PROBABILITY MATCHING OF MEDICAL EVENTS

Marianne Winglee, J. Michael Brick, Richard Valliant, Carmen Vincent, and Amy Lavis, Westat, Inc.; Steven Machlin, AHCPR

Marianne Winglee, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: probability matching, linkage errors, and medical events

1. Introduction

This paper discusses the evaluation of linkage error when a probability-based method is used to link medical events reported by patients to the same events reported by medical providers. The data are from the 1996 Medical Expenditure Panel Survey (MEPS). This survey collected information, from both patients and their medical providers, about medical events that occurred during a reference period. The information provided by the two sources, however, is not always identical. For example, the patients may recall the date of medical services incorrectly or they may disagree with their providers on the medical conditions treated. By linking events reported by patients to the events reported by providers, the data from the two sources can be used together to support estimation of medical expenditures in the United States.

Section 2 provides an overview of the 1996 MEPS and the medical events included for linkage. Section 3 outlines the principles of probability matching and the use of match weights to determine whether a record pair refers to the same event. Section 4 describes the method used to link the medical events in the 1996 MEPS. Section 5 discusses the linkage outcome and the methods to evaluate linkage errors. Section 6 provides a summary.

2. MEPS and Medical Events for Linkage

The 1996 MEPS is a panel survey sponsored by the Agency for Health Care Policy and Research. The main component is a household survey (HHS) designed to assess the health care utilization and expenditures of the civilian non-institutional population in the United States. The 1996 panel consists of almost 10,000 households. These households were interviewed five times over a period of 2 years through computer assisted personal interviews. A qualified household member reported for each member of the household. Medical events were collected for both 1996 and 1997. This study used only the 1996 events.

The payment system for medical services in the United State is rather complex. Household reports of medical expenditures may not be complete because the patients may not know some of payments made through medical insurance plans, and the separate charge components associated with hospital events. To supplement the household data, a medical provider survey (MPS) was conducted with a sample of the providers identified by the households. With the consent of households, providers were contacted by phone and were asked to provide information about all the medical visits associated with sampled patients that occurred in 1996.

The 1996 MEPS panel yielded 8,992 sampled persons with data from both patients and providers. For these persons, a HHS medical-events file was created to include the household reported medical events in 1996 that were associated with providers who participated in the MPS. A corresponding MPS medical-events file was created to include the medical events reported by the providers for sampled patients. There were over 15,500 patient-provider pairs included in these files. The patient-provider pairing was accomplished through careful data management. In some cases, new providers were identified during MPS because the patients might have misidentified some providers, or because some providers might have changed practices (i.e., by merging or splitting with other practices). A patient-provider pair identification code (PAIRID) based on information from the surveys identifies the correct patient-provider pairs.

Table 1 shows the number of medical events in the HHS and the MPS files included for linkage. While the households and the providers were ostensibly reporting the same events, the numbers of events reported by the two sources are not equal for various reasons, including different types of nonresponse and response error. Households reported fewer medical events than providers, and there appears to be some classification differences for outpatient (OP) events that occurred in hospital-based facilities and other medical visits (MV) in physician's offices. The events HS, ER, and OP are hospital-based; MV events are office-based. The IC events in this file do not include nursing home events.

Table 1. Number of medical events for linkage

Event type		Number	
		HHS file	MPS file
Hospital stay	HS	1,847	1,869
Emergency room	ER	3,112	4,124
Outpatient	OP	7,723	15,137
Medical visits	MV	35,519	34,522
Institution care	IC	55	56
Total		48,256	55,708

3. Probability Linkage Method

The probability linkage method is designed to accommodate discrepancies in the responses from two sources, and establish the best overall linkage under conditions of uncertainty. This section outlines some of the mathematical details of the probability linkage method. A probability based method is necessary because of reporting errors from the two sources.

Fellegi and Sunter (1969) provide the theory of probability matching. Briefly, this method involves the following steps. Take each record from one file, A, and compare it with each record from another file B. Assign a weight to each pair based on its likelihood of being a match (corresponds to the same event), and declare a pair to be a match if the weight is sufficiently large.

In the basic setup, the weight assigned to a pair of records is derived from a likelihood ratio that accounts for the closeness of the fields being compared for each pair, assuming that the fields are independent. We use r for a record pair, i for a field compared where i=1,..,I fields. The weight of a record pair w_r is:

$$w_r = \log_2 \left[\frac{\prod_{i=1}^{I} m_i^{y_{ri}}}{\prod_{i=1}^{I} u_i^{y_{ri}}} \right]$$

where m_i = Pr (field i agrees in pair $r \mid r \in M$), M is the set of true matches, u_i = Pr(field i agrees in pair $r \mid r \in U$), U is the set of true non-matches, and y_{ri} = 1 if field i agrees and 0 otherwise. The weight w_r is a type of log-odds or log-likelihood ratio.

By taking the anti-log of w_r, we have

$$2^{w_r} = \frac{L(\mathbf{y}_r | r \in M)}{L(\mathbf{y}_r | r \in U)} \equiv LR(\mathbf{y}_r),$$

where $\mathbf{y}_r$ is the vector of 0's and 1's for disagreements and agreements of the component fields in pair r. $L(\mathbf{y}_r | r \in M) = \prod_{i=1}^{I} m_i^{y_{ri}}$ is the likelihood of a particular configuration of agreements and disagreements among the fields given that the pair is a true match, and $L(\mathbf{y}_r | r \in U) = \prod_{i=1}^{I} u_i^{y_{ri}}$ is the likelihood of the same configuration given that the pair is really a non-match. The transformed weight, a likelihood ratio $LR(\mathbf{y}_r)$, is a measure of the strength of evidence that a pair is a match. When matching MEPS records, we allowed for partial agreements between fields, as described in Section 4, rather than just agreement or disagreement. The mathematical information of this more elaborate application is similar to that above.

Determination of a threshold for classifying a record pair as a true match or a non-match is not straightforward. With the Felligi and Sunter method, the weight of each record pair is compared to an upper (w_u) and a lower (w_ℓ) threshold and the pair is declared to be a link if $w_r \geq w_u$, a potential link if $w_\ell < w_r < w_u$; or a non-link if $w_r \leq w_\ell$. Pairs that are potential links are clerically reviewed and classified. A single threshold is often used in practice since a manual review is not always possible. This threshold is ideally selected to control the linkage errors of (1) declaring a pair to be a link when it is not (i.e., a false positive, FP, error) and (2) declaring a pair to be a non-link when it is (a false negative, FN error).

4. Matching the 1996 Medical Events in MEPS

The comparison of every medical event in the HHS file with every event on the MPS file would require over 2.8 billion comparisons. This is impractical and unnecessary because the medical events were reported for specific patients in the two surveys. Therefore, the data files were partitioned into person-blocks, each block consisting of the events for a person, and only events within the blocks were compared. Within the person-blocks, comparisons were made first for events reported for the same patient-provider pairs; the remaining unmatched events were compared across providers within a person. This two-pass approach gave priority to matching events within patient-provider pairs, and relatively few events were matched across providers. The within-person comparison generated 776,310 record pairs, a significant reduction from the 2.8 billion.

The software package used for matching is AutoMatch (1996). This package uses a model-based method to estimate the conditional probabilities m_i and u_i, for each field through an iterative process, and calculates the log-odds weights for record pairs. A particular patient record can have several provider records that are potential matches, and a decision must be made as to which is the best match. Given a set of weights, the assignment of pairs as matched or non-matched uses a linear sum assignment algorithm. This algorithm selects the set of matched pairs with the maximum sum of weights in a block. The assignment involves only those record pairs with a match weight above a user-specified threshold.

The fields for comparison were selected because they were the fields reported in both HHS and MPS that are found to be effective in identifying the true pairs. The match fields used are: date (year, month, day), duration of hospital stay (number of days), medical conditions (ICD9 codes summarized into

Table 2. Matching rules and weight

Match field	Match rule	Weight
DATE	Exact match	8.52
	Off +/- 1 day	5.71
	Off +/- 2 day	4.90
	Off +/- 3 day	4.09
	Off +/- 4 day	3.28
	Off +/- 5 day	2.47
	Off +/- 6 day	1.66
	Off +/- 7 day	2.84
	Off +/- 8 day	0.03
	Off +/- 9 day	-0.78
	Off +/- 10 day	-1.59
	Off +/- 11 day	-2.40
	Off +/- 12 day	-3.21
	Off +/- 13 day	-4.02
	Off +/- 14 day	-2.83
	Off 21,28,35,42,49 56 days	-3.64
	Off +/- 15 days to 60 days	-5.64
	>2 month, same day of week	-4.64
	>2 month	-6.64
DURATION HOSPITAL STAY (DAYS)	Exact match	5.93
	Off +/- 1 day	5.22
	Off +/- 2 day	4.50
	Off +/- 3 day	3.78
	Off +/- 4 day	3.07
	Off +/- 5 day	2.35
	Off +/- 6 day	1.63
	Off +/- 7 day	0.92
	Disagree	0.20
CONDITION	All elements agree	5.45
	Partial (approx.)	4.00
	All disagree	-1.52
SERVICE	Agree 111	3.69
	Agree 112	3.26
	Agree 121	2.44
	Agree 122	2.93
	Agree 211	3.26
	Agree 212	2.50
	Agree 221	2.44
	Agree 222	1.68
	DISAGREE	-2.77
	Partial 11D	2.68
	Partial 12D	1.86
	Partial 1D1	2.31
	Partial 1D2	1.88
	Partial 1DD	0.89
	Partial 21D	1.92
	Partial 22D	1.10
	Partial 2D1	1.55
	Partial 2D2	1.12
	Partial 2DD	0.54
	Partial D11	2.39
	Partial D12	1.97
	Partial D1D	1.38
	Partial D21	1.57
	Partial D22	1.15
	Partial D2D	0.56
*D=element Disagreed	Partial DD1	1.30
	Partial DD2	0.54
GLOBE FEE	Agree: Yes	2.01
	No	0.02
	Disagree	-0.27

65 categories), services (surgery, radiology, and laboratory tests), and global-fee (indicates whether event is paid for as part of a package).

The date field was compared for exact matches. Differences by a specified number of days, by week and by month were assigned partial agreement weights. For example, record pairs with exact agreement on date received a weight of 8.52. Record pairs with some difference in date received a lesser weight proportional to the size of the difference and adjusted if there was agreement on the day of the week or the month. A total disagreement was assigned a weight of -6.64. As expected, date was the field with the most discriminating power in the study.

The length of hospital stay was compared as a numeric field, prorated to allow for differences of a specified number of days. Since this field was only available for hospital stay events, the disagreement weight for stays that were more than seven days apart was assigned a small positive value to improve the likelihood of matching hospital stay events to hospital stay events.

Medical conditions and services were compared as numeric arrays. The medical condition arrays contain zero to 10 elements because multiple conditions could be reported for the same event. The service arrays have a fixed number of three elements (indicating surgery, radiology, and laboratory tests). The arrays are used to provide a way of handling correlated fields without substantially violating the independence assumption. The arrays can consist of different numbers of elements, yet the overall weight is constrained so it does not exceed the weight that would result from the comparison of a single element. This keeps the weights for array comparisons from dominating weights for other fields.

For the service array, the frequency weight option (in AutoMatch) was used. This feature enables weights to be adjusted depending on the particular values occurring. Rare values can be assigned greater weights because they have greater discriminating power. The services occurred in a small percentage of events: Surgery and radiology were identified in about 10 percent of events; laboratory test occurred in about 18 percent of events.

In Table 2, 3-digit codes are used to show agreement or disagreement for the services. For example, "Agree 112" means that the HHS and MPS records both showed surgery and radiology for the event but no lab tests. "Partial 1D1" means that both files showed surgery and lab tests for the event, but one file showed radiology while the other did not.

Global fee is an indicator of whether an event is part of a payment package. It was compared as a numeric field for exact agreements and disagreements. The frequency weight option was used so agreement on a "yes" response was assigned a much higher weight than agreement on a no response.

As noted in Section 3, the overall weight for a record pair was the sum of the weights across the five fields. For example, for complete agreement on all fields, the total weight for a record pair might be 23.18. This is obtained by adding 8.52 (exact match on date) +5.93 (exact match on duration of hospital stay, if both are hospital stay events), 5.45 (agreement on condition array) +3.26 (if the pattern for the service array was Agree 112) + 0.02 (agreement "no" on global fee). Missing values are assigned a zero weight for all fields.

The match weights shown in Table 2 are derived at the threshold weight of 1 (Section 5 discusses the choice of threshold). These weights are estimated by using the iterative model-based estimation procedures in AutoMatch, and then adjusted to reflect systematic reporting errors. The goal is to attain a set of matched pairs from AutoMatch that closely resembles the characteristics of true matched pairs. For the purpose of developing match rules, we selected a pilot sample of about 800 persons and manually matched their medical events. Data managers familiar with the data conducted the match. They utilized all available information in determining the correct links including data (such as descriptive information) that cannot be used in AutoMatch. For the purpose of this evaluation, the manually matched pairs are considered the "true" matches, even though it is clear that different manual matches were possible.

For example, Table 3 shows the percentage distribution of the AutoMatch selected matched pairs and that of manually matched pairs on event date. Based on both methods, about 70 percent of matched pairs agreed on date exactly.

Table 3.Percent of Matched Pairs by Agreement on Date

	Percent	
Date in record pair	Manual match	AutoMatch
Exact match	70.9	69.3
Off +/-1 day	7.1	6.8
Off +/-2 days	2.9	2.9
Off +/-3 days	1.9	2.3
Off +/-4 days	1.7	1.8
Off +/-5 days	1.3	1.4
Off +/-6 days	1.2	1.5
Off +/-7 days	3.0	3.0
Off +/-8 days	0.7	0.7
Off +/-9 days	0.7	.04
Off +/-10 days	0.9	0.4
Off +/-11 days	0.3	0.3
Off +/-12 days	0.5	0.2
Off +/-13 days	0.4	0.2
Off +/-14 days	0.3	0.4
21,28,35,42,49,56 days	0.4	1.4
Off +/-15 days-2 month	2.2	1.6
>2 month, same day	0.0	2.0
>2 month	0.1	0.3

5. Estimating Threshold and Error Rates

This study used a weight of 1 as the minimum threshold to determine whether a record pair is a match or non-match. This means that a record pair has to have a minimum weight of 1 (or, using the likelihood interpretation, be twice as likely to be a match as a non-match) before it is considered for selection as a matched pair. Other threshold levels evaluated were 0, 1, 2, 3, and 4. We selected the threshold of 1 as a good balance between linkage errors and match outcome. Since the linked pairs of events are used to supplement household responses, this threshold allows better inclusion of MPS data.

Table 4 shows the match outcome and the estimated linkage error at threshold levels 0, 1, and 2. The threshold at 1 attained a match rate of over 86 percent. That is, 86 percent of the medical events in the HHS file, that have a chance of linkage with the events in the MPS file within a person-block, was matched. The estimated FP error is about 4 to 5 percent, the FN error is about 2 to 5 percent. The error estimates are presented as ranges because different estimation methods provided slightly different estimates. The methods used to determine the linkage error are discussed below.

Table 4. Threshold and match outcome

	Threshold		
Outcome	0	1	2
Match pairs:			
Number	36,515	35,585	34,520
Rate	88%	86%	83%
Error rates:			
False positive	9-30%	4-5%	1-4%
False positive	1-4%	2-5%	4-8%

One method described by Jaro (1989) to determine the linkage error uses a weight chart. The basic steps involve ordering all possible configurations of agreement and disagreement of the match fields by w_r. Then plot the cumulative distribution function of weights for matched and unmatched pairs (the M-curve and the U-curve). Use the weight chart to determine thresholds to attain desired levels of FP and FN errors. There appears to be no entirely satisfactory way of estimating these curves.

Ideally the M and U-curves would be estimated from a set of pairs for which the truth is known. The goal, in this case, is to model the error-making behavior of AutoMatch as it is applied to the MEPS matching problem. We would begin with a large set of correctly matched pairs, run them through AutoMatch to obtain a weight for each, and observe what proportion is above or below a given threshold. Similarly, a large set of pairs, known to be true non-matches, would be assigned weights, and again tabulate the proportions of them on

either side of the threshold. The proportion of true matched pairs with weights below the threshold and the proportion of true non-matched pairs with weights above the threshold would then be estimates of the error rates associated with the way in which the matching algorithm is implemented.

This ideal approach is often not feasible because obtaining a "truth" set generally requires manual matching, which, if done on a large scale is expensive. This study used two alternative methods to develop the M- and the U-curves and found discrepancies between the two approaches. One method combines both manual match (described below) and AutoMatch results to provide the M- and U curves. An alternative method uses a simulation to generate the simulated M- and U-curves. Figure 1 shows the curves developed by these two methods at the threshold weight of 1. These are described below more completely.

A manual match M-curve is generated by applying the AutoMatch weights to record pairs identified through manual review. A second manual matching was conducted on a random sample of about 500 persons (over 2,500 events). Data managers conducted the match, and as before, the manual pairs are considered as the "true" pairs for the purpose of evaluation. The manual matched pairs are assigned the weights derived from AutoMatch to generate a cumulative distribution function. As shown in Figure 1, this manual match M-curve crosses the minimum threshold of 1 leaving about 5 percent of the "true" matches with a weight below 1. Therefore, with threshold at 1, we estimate that the FN rate is about 5 percent.

The second component of this chart, the U-curve is generated using samples of about 1,000 events from the two events files. These events were selected using a simple random sampling with replacement design. AutoMatch was used to generate the match weight for all possible pairs (i.e., 1 million pairs, ignoring the blocking within persons). The AutoMatch U-curve is 1 minus the cumulative distribution of the weights of these pairs, and is shown in Figure 1. This curve shows that about 5 percent of the unmatched pairs had a weight greater than 1, suggesting a FP error of about 5 percent. This method ignores the person-blocking and estimates the proportion of times that randomly paired events from the HHS and MPS files would have a weight greater than the threshold. Since the probability of correctly pairing events at random is negligible, we interpret the proportion of weights above the threshold as being entirely due to error.

To ascertain the stability of this solution, a simulation was conducted. The m and u probabilities for individual fields were estimated from AutoMatch. For a matching rule and its subcategories, like date, we tabulated the relative frequencies for pairs that were matched using the software. These relative frequencies were used as estimates of the m_i's described in Section 3. The relative frequencies for unmatched pairs were used as estimates of the u_i's. We also examined the relative frequencies based on the manually matched cases but had too few observations to reliably estimate m and u in all categories.

Ten thousand realizations of multinomial random variables were then generated using the m_i's for matched pairs. The weight w_r was evaluated for each realization using the weight formula in Section 3. Similarly, 10,000 realizations for unmatched pairs were generated using the u_i probabilities and the weight w_r evaluated for each. The resulting weights were used to draw the curves labeled Simulation U and Simulation M in the figure. This method was also not entirely satisfactory. Estimates m and u based entirely on a truth set would be preferable, but could not be made because of the relatively small size of the manually matched set. In addition, there are likely to be multivariate dependencies among the rules that are not accounted for by the multinomial model.

Another evaluation uses a sample-based method described by Bartlett et al. (1993). This method involves selecting a small sample of events for linkage. Links are determined using both manual review and the AutoMatch method. Table 5 shows the result of the comparison on events from a random sample of about 500 persons. For these persons, a total of 37,943 record pairs were compared within persons. Regarding the manual matches as the "true" matches, the FN rate is estimated at about 7 percent (109 divided by 1,501), the estimate of the FP rate is less than 1 percent. Manual matching identified a slightly different set of pairs from AutoMatch. Note that when the weight chart method was applied to this same set of manually matched pairs (in Figure 1), we estimated a somewhat lower error rate of 5 percent.

Table 5. Sample-based error estimates

	Number of pairs	
	Manual	
Automatch	Matched	Non-match
Matched	1,392	336
Non-match	109	36,106
Total	1,501	36,442

6. Summary

This study uses a probability based linkage method to link the medical events in the 1996 MEPS. Over 35,500 medical events reported by patients in the HHS were matched to events reported by the associated provider in the MPS. The match rate is about 86 percent. The linked events provide medical expenditure data from both patients and their providers. These events will be used in a subsequent study to help handle

missing data and to adjust for household response errors when estimating medical expenditures in the United States.

How to best estimate the linkage error, given a limited budget and time schedule, is an open question. This study used several methods to determine the linkage error, including the weight chart approach (two sets of curves) and the sample-based approach. While these methods provided different estimates of the FP and FN rates, both the FP and the FN are generally estimated to be less than 5 percent.

7. References

AutoMatch. (1996). Generalized Record Linkage System, User's Manual. MatchWare Technologies, Inc.

Bartlett, S., Krewski, D., Wang, Y., and Zielinski, J.M. (1993). Evaluation of Error Rates in Large Scale Computerized Record Linkage Studies. *Survey Methodology*, **19**, 1, 3-12.

Fellegi, I.P., and Sunter, A.B. (1969). A Theory of Record Linkage. *Journal of the American Statistical Association*, **64**, 1183-1210.

Jaro, M.A. (1989). Advances in Record Linkage Methodology as Applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, **84**, 414-420.

Acknowledgment

The authors would like to thank Steve Cohen from AHCPR for his review and comments of this manuscript. The views expressed in this paper are those of the authors and no official endorsement by the DHHS or AHCPR is intended or should be inferred.

Figure 1