients with

no prior arrest, the odds of arrest following admission increase (because counts in cell D increase while counts in cell C decrease). The net result is a dramatic change in the Odds Ratio for "prior arrests" from .20 (Table A1 Scenario #2) to .38 (Table A1 Scenario #2).

Example #2: For the 1,000 subjects in Example #1, assume that the initial data linkage produced a relationship between the WC performance indicator and arrests in the year following admission as detailed in Table A2 (Scenario #1).

Of the 225 clients linked to an arrest in the year following admission:

- 24.4% (55/225) were Not Initiated or Engaged (cell B)
- 8.9% (20/225) were Initiated Only (cell D)
- 66.7% (150/225) were Initiated and Engaged (cell F)

If the alternative linkage algorithm found an additional 75 arrests in the year following treatment admission and the failure to link these records by match-merge was a random error, then the newly linked records would be distributed into cells B, D, and F following the distribution shown above (i.e. Cell B=55+(.244*75)=73, Cell D=20+(.089*75)=27, Cell F=150+(.667*75)=200).

The odds of arrest in all WC categories increase (substantially in some cases) but the Odds Ratio for the "Initiation Only" and "Initiation and Engagement" categories is minimally impacted (Table A2, Scenario #2).

****** TABLE A2 ******

Table A1

Hypothetical Example:

Impact of Missed Record Linkages on Relationship Between

Arrests in Prior to Admission and Arrests Following Admission (n=1,000 clients)

Scenario #1: 450 clients linked to arrest dataset (sum of cells A, B and D).

Arrested in Year Prior to Admission	Arrested in Year following Admission		Odds of Arrest after Admission	Odds Ratio
	No	Yes		
Yes	225 (cell A)	150 (cell B)	0.67	Reference
No	550 (cell C)	75 (cell D)	0.14	0.20

Scenario #2: Additional 150 links to arrest data.

New links proportionally distributed into cells A, B, and D according to distribution in cells A, B, and D in Scenario #1 (i.e., assumes links are missed at random).

Arrested in Year Prior to Admission	Arrested in Year following Admission		Odds of Arrest after Admission	Odds Ratio
	No	Yes		
Yes	300 (cell A)	200 (cell B)	0.67	reference
No	400 (cell C)	100 (cell D)	0.25	0.38

Table A2

Hypothetical Example:

Impact of Missed Record Linkages on Relationship Between

WC Performance Indicator and Arrests in Following Admission (n=1,000 clients)

Scenario #1: 225 clients linked to arrest data following admission (sum of cells B, D, and F).

WC Performance Indicator	Arrested following Admission		Odds of Arrest after Admission	Odds Ratio
	No	Yes		
Not Initiated or Engaged	150 (cell A)	55 (cell B)	.37	reference
Initiation Only	50 (cell C)	20 (cell D)	.40	1.09
Initiation and Engagement	575 (cell E)	150 (cell F)	.26	0.71

Scenario #2: Additional 75 links to arrest data in the year following admission.

New links proportionally distributed into cells B, D, and F according to distribution in cells B, D, found in Scenario #1 (i.e., assumes links are missed at random).

WC Performance Indicator	Arrested in Year following Admission		Odds of Arrest After Admission	Odds Ratio
	No	Yes		
Not Initiated or Engaged	132 (cell A)	73 (cell B)	0.55	reference
Initiation Only	43 (cell C)	27 (cell D)	0.63	1.14
Initiation and Engagement	525 (cell E)	200 (cell F)	0.38	0.69

References

Campbell, K., Deck, D., Cox, A., & Broderick, C. (2006). The Link King User Manual. Retrieved April 1, 2008, from The Link King Web site: www.the-link-king.com\user_manual.zip.

Campbell, K., Deck, D., & Krupski, A. (2008). Record Linkage Software in the Public Domain: A Comparison of Link Plus, The Link King, and a "Basic" Deterministic Algorithm. *Health Inform J, 14*(1).

Christen, P., & Goiser, K. (2006). Quality and Complexity Measures for Data Linkage and Deduplication. In F. Guillet and H. Hamilton (eds), Quality Measures in Data Mining, Vol 43 of Studies in Computational Intelligence, (pp. 127-152). Heidelberg, Germany: Springer.

Clark, D. (2004). Practical introduction to record linkage for injury research *Inj Prev 10*(3), 186-191.

Contiero, P., Tittarelli, A., Tagliabue, G., Maghini, A., Fabiano, S., Crosignani, P., & Tessandori, R. (2005). The EpiLink record linkage software. *Methods Inf Med, 44*(1), 66-71.

Dal Maso, L., Braga, C., & Franceschi, S. (2001). Methodology used for Software for Automated Linkage in Italy (SALI). *J Biomed Inform, 34*, 387-395.

Garnick, D., Lee, M., Chalk, M., Gastfriend, D., Horgan, C., McCorry, F., McLellan, A., & Merrick, E. (2002). Establishing the feasibility of performance measures for alcohol and other drugs. *J Subst Abuse Treat 23*(4), 375-385.

Garnick, D., Lee, M., Horgan, C., & Acevedo, A. (2008). Adapting Washington Circle Performance Measures for Public Sector Substance Abuse Treatment Systems. Unpublished Manuscript.

Gill, L., Goldacre, M., Simmons, H., Bettley, G., & Griffith, M. (1993). Computerized linking of medical records: methodological guidelines. *J Epidemiol Community Health 47*, 316-319.

Gomatam, S., & Carter, R. (1999). *A Computerized Stepwise Deterministic Strategy for Record Linkage*. Unpublished Manuscript.

Gomatam, S., Carter, R., Ariet, M., & Mitchell, G. (2002). An empirical comparison of record linkage procedures. *Stat Med 21*, 1485-1496.

Grannis, S., Overhage, J., & McDonald, C. (2002). *Analysis of identifier performance using a deterministic linkage algorithm.* Paper presented at the Proceedings of American Medical Informatics Association Symposium, Philadelphia, PA.

Jaro, M. (1995). Probabilistic linkage of large public health data files. *Stat Med 14*, 491-498.

Jones, L., & Sujansky, W. (2004). *Patient Data Matching Software: A Buyers Guide for the Budget Conscious* (ISBN 1-932064-74-5). Oakland, Ca: California Health Care Foundation.

Kendrick, S., Douglas, M., Gardner, D., & Hucker, D. (1998). Best-Link Matching of Scottish Health Data Sets. *Methods Inf Med 37*(1), 64-68.

Newcombe, H., Kennedy, J., Axford, S., & James, A. (1959). Automatic Linkage of Vital Records. *Science 130*, 954-959.

Phillips, L. (2000). *The Double Metaphone Search Algorithm*. Retrieved April 1, 2008, from Dr. Dobbs Portal Web site: www.ddj.com/cpp/184401251.

Wajda, R., Layefsky, M., & Singleton, J. (1991). Record linkage strategies: Part II. Portable software and deterministic matching. *Methods Inf Med, 30*, 210-214.

Weiner, M., Stump, T., Callahan, C., Lewis, J., & McDonald, C. (2003). A practical method of linking data from Medicare claims and a comprehensive electronic medical records system. *Int J Med Inform, 71*(1), 57-69.

Whalen D, P. A., Graver L, Busch J. (2001). *Linking Client Records from Substance Abuse, Mental Health, and Medicaid State Agencies (*SAMHSA Publication No. SMA-01-350). Rockville (MD): Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration.

Yancey, W. E. (2005). *Evaluating String Comparator Performance for Record Linkage* (Statistics #2005-05). Washington, DC: U.S. Census Bureau, Statistical Research Division.