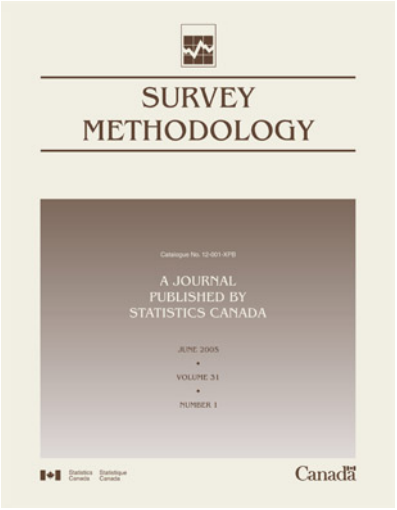
Catalogue no. 12-001-XIE

Survey Methodology

June 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2005

Published by authority of the Minister responsible for Statistics Canada

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 1, pp. 3-11
Statistics Canada, Catalogue No. 12-001

A Case Study in Record Linkage

M. Winglee, R. Valliant and F. Scheuren [1]

Abstract

Record linkage is a process of pairing records from two files and trying to select the pairs that belong to the same entity. The basic framework uses a match weight to measure the likelihood of a correct match and a decision rule to assign record pairs as "true" or "false" match pairs. Weight thresholds for selecting a record pair as matched or unmatched depend on the desired control over linkage errors. Current methods to determine the selection thresholds and estimate linkage errors can provide divergent results, depending on the type of linkage error and the approach to linkage. This paper presents a case study that uses existing linkage methods to link record pairs but a new simulation approach (SimRate) to help determine selection thresholds and estimate linkage errors. SimRate uses the observed distribution of data in matched and unmatched pairs to generate a large simulated set of record pairs, assigns a match weight to each pair based on specified match rules, and uses the weight curves of the simulated pairs for error estimation.

Key Words: File matching; Linkage error rates; Match weight; Selection threshold; Medical records.

1. Introduction

The basic record linkage framework by Newcombe Kennedy, Axford and James (1959) and Fellegi and Sunter (1969) uses a match weight to measure the likelihood of a correct match and a decision rule to classify record pairs. The optimal decision rule uses two match weight thresholds for selection (an upper threshold above which a link is treated as a match and a lower threshold below which a link is treated as a nonmatch). The choice of these thresholds depends on the acceptable pre-set linkage error rate and the requirement to minimize the number of links with indeterminate status between the two thresholds. Nowadays, practitioners of computerized linkage systems often use a single selection threshold to avoid manual intervention of the indeterminate links. Linkage decisions are typically made automatically after the system is "tuned" to achieve pre-set error levels. The challenge is that current methods to determine the selection threshold and to estimate linkage errors can produce divergent results depending on the type of linkage error, the choice of comparison space, and the estimation method.

This paper shares our experience with fellow practitioners who need a method to guide linkage selection and error estimation. Our case study used medical event files from the US Medical Expenditure Panel Survey (MEPS). MEPS collects medical expenditure data from both household respondents and their medical providers. The purpose is to combine the data from both sources for supporting annual estimations of medical utilization and expenditures (see Agency for Healthcare Research and Quality 2001 for more details on MEPS).

Here we discuss the linkage with three sets of annual medical event files – MEPS 1996, MEPS 1997, and MEPS 1998. Each set consisted of a household file containing events reported by household respondents for a given year and a medical provider file containing the corresponding events reported by medical providers of the household respondents. On average, approximately 50,000 medical events were reported for close to 10,000 persons, and around 15,000 person-provider units each year.

We used two model-based alternatives for linkage error estimation. One of these uses simulation to develop a distribution of the weights for various levels of agreement. This technique, called SimRate, begins by generating weight distributions for matched and unmatched record pairs. Using these, SimRate can then provide estimates of linkage error rates for different threshold levels. The error rates can then be used as a guide to action and a way to measure success. SimRate is contrasted with a second modeling approach created by Belin and Rubin (1995). As we hope to show, there is a role for both approaches; each has strengths as illustrated in the comparisons.

2. Mixture Models and Simrate Approaches

The mixture modeling method of linkage error estimation, as presented in Belin and Rubin (1995), has several attractive features. It is flexible in a sense that the weight creation process does not have to be considered directly. Hence, this method can be applicable to many different ways of creating weights. Once a model is specified, error

1. M. Winglee, Westat, Statistical Group, 1650 Research Boulevard, Rockville, MD 20850-3195, U.S.A.; R. Valliant, Joint Program for Survey Methodology, University of Maryland and University of Michigan; F. Scheuren, NORC, University of Chicago.

rates can be examined for a continuum of potential threshold values and confidence bands can be constructed to monitor the precision of error estimates (see section 7).

Mixture modeling does have limitations. While the method provides a particular kind of error rate – the proportion of linked records that are actually unmatched pairs, overall false positive and false negative error rates cannot be estimated since nonlinked pairs are not considered. The error rate that is estimated is conditional on the set of linked pairs of records. Furthermore model parameters may be hard to estimate if the weight distributions for the matched and unmatched sets are not separable (see Winkler 1994).

A key assumption in the Belin–Rubin approach is that it is possible to transform the distributions of the weights in the matched and unmatched sets to make them normal. Now a real difficulty exists here in that the transformed weights may be far from normal when the weight distribution for either the matched or unmatched sets is multimodal.

Another critical requirement is to have a training data set whose characteristics are very similar to those that are to be matched. Without a good training data set, the input parameter estimates for the mixture model may be poor, affecting the final estimated error rates obtained. Based on our application using annual medical event data repeated over three years, the parameters were not stable over time. This instability necessitated a training set for each year, making the Belin–Rubin approach impractical in our application because of the cost and time it required.

The simulation approach, SimRate, like mixture modeling, has the ability to examine different thresholds, allowing the user to monitor both the sensitivity and specificity of the decision rule for selecting linked pairs. As long as the process used to create match weights can be realistically modeled, customized methods of weight assignment like the one used in the current case study can be accommodated. The method does require the generation of pairs of records using the distribution of characteristics for the matched and unmatched sets. Some effort is needed to realistically generate the populations of pairs. In our work we have been successful with multinomial models for generating these populations.

3. Threshold Weight and Linkage Error Estimation

Several methods are available in the literature for selecting true matches and for estimating linkage errors (*e.g.*, Bartlett, Krewski, Wang and Zielinski 1993, Armstrong and Mayda 1993, Belin 1993, Belin and Rubin 1995 and Winkler 1992, 1995). See Fellegi (1997) for an overview of evolutions in record linkage, Tepping (1968) and Larsen and Rubin (2001) for other linking methods, and Scheuren (1983) for a capture-recapture method to estimate omission error.

Comparison of estimates from the different approaches is complicated by the fact that each approach tends to focus on different error components. In fact, the methods used in the linkage literature to construct linkage error rates are somewhat inconsistent. We illustrate this problem below.

Table 1 shows a 2×2 contingency table tabulating the numbers of true matched and unmatched pairs and declared linked and nonlinked pairs selected by linkage systems. Estimates of linkage error rates can be constructed relative to the true totals shown in the columns. An estimate of false positive linkage error rate under the Fellegi and Sunter framework is $\hat{\mu} = P(A_1 \mid U) = n_{12} / n_{\bullet 2}$ and that of false negative linkage error rate is $\hat{\lambda} = P(A_3 \mid M) = n_{21} / n_{\bullet 1}$ (see also Armstrong and Mayda 1993). These are the rates that SimRate is designed to estimate. They answer the question – "Of the set of true matched (or unmatched) pairs, what proportion is not correctly identified?"

Table 1
A Contingency Table for Evaluating Linkage Errors

	True set		
Declared set	Match (M)	Unmatch (U)	Declared total
Link (A_1)	n_{11} true positive	n_{12} false positive	$n_{1\bullet}$
Nonlink (A_3)	n_{21} false negative	n_{22} true negative	$n_{2\bullet}$
True total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

Some linkage evaluations have also considered rates relative to the declared totals in the rows. For instance, Gomatam, Carter, Ariet and Mitchell (2002) used $n_{12} / n_{1\bullet}$ and labeled it the positive predictive power of the linkage system. Others, however, have labeled this as the false match rate (Belin and Rubin 1995) or false positive declared rate (Bartlett *et al.* 1993). Rates constructed in this manner answer the question – "Of the declared linked (or nonlinked) pairs, what proportions are wrong?" Both questions are important in selecting matched pairs and should be addressed. That is one of the appeals in employing both SimRate and Belin–Rubin, if possible.

4. Simrate Weight Distribution Methods to Estimate Linkage Error

How to best estimate the linkage errors, given a limited budget and time schedule, is a difficult question. Accurate estimation of linkage errors should depend on at least two factors – the power of the identifying fields to unambiguously identify events that are true matches and the linkage method used. Taken together it is then possible, in a given setting, to specify linkage categories, estimate agreement probabilities, and determine match weights.

Following Newcombe and Kennedy (1962) and Jaro (1989), we adopt a weight distribution approach in our application that can take all these factors into consideration. The basic step is to first compute the match weight and order all possible configurations of agreement and disagreement outcomes of the comparison fields by match weight. Then we plot the cumulative distribution function of the weights for matched and unmatched pairs, and use the resulting weight chart to determine thresholds to attain desired levels of false positive and false negative error rates.

An ideal method to develop these curves might be to begin with a set of record pairs for which the truth is known. If resources are available, we could use a large set of true matched pairs, order them by match weight, and observe what proportion is above or below a given threshold. Similarly, we could take a large set of pairs, known to be true unmatched pairs, order them by weights, and again tabulate the proportion on either side of the threshold. The proportion of true matched pairs with weights below the threshold and the proportion of true unmatched 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.

One method to approximate this "ideal" approach (see also Bartlett *et al*. 1993) is to sample record pairs and use manual review to determine the true match status. Once the true pairs are known, we can attach the match weights from whatever linkage system is being used and then develop cumulative weight distributions, as discussed above. This method is, of course, subject to the well-known time and other resource limitations of manual review and is seldom practical with a large sample.

An alternative method is to generate the cumulative weight distributions through simulation. That is the heart of the SimRate approach. To explain in some detail, denote a record pair by r and a comparison field by v ($v = 1, \ldots, V$ fields). The comparison outcome situations in our application included partial agreements and multiple outcome categories beyond the basic agreement and disagreement categories (see also Newcombe 1988). Therefore, we denote that each field v has $i = 1, \ldots, c_v$ outcome categories. The outcome indicator is $\mathbf{y}_{rv} = (y_{rv1}, \ldots, y_{rvc_v})$, a vector of indicators showing the category into which pair r falls. One of the values of y_{rvi} will be 1 and the others 0 for each field.

The particular theory supporting the SimRate approach is to assume that $\mathbf{y}_{rv}$, has one multinomial distribution if pair r is a matched pair and a different multinomial distribution if it is an unmatched pair. We can then model the $\mathbf{y}_{rv}$ vectors as having a multinomial distribution with parameters $\mathbf{m}_v = (m_{v1}, \ldots, m_{vc_v})$ if the pair is a matched pair and parameters $\mathbf{u}_v = (u_{v1}, \ldots, u_{vc_v})$ if the pair is an unmatched pair. Then the probability $m_{vi} = P$ (field v category i agrees in pair $r | r \in M$) is the conditional probability of agreement for field v category I, given that the record pair r is in the set M of true matched pairs. In contrast, the probability $u_{vi} = P$ (field v category i agrees in pair $r | r \in U$) is the conditional probability of agreement for field v category I, given that the record pair r is in the set U of true unmatched pairs. Assuming independence of the matching variables, $v = 1, \ldots, V$, we can specify the joint probability of $\mathbf{y}_r = (y_{r1}, \ldots, y_{rV})$ if a pair r is a match, as

$$\mathrm{P}(\mathbf{y}_r | r \in M) = \prod_{v=1}^{V} \prod_{i=1}^{c_v} m_{vi}^{y_{rvi}}.$$

The corresponding probability of the same configuration of data, if the pair is really an unmatched pair, is

$$\mathrm{P}(\mathbf{y}_r | r \in U) = \prod_{v=1}^{V} \prod_{i=1}^{c_v} u_{vi}^{y_{rvi}}.$$

SimRate uses Monte Carlo simulation methods to generate a large number of realizations of matched pairs and unmatched pairs using estimates of the probabilities m_{vi} and u_{vi}. For each simulated pair, a match weight w_r, which applies to a given configuration of data, is calculated. For a given realization $\mathbf{y}_r$, a weight w_r is computed for the pair by summing the weights for the randomly generated categories that the pair fell into. The match weight w_r of a record pair is typically estimated as

$$w_r = \log_2 \left[\frac{\prod_{v=1}^{V} \prod_{i=1}^{c_v} m_{vi}^{y_{rvi}}}{\prod_{v=1}^{V} \prod_{i=1}^{c_v} u_{vi}^{y_{rvi}}} \right].$$

See section 6 on the match weights used in our simulation.

The cumulative distribution of these weights for the simulated matched pairs is then plotted as "Sim–*M*". Similarly, the reverse cumulative distribution for the unmatched pairs is plotted to generate "Sim–*U*" (see Figure 1, section 8, for an example of the simulation curves used in this study). The simulated proportion of matched pairs whose weights are below the cutoff is the estimate of the false negative error rate. The simulation proportion of unmatched pairs whose weights are above the cutoff is the estimate of the false positive error rate.

This approach requires that empirical estimates be made of the distributions among the matching variables of both true matched and true unmatched pairs. Even though the weight algorithm may involve the assumption of independence among matching variables, the actual data may show dependence. As long as artificial pairs can be generated that realistically follow the observed distribution of the data (incorporating any dependencies), then this method should provide suitable error rate estimates.

In our case study, we modeled data fields as having independent multinomial distributions, but this may not be reasonable in other applications. The SimRate concept can apply to any algorithm where weights and a cutoff point are used for classification. Thus, methods other than Fellegi and Sunter (1969), like Belin and Rubin (1995), might also be evaluated in this way. If methods are needed to deal with dependent categorical variables, the multivariate multinomial distributions in Johnson, Kotz, and Balakrishnan (1997, Chapter 26) may be appropriate. However, in applications similar to ours, the simplest procedure for accounting for dependence is to form cross-classifications of the variables that are related and to estimate probabilities for each cell in a cross-table. For example, if two variables with c_1 and c_2 categories are associated, then we can estimate the joint probability, p_{ij}, for each cell in the $c_1 * c_2$ table and use those in the simulation. Sparse data will naturally limit the number of cells for which this is feasible. But in the presence of sparse data, the penalty for model failure must be small.

5. Record Linkage of MEPS Medical Events

Record linkage of MEPS medical events used five identifying fields: event dates (year, month, day, and day-of-week), medical condition codes, procedure codes, global-fee codes, and lengths (number of days) of hospital stay. These fields are described in more detail in Winglee, Valliant, Brick and Machlin (2000). A training sample from MEPS 1996 was employed to derive match rules and outcome categories and to estimate the probabilities of agreement for each category, allowing for partial agreement and value specific outcomes. The same match rules were repeated each year with minor adjustments of the matching parameters.

For the training set we used the linkage system Automatch (Matchware 1996) and the unique match algorithm to select linked pairs. In "unique" matching, a File A record is optimally linked to only one File B record (Jaro 1989). In addition, we used the many-to-many match algorithm to generate a random sample of nonlinked pairs to facilitate linkage error estimation. However, the methods for estimating error rates, described below, apply to any software that implements the linkage methods based on match weights. They are not specific to Automatch.

The tradeoff in determining the selection threshold for MEPS was between getting a high match rate and limiting mismatch linkage errors. A high threshold weight would minimize false positive (mismatch) errors at the expense of lowering the match rate and losing valuable data collected from medical providers. On the other hand, a low threshold would increase false positive error and may affect the allocation of expenditure data in a way that would require special analytic techniques to overcome and even then only with uncertainty. Since both data sources had reported on ostensibly the same medical events for the same persons over the same period, the strategy was to maintain a reasonably high match rate and to conduct a manual review of a limited number of questionable linked pairs after selection to assess the analytic impact of falsely accepting them. Based on this decision the average match rate for the annual MEPS medical records files was about 85 percent.

The 1996 MEPS training sample *M* curve, labeled the "Tra–*M*" curve, was generated by applying match weights to "true" matched pairs for a random sample of 500 persons in MEPS 1996. For these persons, the manual review files contained 2,507 events from household respondents and 2,804 events from medical providers. Knowledgeable data managers reviewed the events and selected 1,501 pairs. We considered these as the true matched pairs in this evaluation. The manually matched pairs were assigned the weights derived from our match specification to generate a cumulative distribution function.

The 1996 training sample *U* curve, labeled the "Tra–*U*" curve, was generated using a random sample of unmatched pairs. We used a simple random sampling with replacement method to select 500 events each from the matching files and employed a many-to-many match algorithm to generate all 250,000 possible event pairs. For these randomly selected sets of pairs, the chance of there being any correctly matched pairs is negligible; thus, the entire set was taken to consist of unmatched pairs. We applied the match weights from our matching specification and plotted the "Tra–*U*" curve equal to 1 minus the cumulative distribution of the weights of these pairs. Figure 1 in section 8 shows both the Tra–*M* and Tra–*U* curves for the 1996 MEPS. The curves shown in this figure were smoothed using a nonparametric lowess function (Chamber, Cleveland, Kleiner and Tukey 1983) in S–PLUS 2000 (1999).

6. Simrate Implementation in MEPS

The SimRate weight distribution method used Monte Carlo simulation methods to generate separate sets of 10,000 simulated matched and unmatched pairs for creating the weight curves. To generate the "Sim–*M*" weight distributions we estimated the probabilities m_{vi} from linked pairs assigned by a unique matching algorithm. We used the "tuned" linkage system to select matched pairs from the 1996 annual matching files and tabulated the observed frequencies for each outcome category for each of the five matching fields. The proportion of pairs that fell into category *i* of field *v* was then used as the estimate $\hat{m}_{vi}$ of the probability m_{vi}.

For the unmatched pairs and the "Sim–*U*" curve, the u_{vi} probabilities for unmatched pairs were estimated using the same sample of unmatched pairs used in creating the "Tra–*U*" curve. The difference is that we used these pairs to observe the relative frequencies for each outcome category for each of the five matching fields among unmatched pairs. The proportion of pairs that fell into category *i* of field *v* was then used as the estimate $\hat{u}_{vi}$ of the probability u_{vi}.

For a simulated matched pair, a realization of the multinomial random variable y_{rv} was generated for each match field. For example, a configuration like (agreement on event date, agreement on length of hospital stay, agreement on the array of condition codes, joint agreement by type of procedure, and value specific agreement for a global-fee indicator) was generated using the match probabilities $\hat{m}_{vi}$ for each outcome category. Similarly, for each unmatched pair, a realization was generated of a category for each of the five fields using the unmatched probabilities $\hat{u}_{vi}$ discussed above.

For a given realization $\mathbf{y}_r$, a weight w_r was computed for the pair by summing the weights for the randomly generated categories that the pair fell into. The actual weights used in our simulation were adjusted ones that we specified rather than ones defined directly by the matching software (see Winglee, *et al.* 2000). Thus, we are simulating the way in which matching would actually be implemented. To do this we calculated the match weight for both the matched and unmatched sets of 10,000 pairs and plotted the simulated match weight functions.

Table 2 shows examples of some the partial agreement categories used for matching event date and the estimates of $\hat{m}_{vi}, \hat{u}_{vi}$, and w_r used in SimRate simulation. We defined a total of 19 outcome categories for matching by event date, 9 categories for duration of hospital stay, 27 categories by medical procedures, and 3 categories each for medical conditions and global fee. For example, for the outcome category exact agreement on event date, the estimate of $\hat{m}_{vi}$ was 0.69, meaning that 69 percent of the linked pairs had exact agreement on event date. The estimate of $\hat{u}_{vi}$ for this outcome category was 0.003, showing that only 0.3 percent of the unlinked pair showed agreement on this field. The match weight for exact agreement on date of event was 8.52 and that for complete disagreement (difference of more than two weeks apart and on different day of week) was –6.64. (see Winglee, *et al.* 2000 for the match weights by match fields and outcome categories).

We selected the match fields that were approximately independent in this case study. For example, we found no functional association between the date of medical events and other match fields like medical condition and length of hospital stay. For fields such as the indicators for surgery, radiology, and laboratory procedures, we used chi-square tests and found some dependence between the concurrence of surgery and radiology. To handle this situation, we estimated the joint probabilities and specified match rules to treat these procedure flags as a single match field (see section 4 above). Hence, we could then apply the independent multinomial distribution for simulation.

Table 2
Estimates of Multinomial Probabilies for Matched Pairs ($\hat{m}_{vi}$) and Unmatched Pairs ($\hat{u}_{vi}$), and Match Weights (w_{vi}) for the Match Field Event Date

Match rule for Event Date	$\hat{m}_{vi}$	$\hat{u}_{vi}$	w_{vi}
Missing	0.031	0.046	0.00
Exact match	0.693	0.003	8.52
Off +/– 1 day	0.068	0.006	5.71
Off +/– 3 day	0.023	0.005	4.09
Off +/– 5 day	0.014	0.005	2.47
Off +/– 7 day	0.030	0.006	2.84
Match by day of week only	0.014	0.034	– 3.64
Disagree	0.003	0.547	– 6.64

Table 3 shows the results of linkage error estimates from SimRate and the training curves at the threshold weight of $w = 1$ for MEPS 1996, MEPS 1997, and MEPS 1998. SimRate was easy to repeat each year. Repeating the manual-based weight curves, however, depended in part on manual review and we had only one reliable training sample, that for 1996. Note that the linked pairs used in SimRate will naturally generate some percentage of false positives and false negatives, *i.e.*, some matched and unmatched pairs are incorrect. Thus, the $\hat{m}_{vi}$ probabilities computed in this way for the identified fields are subject to error. It would have been preferable to estimate the *m* probabilities from a "truth" set where we were confident that all matches were correct. However, the manually matched training sets we were able to produce were too small to yield stable estimates in all of the detailed match categories and manual selection is also imperfect. This difference may explain in part the slightly higher overall error rate estimates from SimRate than from the training sample weight curves.

Table 3
Weight Curve Methods to Estimate Linkage Error Rates at Threshold Weight 1, MEPS 1996 – 1998

Method	Error Rate	1996	1997	1998
SimRate simulation curves	False negative	5.2	6.5	5.8
	False positive	9.0	6.9	7.6
Training sample curves	False negative*	3.3	3.3	3.3
	False positive**	5.5	6.4	5.7

* Estimates from the 1996 Tra–*M* curve were used for all three years.

** Estimates from the 1996 Tra–*U* curve used samples of 500 records from each match file and a total of 250,000 unmatched pairs. The 1997 and 1998 estimates used different Tra–*U* curves employing samples of 1,000 records from each match file and a total of 1,000,000 unmatched pairs.

7. Mixture Model Implementation in MEPS

A mixture modeling approach by Belin and Rubin (1995) views the distribution of observed match weights from a computerized linkage system as a mixture of weights for true matches and false matches. In principle, the mixture model method has two attractive features suitable for MEPS. First, it can handle repeated applications efficiently. When global parameter estimates of the transformed parameters and the ratio of the variances of the two distributions are available, these estimates can be applied to similar data for estimation. Since the MEPS record linkage is done annually, global estimates derived from early training samples could conceivably be applied for linkage error estimation in later years when manual review samples were not available.

The second advantage is that the mixture model can draw from multiple sets of parameter estimates from different training samples and can reflect variations. This feature is especially appealing for MEPS because manual review is a complex process and not necessarily always accurate. Hence, an alternative is to view the computer system selection as the truth and use them to provide an alternative set of parameter estimates. This process can also be repeated using training samples from more than one year.

Our application of the Belin–Rubin approach used the same training samples from MEPS 1996 and a second training sample of the same size from 1997. Following Belin–Rubin's examples, we applied the mixture modeling method using manually identified true and false match pairs from a one-to-one matching system (note that such systems provide relatively few false match pairs for estimation). We computed model estimates for MEPS 1996 and MEPS 1997 assuming the manual selection to be the truth, and for testing the behavior of the model, we computed a second set of estimates assuming computer system selected match pairs to be the true pairs.

Implementation involved two procedures – the Box and Cox (1964) procedure for global parameter estimation and the Calibrate procedure (Belin and Rubin 1995) to fit a mixture model for error rate estimation. Before applying Box–Cox, the weights were rescaled between 1 and 1,000. The Box–Cox transformation discussed by Belin and Rubin (1995) was

$$\Psi(w_r) = \frac{w_r^{\gamma} - 1}{\gamma \overline{w}^{\gamma-1}}$$

where w_r is the match weight for pair r, $\overline{w}$ is the geometric mean of the w_r weight, and γ is a parameter that is dependent on whether the pair is in the matched or unmatched set.

For the mixture model procedure to be effective, the transformed weights should be approximately normally distributed. The untransformed weight distribution with our data showed bimodality and almost no overlap in match weight between matched and unmatched pairs (bimodality was also observed in Belin–Rubin 1995). For example, application of their transformation procedure to the 1996 MEPS system pairs resulted in parameter estimates of $\overline{w} = 585.7$ and $\gamma = 1.15$ for the true matched pairs and $\overline{w} = 113.1$ and $\gamma = 0.48$ for the false matched pairs. The transformed weights, however, showed relatively little improvement towards normality. Since the match weights are the log of a product, or the sum of logs, we might hope that the weights would be normally distributed if there were many components in the sum. However, we had only five fields to use for matching. The small number of fields may have accounted in part for the lack of normality with our transformed data.

Table 4 shows the results of applying the Belin–Rubin mixture model to MEPS 1996. This table shows the model estimated false match rates, the 95 percent confidence interval of the estimated rate, and the actual observed false match rate at the threshold weight of 1. Using the manual review pairs as the true matched pairs, the model estimate of the expected false match rate at the threshold of $w = 1$ was 9.1 percent, with a 95 percent confidence interval ranging between 6.0 and 12.2. The actual observed false match error rate, however, was 14.5 percent, which is higher than the upper 95 percent confidence bound. Note that these are rates of the form $n_{12} / n_{1\bullet}$ in Table 1. These are not the same rates estimated by SimRate and the weight curve approach.

Table 4
Mixture Model Linkage Error Estimates

MEPS 1996	Percentage false match error			
	Expected rate	Lower Bound*	Upper Bound*	Observed rate
Manual match	9.1	6.0	12.2	14.5
System match	0.9	0.6	1.2	0.0

* The lower and upper bounds are the 95 percent confidence interval of the expected error rate.

Since manual review may not always be accurate, an option, for the purpose of evaluation, is to treat the computer system linked pairs as the truth matched pairs, and use them for modeling. Under this assumption, the model estimate of the expected error rate is 0.9, and a 95 percent confidence interval between 0.6 and 1.2. The actual observed rate in this case, 0 percent, was a hypothetical outcome treating the computer-linked pairs as correct. Of course, in reality there will be some nonzero level of error so that the mixture model confidence interval is not necessarily wrong.

We generated global parameter estimates using both the training sample manual selections and system selections for

MEPS 1996 and MEPS 1997 and used them as four sets of inputs to provide global estimates for modeling linkage error for MEPS 1998. This should be possible because the data remained similar and record pairs were selected using the same match rules for all 3 years. A difference was that manual review was not conducted for MEPS 1998 and we could not use the Box–Cox procedure for global parameter estimation for 1998 (because there was no separate manual indicator for true and false pairs). For this application, we use a bootstrap method in the Belin and Rubin Calibrate procedure to draw from multiple parameter sets to reflect uncertainties in estimation. This application, however, did not converge after 150 iterations of the estimation procedure. We could only conclude that the global parameter estimates from earlier training samples failed to generalize and provide error rate estimates for repeated linkage applications.

8. Concluding Comments and Analytic Implications

The process of threshold selection and linkage error estimation is an iterative process involving repeated cycles of observation, estimation, and modeling. Our case study employed modeling approaches for estimating linkage errors and for monitoring the predictive power of the linkage system. Both methods provided valuable information for determining the linkage selection and for evaluating the quality of the declared matched pairs as we found in MEPS.

The weight curves approach of estimation has the appeal that one can choose a selection threshold to attain the acceptable linkage error level. For example, Figure 1 shows the training sample and the SimRate simulation weight curves based on the 1996 MEPS matching files. A vertical line is drawn at the selection threshold weight of $w = 1$; the error levels for 1996 MEPS (shown in Table 3) were then estimated by the cumulative percentage at threshold level. By sliding this threshold, one can aim to minimize the total linkage error by selecting a threshold at the crossing point of the *M* and *U* curves. In this case study, the optimal threshold suggested by both sets of weight curves is fairly consistent. We included a likelihood ratio scale in this figure to provide a rough likelihood interpretation of the match weight. For example, at the match weight of $w = 1$, the likelihood ratio score is 2. This means that for records with a match weight of $w = 1$ or above, the relative likelihood of being true pairs is at least 2 to 1.

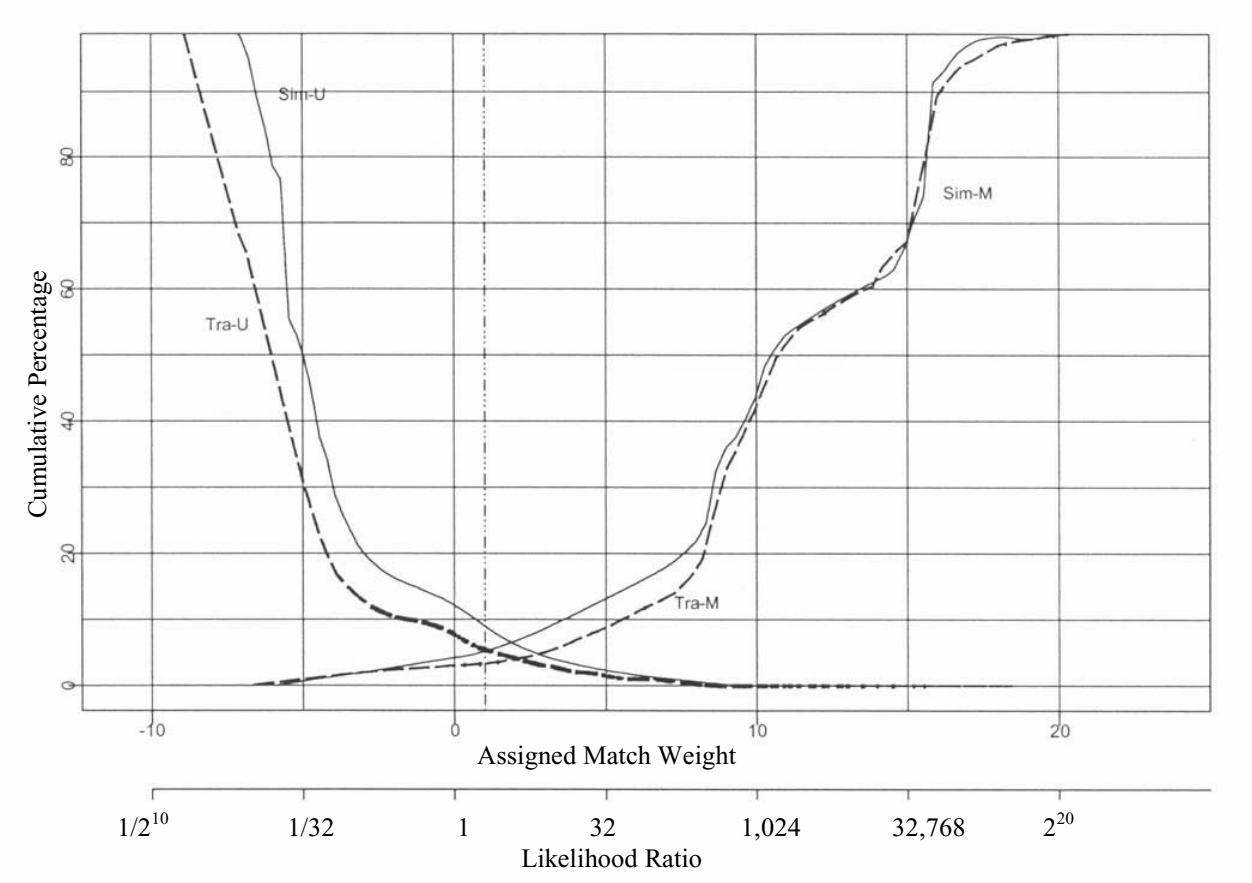


Figure 1. Weight Curves for MEPS 1996 using the SimRate and Training Sample Methods; the dashed vertical reference line shows the threshold value of 1.

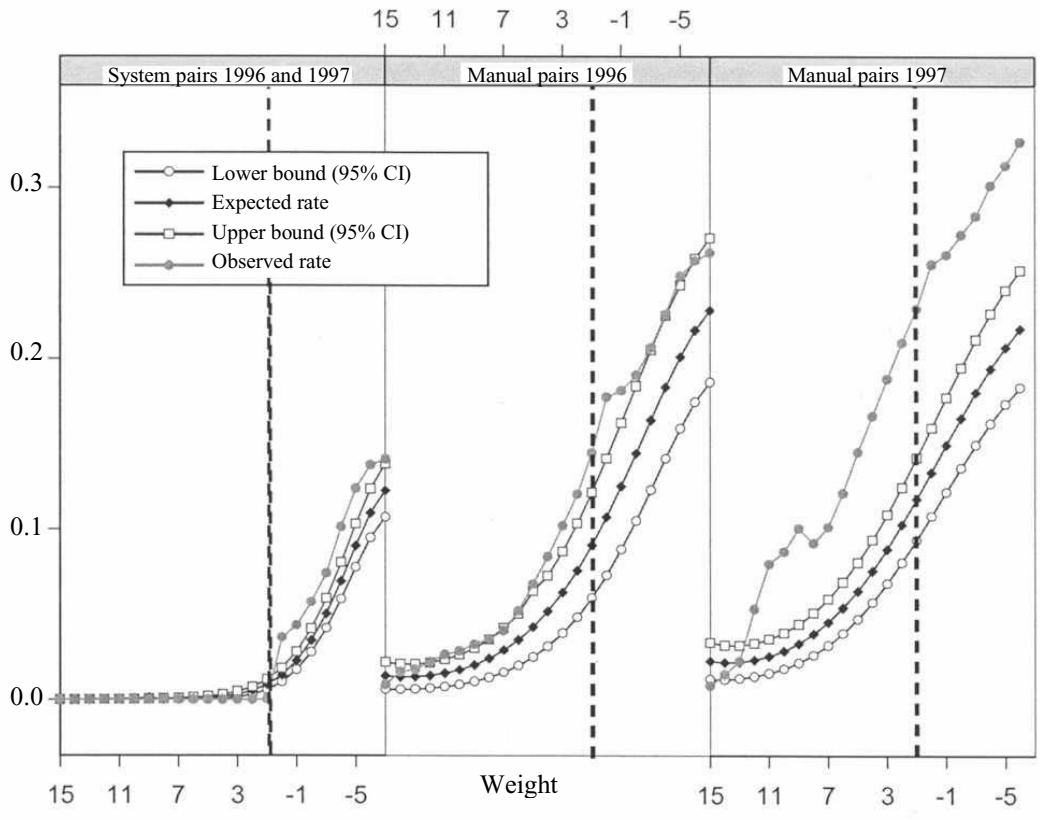


Figure 2. Mixture Model Estimates of False Match Rates by Weight, 1996 and 1997 MEPS Training Samples (a vertical line is drawn at weight = 1, which is threshold).

For linked pair quality, Figure 2 shows the distributions of false match rate estimates from mixture modeling. This figure shows the model estimated false match rate, the upper and lower 95 percent confidence bounds of the error rate estimates, and the actual observed rates. Panel 1 shows the estimates treating the computer system linked pairs as the true matched pairs. Panels 2 and 3 show the estimates from the 1996 MEPS and 1997 MEPS training samples. The difference between Panels 2 and 3 shows the inconsistency of manual selection by different reviewers in our application. In all three panels, the 95 percent confidence interval of the model estimates failed to cover the true observed values. Ideally, one would use both Figure 1 and Figure 2 together to guide the choice of selection thresholds.

We have found SimRate to be an informative and flexible tool for determining selection thresholds and estimating error rates in our application. Given multinomial or other models for the matching variables, the SimRate method provides error rate estimates that would be obtained from repeated application of the matching algorithm to a large number of candidate record pairs. It is also flexible in accommodating the choices of comparison sets of pairs for computing rates.

While our application achieved the matching and error rate estimation goals for MEPS, more work might be done prior to or during the analysis stage. Space does not permit us to develop these in the context of the current case study but two general approaches might be mentioned. First, it is possible to reweight the final results and adjust for false nonmatches – treating them in a manner analogous to unit nonresponse (*e.g.*, as in Oh and Scheuren 1980). To handle mismatches, the ideas in Scheuren and Winkler (1993 and 1997), and Lahiri and Larsen (2002) might be worth consulting. Whether these added steps are needed, of course, depends on the final uses to which the linked data will be put.

Acknowledgements

The basic linkage research, reported on here, was conducted under contracts 290–99–0002 and 290–94–2002 sponsored by the Agency for Healthcare Research and

Quality and the National Center for Health Statistics. The authors would like to thank Steven B. Cohen, Steven Machlin, and Joel Cohen of the Agency for Healthcare Research and Quality for their comments on various stages of this research and Thomas Belin for his suggestions on an earlier draft.

References

Agency for Healthcare Research and Quality (2001). MEP – Medical Expenditure Panel Survey. <http://www.ahrq.gov/data/mepsix.htm>.

Armstrong, J.B., and Mayda, J.E. (1993). Model-based estimation of record linkage error rates. *Survey Methodology*, 19, 137-147.

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, 3-12.

Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations (with discussions). *Journal of the Royal Statistical Society*, Series B, 26, 206-252.

Belin, T.R. (1993). Evaluation of sources of variation in record linkage through a factorial experiment. *Survey Methodology*, 19, 13-29.

Belin, T.R., and Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P. (1983). *Graphic Methods for Data Analysis*, Duxbury Press, Boston.

Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Fellegi, I.P. (1997). Record linkage and public policy – A Dynamic Evoluation. *Proceedings of the International Workshop and Exposition*, *Federal Committee on Statistical Methodology, Office of Management and Budget,* Washington, DC.

Gomatam, S., Carter, R., Ariet, A. and Mitchell, G. (2002). An empirical companion of record linkage procedures. *Statistics in Medicine,* 21, 1485-1496.

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.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: John Wiley & Sons, Inc.

Lahiri, P., and Larsen, M.D. (2002). Regression analyses with linked data. (Draft manuscript).

Larsen, M.D., and Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96, 32-41.

Matchware Technologies Inc. (1996). *AutoMatch: Generalized Record Linkage System User's Manual*. Silver Spring, MD: Matchware Technologies, Inc.

Newcombe, H.B. (1988). *Handbook of record linkage: Methods for health and statistical studies, administration, and business*. Oxford University Press, New York.

Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

Newcombe, H.B., and Kennedy, J.M. (1962). Record linkage: Making maximum use of the discriminating power of identifying information. *Communications of the Association for Computing Machinery*, 5, 563-567.

Oh, H.L., and Scheuren, F.(1980). Fiddling around with nonmatches and mismatches, *Studies from Interagency Data Linkages Series.* Social Security Administration, Report No. 11.

Scheuren, F. (1983). Design and estimation for large federal surveys using administrative records. *Proceeding of the Section on Survey Research Methods,* American Statistical Association, 377-381.

Scheuren, F., and Winkler, W.E. (1993). Regression analyses of data files that are computer matched. *Survey Methodology*, 19, 35-58.

Scheuren, F., and Winkler, W.E. (1997). Regression analyses of data files that are computer matched, II. *Survey Methodology*, 23, 157-165.

S-Plus 2000 (1999). MathSoft, Inc. Data Analysis Products Division, Seattle, Washington.

Tepping, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

Winglee, M., Valliant, R., Brick, J.M. and Machlin, S. (2000). Probability matching of medical events. *Journal of Economic and Social Measurement*, 26, 129-140.

Winkler, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 829-834.

Winkler, W.E. (1994). *Advanced Methods for Record Linkage*. Bureau of the Census Statistical Research Division, Statistical Research Report Series, RR 94/05.

Winkler, W.E. (1995). *Matching and record linkage*. In *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College and P.S. Kott). New York: John Wiley & Sons, Inc., 355-384.