MATCHING HOUSEHOLD AND MEDICAL PROVIDER DATA IN THE NATIONAL MEDICAL EXPENDITURE SURVEY

Ayah E. Johnson Ph.D, Barbara Lepidus Carlson MA, Center for General Health Services Intramural Research,Agency for Health Care Policy and Research
Ayah E. Johnson Ph.D., Parklawn Building, Room 18A-55, 5600 Fishers Lane, Rockville, MD 20857.

The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services or the Agency for Health Care Policy and Research is intended or should be inferred. The authors wish to thank Ian Whitlock and Robin McEntyre for their support in this project.

1.0 Introduction

Computerized record linkage is bringing together data from separate sources relating to the same entity. An entity could be a person, a medical doctor, a business or another defined unit. Two potential errors are associated with the linkage process. The first is the linkage of records that correspond to different entities, and the second is the failure to link records that correspond to the same entity.

This is an empirical evaluation of the size of the two types of errors before and after the linkage is executed using data from the National Medical Expenditure Survey (NMES), and the gains in terms of the additional number of links if one relaxes the control over the two types of errors.

2.0 Background

Medical care recipients are not always a reliable source of information on their medical expenditures. They either over- estimate or underestimate the expenditures or in many instances don't know the cost of medical services. A concrete example of those who cannot report adequately their medical expenditures are Medicaid recipients. Provided that this assumption is true, national medical expenditure estimates based on household reported data will be biased due to response error or nonresponse on the part of the interviewee. In order to compensate for this response error and to address the issue of missing data, NMES collected expenditure data from a sample of both household respondents and their medical providers. The assumption being that medical providers such as doctors have accurate records for utilization and expenditures for their patients.

The Household component (HHS) of NMES collected detailed data on the occurrence of health care related events, utilization of medical services, and associated expenditures. A subset of the HHS respondents was selected for the Medical Provider Survey (MPS). The targeted subset were those individuals who were most likely to misreport or not possess adequate knowledge about their medical expenditures (Cohen, DiGaetano and Brick, 1989). This targeted group consisted of persons who had:

o <u>Hospital related care</u> including inpatient stays, outpatient and emergency visits, and clinic visits excluding visits to a school or company clinics;

o <u>Medicaid eligibility</u> that completed the first round of four which were conducted with the household respondent; and,

o <u>Medical providers</u> associated with an admission to nursing homes, and facilities for the mentally retarded.

To allow for methodological comparisons on reporting differentials between household and medical provider reported data at the person level, the MPS survey also included a 25 percent sample of all NMES HHS dwelling units. Consequently, this Medical Provider Survey was designed to obtain medical provider[1] reported-charge data for household reported medical care events.

A computerized matching algorithm developed at Statistics Canada referred to as CANLINK[2] (Canadian Linkage System) was used to match the HHS and MPS data bases. This matching algorithm pairs medical events as reported by the household respondents to the medical-provider's records collected for the same person. A decision is made as to whether the medical events for this person match (H_0 --the null hypothesis) or do not match (H_1 -- the alternative hypothesis). If they are classified as a matched pair, then expenditures reported by the medical provider are moved to the household data base, and are used to get national estimates for medical expenditures.

3.0 Matching Household to Medical Providers

The matching algorithm requires that each pair of records from each data-base be considered for matching. In order to decide whether the pair of records refers to the same entity (e.g. the same doctor visit) a set of common but independently obtained data fields are compared, one at a time. The result of each comparison is an outcome. An outcome could be an agreement, a partial agreement, a missing data status, or a disagreement. A set of probabilities is attached to the outcome. If there is more than one common data field being compared, the probability of the vector of outcomes is the product of the individual probabilities. This probability associated with a given vector of outcomes indicates the likelihood that the medical event reported by the household respondent and the medical event reported by the provider are the same event -- in which case expenditure for that event will be extracted from the medical provider data base to the household data-base.

The probability for a given vector of outcomes depends on the individual probabilities of outcomes, but initially it does not depend on the values of the common data fields in the two data-bases. These individual probabilities of outcomes are based on a representative pretest sample of the matching algorithm. Since they do not depend on the actual values of the data fields being compared, they are called <u>prior probabilities</u>. Analogously, <u>posterior probabilities</u> of vectors of outcomes are estimated conditioned on the actual values of the common data fields. The posterior probability can be obtained after the matching algorithm has been completed.

For the HHS and the MPS data bases ten common data fields (n=10) were used: (1) the medical provider identification number; (2) date of the visit; (3) medical condition; (4) place of visit; (5) reason for the visit; (6) length of a hospital stay; (7) a repeat visit indicator, for those cases where there are multiple similar visits with the same fee, such as allergy shots, hypertension checks, and psychiatrist visits; (8) X-ray services (9) surgical services; and (10) throat culture test.

Table 1.0 provides the list of common fields which were chosen for comparison, a brief description of the rules, and the possible outcomes that the matching rules encompassed. The description of the rule basically depicts the result of the comparison. For example, if both the person and the medical provider indicated that surgery was performed, then the outcome for this event and for this data field is an "agreement" that the person had surgery. However, if the provider reported that no surgery was performed, while the household reported surgery, the outcome would be a "disagreement". In addition, Canlink also allows the user to specify levels of partial agreement. For example, if both the respondent and the medical provider agree that there was an inpatient stay but they do not agree on the length of stay, the algorithm allows the user to consider it as a partial agreement by assigning to that outcome a probability that is lower than the probability of a complete agreement, but a higher probability than is given for a complete disagreement between the common data fields.

Table 1. Canlink Matching Rules

RULES	DESCRIPTION OF THE RULE[1]	OUTCOMES
1. Repeat Visit Indicator	a.It is (not) part of repeat visit series	AGREE
	b.Repeat visit indicator is missing	MISSING
	c. Otherwise	DISAGREE
2. Place of visit	a. Same place of visit	AGREE
	b. Place of visit is missing	MISSING
	c. Otherwise	DISAGREE
3. Reason for visit	a. Same reason for visit	AGREE
	b. The reason is missing	MISSING
	c. Otherwise	DISAGREE
4. X-rays service	a. Received (did not) X-Rays	AGREE
	b. Missing data	MISSING
	c. Otherwise	DISAGREE
5. Throat culture	a. Received (did not) throat culture	AGREE
	b. Missing data	MISSING
	c. Otherwise	DISAGREE
6. Surgical Service	a. Had (did not) have surgery	AGREE
	b. Missing data	MISSING
	c. Otherwise	DISAGREE
7. Length of stay	a. Length of stay> 0, and number of days hospitalized are equal	AGREE
	b. Length of stay=0 & length of stays is equal.	PA2
	c. Difference of ±1, or 2 days but not missing or 0.	PA3
	d. Missing and not 0, or difference > 2 days & not missing or zero)	PA4
	e. Length of stay=0 or 1 day	PA5
	f. Length of stay=0 and either missing or stay > 1 day	DISAGREE
8. Medical condition ICD-9 codes	a. 3 digits ICD-9 exact match	AGREE
	b. Collapsed 3-digits ICD-9 match	PA1
	c. Match on letter code	PA2
	d. No letter code agree	DISAGREE
9. Date rule	a. Exact match	AGREE
	b. Difference of ± 1 day	PA1
	c. Difference of 2 to 6 days and same RR[2]	PA2
	d. Difference of 7 days and RR is equal	PA3
	e. Difference of 14,21,28 days and same RR	PA4
	h. Difference is ±1 day and RR is 1	PA5
	e. Difference is 2 to 6 days and RR is 1	PA6
	j. Difference is >15 days and RR is equal	PA7
	k. Difference is ±1 and RR is -1	PA8
	l. Difference is 2 to 7 days and RR is -1	PA9
	m. Difference is >15 days and RR is -1	PA10
	n. HHS missing date and RR is not 0	PA11
	o. Date is different, and RR is ±2,±3	DISAGREE
10. Medical Provider ID	a. Have the same provider	AGREE
	b. Missing data	MISSING
	c. Otherwise	DISAGREE

[1] Household reported data was compared to medical provider data.

[2] RR is the relative interview round for the HHS respondent and for which the provider reported expenditure information.

[3] PAi is the ith level of partial agreement.

3.1 Matching Household and Medical Provider Pairs

In theory the matching algorithm requires that every record from the Household data base be compared to every record in the Medical Provider data base. This strategy, although complete, would require a very large number of comparisons, the vast majority of which would be non-matches. For the matching of household and medical provider data, that number of pairs exceeded one billion. An alternate strategy was developed to make the size of all potential links more manageable. All records from each of the two data bases, the Household and the Medical Provider, were "blocked" by person. That restricted the records to be paired only if they were reported by or regarding the same person. Within this subspace of potential pairs, the number of potential links still exceeded one million. Based on examination of the data, about 75 percent of these potential links had no chance of being linked because they had no chance of being a pair relating to the same entity, or because they had competitors with much higher weights. Those links were dropped from consideration because their probabilities of being the same entity indicated that they were a definite non-link; and, because dropping them meant a significant reduction in computer cost when using CANLINK. Thus the number of pairs that were considered for linkage after the initial comparison between the two data bases was 253,569.

The matching algorithm is iterative. In the first phase paired records are compared, person by person according to set comparison rules. The probability of an outcome during this phase is a global scoring measure irrespective of whether the values, of the field such as whether a hospital stay was two days or three days, agree or disagree. It is also a measure estimated or based upon a pretest sample. A decision is made as to whether the pairs of records are definite, possible or rejected matches. The probability of outcomes are adjusted to reflect the number of times an outcome occurred among the definite links of the household and medical provider reporters.

In the second iteration these new probabilities of the outcomes are used; frequency weights which are measures of the likelihood of certain values for given fields occur are introduced. The matching algorithm is activated again and further cuts in the data are made. Again the probabilities of the outcomes are updated to incorporate not only the proportion of times a field agreed but on which value it agreed. For example if a person and a provider reported that there was a length of stay, and the length of stay was two days, these new probabilities will incorporate both pieces information: the actual agreement on a hospital stay and the agreement that it was two days. Additional iterations are made until the probability of outcomes stabilizes. For NMES it took only one additional iteration to converge.

Given this process, the concept of posterior probabilities of outcomes can be attributed to any one of these three phases. The one of most interest is of course the last one since it depicts the critical errors associated with the matched linked file that is used for identifying medical expenditure data to be moved.

3.2 Errors During the Matching Process

This linkage process is subject to two types of errors: (1) an erroneous non-match, by failing to link records that correspond to the same medical event (rejecting H_0 when H_0 is in fact true); and (2) an erroneous match, by linking two records that correspond to two different events (accepting H_0 when it is false). Those are the type I and the type II errors (α and β respectively). Two thresholds are computed to control for these two types of error. The decision process is based on a weight (a score based on the probabilities) that is given to each pair of records. If the weight for a pair is above the upper threshold, the pair is classified as a match; if the weight is below the lower threshold, it is considered a non-match; and, if the weight is between the two thresholds, it is considered a possible match. The definition of an optimal linkage rule is one which achieves specified values of α and β errors while minimizing the number of pairs classified as possible links (Felligi and Sunter, 1981). Note that if α and β are low and the number of possible links is zero we have achieved the optimal linkage between the two data bases.

In this paper, the objective is to examine the variations in the thresholds levels and in the number of definite, possible, and rejected links as a function of α and β. In addition, error levels which were set a priori, α and β respectively, are compared with the critical error levels $\hat{\alpha}$ and $\hat{\beta}$, which are achieved after the linkage of the data-bases is completed. These error levels are relevant in the sense that they can help us evaluate the reliability for this matching algorithm of the HHS and the MPS data bases. The computation of the critical error levels is done using a computer algorithm outside CANLINK after the subspace of the 253,569 potential links is identified.

In order to meet these objectives, we: (1) determine the thresholds C_1, C_2 for given values of α and β based on the set of all possible permutations of the n-tuple vector of outcomes, X. This process is executed prior to the actual matching of the data bases; (2) compute the critical error levels, $\hat{\alpha}$ and $\hat{\beta}$, using the prior probabilities of outcomes; (3) compute the critical error levels, $\hat{\alpha}$ and $\hat{\beta}$, using the posterior probabilities of an outcome vector for the different iterations of the matching algorithm; (4) measure the effect of the prior and posterior probabilities on the values of the upper and lower thresholds; and (5) assess the procedure of setting thresholds based on the permutation of outcomes which are determined by the comparison rules, and separate from the data bases. In addition we want to examine changes in the number of definite, possible, and non-matches while varying levels of α and β, in an attempt to achieve an optimal linkage rule.

4.0 Definitions and Terminology

Let O_j denotes the outcome of comparing the jth data field on both files, then:

$$O_j = \begin{array}{ll} D & \text{Disagreement} \\ M & \text{Missing} \\ PA_m & \text{m partial levels of agreement, or} \\ A & \text{Agreement} \end{array}$$

These results of the comparison of all common data fields are defined as an n-tuple vector, X, where n is the number of attributes (common data fields) and where each component of the vector is the outcome from a comparison of the attribute in the data files:

$$X = (O_1, O_2, \ldots\ldots O_n).$$

O_i is a vector of length k, and k varies with each data field, j; so, k= 1, 2, ...n_j. When comparing the household data to the medical provider data, n=10.

Let p_{ini}, denote the vector of prior probabilities of an outcome (n_i = 1,2,..k) given H_0, assuming that the common data fields are independent:

$$P(X=x \mid H_0) = \Pi\, p_{inj} \qquad i=1,2..,10 \quad n_j=1,2,\ldots k.$$

A similar definition applies if H_1 is true.

Once the vector of outcomes has been determined, the matching algorithm estimates how likely it is that a pair of medical records refers to the same event by comparing corresponding fields one at a time to see whether the values agree or disagree.

The statistic that is used to quantify the strength of a match between the household reported event and the medical provider counterpart, is the "log of the odds ratio". The odds ratio is defined as the ratio of the probability of the pair of records being truly matched, to the probability of the records being truly unmatched:

$$\text{Log}_2(\text{odds ratio}) = \frac{\text{Log}_2[P(X=x \mid H_0 \text{ is true})]}{[P(X=x \mid H_1 \text{ is true}]} \qquad (1)$$

If the odds ratio is less than one, it will argue for classifying the pair in the truly unmatched set (U); if it is greater than one, it is more likely that it is a match, and it is classified in the matched set (M). The log transformation of the odds ratio is used to simplify the computational aspect of this process.

The odds ratio can be computed for each of the outcomes, and for the n-tuple vector of outcomes, X, assuming that: (1) each of the data fields is independent, and (2) some prior probabilities for each of the outcomes occurring can be determined[3]. The result of the log transformation of the odds ratio for the vector is called a "weight", denoted by T(x).

The total weight is compared against two threshold values C_1 and C_2. If the total weight is above the upper threshold, $[C_2, \infty)$, it is assigned the status of a "definite" match (accepting H_0). If the total weight is below the lower threshold, $(-\infty, C_1]$, it is assigned the status of a definite non-match. Finally, if it is between the two thresholds (C_1, C_2) it gets a temporary status of possible match which is resolved during the linkage process or manually.

4.1 Determining Thresholds for Set Values of α and β

In the first stage of this analysis we examine the changes in the value of the thresholds C_1 and C_2 as a function of different levels of errors, α and β.

Let's denote the number of possible outcomes by n_j, n_j=1,2,...k, where the value of k varies with the number of possible outcomes for each data field. The total number of configurations that the vector of all possible outcomes could take is: $n_1 * n_2 *\ldots* n_k$. Given this set of potential vectors, X, the type I and the type II error can be expressed as follows:

$$\alpha = P(\, T(x) <= C_1 \mid H_0 \text{ is true}) = \Sigma^{h}_{i=1} Mx(i) \qquad (2)$$
$$\beta = P(\, T(X) >= C2 \mid H_1 \text{ is true}) = \Sigma^{n}_{i=s} Ux(i) \qquad (3)$$

Where Mx is the product of the individual probabilities of outcomes given that H_0 is true, Ux is the product of the probabilities of outcomes assuming H_1 is true; and where:

h is the highest value of i with $T(x) <= C_1$;
s is the smallest value of i with $T(X) >= C_2$; and,
n is the number of outcomes.

The determination of C_1 and C_2 is done by solving equations (2) and (3) for fixed values of α and β using the set of permutations of the outcomes of the 10 rules defined in Table 1 and the different probabilities of outcomes adjusted after each phase of the matching process.

4.2 Deriving Thresholds for Matching the HHS and MPS

A SAS program was developed to determine the values of Mx, Ux, and T(X) (Ian Whitlock, Westat, 1989)[4]. This program was enhanced so that different threshold values are computed as a function of different values of α and β. For given sets of error levels the number of definite, possible, and rejected links is computed using the space of all possible permutations of the decision rules defined by the user. This is actually a way of determining thresholds prior to matching of the data bases, using the distribution of the vector of outcomes.

Table 2 summarizes the results. The objective is to investigate the possibility of relaxing the requirements on β to gain more links. Once the number of definite links have been identified at an acceptable level of β, the strategy is to increase α so that we minimize the number of possible links between the HHS and the MPS. Additional links implied knowledge on expenditure information reported by the provider for the medical events.

In addition, estimates of the number and the proportion of pairs classified as possible links were important to obtain since they can be resolved in one of two ways: (1) by relaxing the restrictions on both error levels; or (2) by manual intervention. The first was more desirable since it implies that CANLINK could in a systematic fashion provide more links, and minimize subjective and time consuming decisions encountered when the process is done manually.

Table 2 shows that for a fixed α, as β is incremented the number of possible links decreases, the number of definite links increases by the same magnitude, and the number of rejected links does not change. The number of definite matched pairs increased by 2,773 when we allowed the probability of including a match that should not be one to increase from one to four percent.

Table 2.

Error Levels		Threshold Values		Number of Permutations		
α	β	C_1	C_2	Definite	Possible	Rejected
0.010	0.010	-22.6	30.8	4,473	6,053	10,210
	0.025		17.4	5,868	4,658	
	0.040		5.6		7,246	3,280
0.025	0.010	-3.5	30.8	4,473	3,838	12,425
	0.025		17.4	5,868	2,443	
	0.040		5.6	7,246	1,065	
0.040	0.010	5.6	30.8	4,473	2,773	13,490
	0.025		17.4	5,868	1,378	
	0.040		5.6	7,246	0	

Conversely, if one fixes β, and lets α vary, the number of possible links decreases, the number of rejected links increases, and the number of definite links does not change. In this case, the number of undecided cases has decreased but we do not get any additional definite matches. Thus the tradeoff between the α and β can be translated as increasing the number of definite links versus increasing the number of rejected links.

This behavior is clear from the theory noted in the previous section. The change in the allowable error of one type while holding the other fixed affects only one of the thresholds; thus it is equivalent to classifying the possible links as either matches or as non-matches.

Table 2 also shows the magnitude of the two types of errors and the respective threshold values when the two errors are equal. At that point, no possible matches exist, and the space of all potential pairs is divided into the "Matched" and the "Unmatched" set. Thus using the distribution of the vector of the decision rules, the equilibrium is reached at the cut-off points $C_1 = C_2 = 5.6$, with $\alpha = .04$ and $\beta = .04$. Since both errors are at acceptable levels the search process for thresholds is determined a priori at that equilibrium point. At that equilibrium, 35 percent of the pairs are classified as definite matches while the remaining permutations were rejected.

5.0 Critical Levels of Errors, $\hat{\alpha}$ and $\hat{\beta}$, During Matching

Reliability of the matching algorithm for the household and the medical provider survey may be defined as the proportion of false matches and erroneous non-matches. In general, the number of unmatched pairs is larger than the number of the matched pairs. Therefore it is desirable to make α smaller than β. In the hypothesis testing context, we impose a small α and we exercise no control over β although we know that it is larger than α.

In this matching project, an optimal decision is one that controls for both types of errors, and minimizes the number of possible matches, and maximizes the number of definite matches. The larger the number of definite matches the greater is the ability to obtain expenditure data from the medical providers that was thought to be more reliable than household reported data. In order to compute the magnitude of these errors in reality, as noted above, there is a need to define the space of potential links that CANLINK separates into the three groups of definite, rejected and possible links. The first and most inclusive space of potential links is the set of all possible pairs that can be created:

$$\text{HHS X MPS} = \Gamma = \{\ (a,b)\text{: } a \ \epsilon\ \text{HHS, } b \epsilon\ \text{MPS}\}.$$

This number of potential pairs within this most inclusive definition is over one billion pairs. This space is too inclusive since it allows for events that are certainty non-matches to be considered for matching. For example visits to the doctor reported by one household respondent are allowed to match to doctor visits reported by another household respondent's doctor. As noted above, it is more realistic to block the events on both data bases so that potential pairs are considered only for the same persons. The new restricted space of one million potential links can then be defined as:

$$\Gamma^* = \{(a,b)\text{: } a\ \epsilon\ \text{HHS, } b\ \epsilon\ \text{MPS, and both a and b are reported by or about the same person}\}.$$

As the matching algorithm proceeded, it was found that a large number of potential links had a weight that was smaller than -50. These pairs were deemed to be non-links, and were therefore eliminated from the matching process. As expected this strategy improved the efficiency of the system by considering a much reduced set of links for further processing. Moreover the inclusion of these non-matches could give us a false sense of a very low β. The practical need is to estimate β for those cases that have some chance of matching, not for those cases that have known but minuscule chance of matching. Thus the subset that is used to identify all definite links between the two data bases is:

Γ^{**} = {(a,b): a ϵ HHS, b ϵ MPS, and both a and b are reported by or about the same person, and their total weight > -50}.

Critical error levels are computed. A critical error level $\hat{\alpha}$ is the probability of obtaining a value of the test statistic as extreme, or more extreme than the one actually observed when H_0 is true. A similar definition applies for the critical error level of $\hat{\beta}$. The critical error levels give a measure of what actually transpired during the matching process using the subspace of potential links that were considered for matching. Since CANLINK is an iterative matching algorithm the critical error levels can be evaluated after each iteration. For the matching of the household and the medical provider data there was an initial evaluation based on the distribution of the vector of possible outcomes, and three iterations, with the last iteration providing the actual critical error levels.

In Table 3 the critical error levels and the thresholds are computed first using the distribution of all possible outcomes for the different decision rules (the first four columns). Then the critical error levels for each of the iterations in the matching process, noted as "B", "C" and "D", are displayed. The subspace of potential pairs in this case is Γ^*, with a minimum weight of -50, and the number of potential pairs is $|\Gamma^*|$ = 253,569.

Table 3

Prior Probabilities		Thresholds[1]		Posterior Probabilities					
α	β	C1	C_2	$\hat{\alpha}_B$	$\hat{\alpha}_C$	$\hat{\alpha}_D$	$\hat{\beta}_B$	$\hat{\beta}_C$	$\hat{\beta}_D$
0.01	0.010	-22.6	30.8	0.025	0.016	0.008	0.014	0.021	0.021
	0.025		17.4				0.076	0.082	0.082
	0.040		5.6				0.130	0.136	0.135
0.025	0.010	-3.5	30.8	0.050	0.035	0.027	0.014	0.021	0.021
	0.025		17.4				0.076	0.082	0.082
	0.040		5.6				0.013	0.136	0.135
0.040	0.010	5.6	30.8	0.068	0.055	0.055	0.014	0.021	0.021
	0.025		17.4				0.076	0.082	0.082
	0.082		5.6				0.130	0.136	0.135

[1] These thresholds were set a-priori based on the distribution of the outcomes

These results indicate that both α and β vary after each iteration, but the variation is at acceptable levels. The posterior probability of a non-match when it should be a match, $\hat{\underline{\alpha}}_A$, increases by two-fold or a little less after the first iteration, but converges to the prior probabilities by the third iteration. The greatest difference is for the α =0.04. The actual value by the end of the process is 0.048, which is a small and acceptable increase. Although $\hat{\alpha}$ is stable, $\hat{\beta}$ is much higher than estimated a priori: it is between 1.4 to 3.38 times higher for different values of α. The second and the third iterations show little change in the values of $\hat{\beta}$, indicating that the process converged.

Since the matched set in practice was identified by setting the thresholds at C_1 = C_2 = 5, the estimates of critical error levels for the matched set are $\hat{\alpha}$ =.048 and $\hat{\beta}$ =.135 respectively. This is an indication that we were able to control for the two errors in matching. There are about five percent of the pairs which did not match and should have ($\hat{\alpha}$), and 14% of the pairs that matched but should not have ($\hat{\beta}$).

If the objective was to force the two types of errors to be at certain levels, the thresholds would need to be different. The thresholds satisfying the α and β a priori, are compared to those determined based on the distribution of the vector of outcomes (Table 4). This table shows that in order to obtain the same levels of errors set a priori the thresholds need to be adjusted considerably. A higher lower bound (C_1) is needed to achieve an α of 0.01. On the other hand, to achieve an α of 0.025 or an α of 0.04, C_1 should have been lower: -4.6 not -3.5 and 4.4 and not 5.6. This information could be useful if one wishes to increase the number of definite classifications, and reduce the amount of indecision identified by the algorithm.

Table 4.

Prior		**Thresholds**[1]		**Thresholds**	
α	β	C_1	C_2	C1(D)	C2(D)
0.010	0.010	-22.6	30.8	-17.1	41.1
	0.025		17.4		29.8
	0.040		5.6		25.6
0.025	0.010	-3.5	30.8	-4.6	41.1
	0.025		17.4		29.8
	0.040		5.6		25.6
0.040	0.010	5.6	30.8	4.4	41.1
	0.025		17.4		29.8
	0.040		5.6		25.6

[1] These are thresholds computed prior to production

To control for β, the probability of linking a pair that should not be linked, the upper threshold should have been uniformly higher. For β=0.01 it is 1.3 times higher, for β=0.025 it is 1.7 times higher, and for β =0.04 it 4.5 times higher than the thresholds set a priori. This explains the higher level of the β than was estimated based of the prior cut-offs.

Table 5 examines the number of links that were classified as acceptable, possible, or rejected. The first set of numbers was obtained using the thresholds determined a priori, and the second set were computed using the last set of thresholds [C_1(D), C_2(D)]. It is clear that if the objective was to match while controlling for fixed type I and type II errors , the number of definite links would be lower than expected and the number of rejected links would be higher. Moreover the table indicates that as β is set at a higher level, from 0.01 to 0.04, the number of definite matches increases by 48,704 using the prior cut-off points, and it increases by 25,259 when using the posterior cut-off points.

When α changes from 0.01 to 0.04 the number of rejected links increases in smaller but significant proportions: 33,349 using the prior cut-off and 24,240 using the posterior cut-off points. Also the weight that will lead to the point of equilibrium where the number of possible links is zero, is 15 when using the posterior cut-off bounds and not the 5 that we set a priori. This number was obtained from examining critical error levels curves and identifying the intersection point.

Table 5

Links/ on Prior Cut-off			Links(d)/Posterior Cut-off		
Rejected	Possible	Definite	Rejected	Possible	Definite
112,928	82,053	58,588	119,496	91,486	42,587
	54,946	85,695		72,773	61,300
	33,349	107,292		66,227	67,846
133,710	61,271	58,588	132,250	78,732	42,587
	34,164	85,695		60,019	61,300
	12,567	107,292		53,473	67,846
146,277	48,704	58,588	143,736	67,246	42,587
	21,597	85,695		48,533	61,300
	0	107,292		41,987	67,846

6.0 Matches of Medical Events after Linkage

The previous sections indicated that both the probability of matching when it is a non-match, and the probability of a non-match when it should be one are relatively low for upper and lower thresholds of 5. This measure, although useful, does not supply the whole answer to the question of how well the matching algorithm worked.

During the matching process we did not distinguish between the type of events per se, such as doctor visits, outpatient visits, emergency room encounters, and hospital stays. For matching purposes all records were in one data file. It was important to check that the linkage was actually done for the same person by the same provider and for similar types of visits. Using the link file, Table 6 summarizes the number of linkages that were correctly linked on this dimension and those that were misclassified. The definition of correctly classified implies for example, that doctor visits were matched to doctor visits. Otherwise they are considered as misclassified.

Table 6

	Number of Events Classified	
Medical Encounter	Correctly	Incorrectly
Doctor Visits	31,758 (99.8%)	67 (0.2%)
Outpatient	6,642 (98.0%)	209 (2.0%)
Emergency Room	3,779 (94.0%)	119 (6.0%)
Hospital Stays	3,416 (91.0%)	339 (9.0%)

The misclassification cannot always be attributed to the matching algorithm. It can be a result of a respondent error in identifying the place where he/she had a medical encounter. In general, most of the cases seemed to have been matched for a similar encounter. However, encounters to emergency rooms and hospitals that are in the misclassified column should be examined to ensure that they are not due to the matching algorithm since they are less likely to be misclassified by the respondents.

7.0 Conclusions

The Canlink matching algorithm was used to match data from two surveys, concerning the same entity -- the same person. The first was a survey of household respondents and the second was a survey of their medical providers. Researchers (Winkler, Felligi and Sunter, Kirkendall) have discussed at length the two types of errors, the need to control for these errors and the various definitions of what is an optimal decision rule. In this study we have investigated empirically the size of these two types of errors before and after the matching algorithm was executed, and we have analyzed the gains in terms of the additional number of links gained if one relaxes the control on the types of errors. The results indicate differences in using the prior and posterior probabilities. If the objective to fix the two types of errors one has to examine the posterior probabilities and use posterior upper and lower bounds. However, if the objective is to match two data bases and maintain reasonable and acceptable levels of the two type of errors, as was the case for NMES, determining the thresholds based on the distribution of the vector of outcomes a priori worked.

Additional issues for future research are whether we could have increased the number of definite links if we added or changed the variables identifying the entities in each of the file, or change the rules to allow for more detailed comparison of the data fields.

REFERENCES

1. Generalized Iterative Record Linkage System (1985). Research and General Systems. Informatic Services and Development Division Statistics Canada, Ottawa, Onterio.

2. Record Linkage Techniques (1985). Proceedings of the Workshop on Exact Matching Methodologies, Arlington Virginia.

3. Newcombe, H.B. and Abbat J.D. (1983). "Probabilistic Record Linkage in Epidemiology." Red Book Series Report No. 5. Eldorardo Resource Limited, Suite 400, 255 Albert street, Ottawa Ontario, KIP 6A9.

4. Jabine T.B. and Sheuren F.J. (1986). Record Linkages for Statistical Purposes: Methodological Issues. Journal of Official Statistics, Vol. 2. No. 3, pp 255-277.

5. Kirkendall, N.J. (19). Weights in Computer matching: Applications and an Information Theory Point of View. Energy Information Administration.

6. Report on Exact and Statistical Matching Techniques (1980). Prepared by the Subcommitee on Matching Techniques. Federal Committee on Statistical Methodology.

7. Tepping B.J. (1968). A Model for Optimum Linkages of Records. Journal of American Statistical Association, 63: pp 1321 - 1323.

Endnotes

1. A medical provider is any Medical Doctor (D.O.) who provides direct patient care; any other medical provider providing care under the supervision of an M.D. or a D.O; any person providing home health services.

2. CANLINK is also known as "GIRLS"-- Generalized Iterative Record Linkage System.

3. The probability of an outcome for a given data field is usually based on experience or on a sample matching test that is done prior to production.

4. The program was not written for NMES, and has not been published.