THE URGE TO MERGE: 
A COMPUTATIONAL METHOD FOR LINKING DATASETS 
WITH NO UNIQUE IDENTIFIER

Robert M. Bell, RAND, Santa Monica, CA
Joan Keesey, RAND, Santa Monica, CA
Toni Richards, RAND, Santa Monica, CA

ABSTRACT

This paper describes a procedure used to link California vital statistics records for very low birthweight (VLBW) infants (weight less than 1500 grams) to Medicaid claims data. Because the two main data files share no unique identifier, record linkage required combining evidence across several identifying variables such as name, birth date, and delivery hospital. We used record linkage theory to compute scores that measure the likelihood of a true match, i.e., that two records correspond to the same delivery. These scores weight appropriately the various pieces of evidence for or against a match. The scores allow for missing data, data errors, and other factors that may cause a true match to disagree on some variables.

The resulting procedure is efficient, accurate, and noniterative. It eliminates the necessity of repeated merge steps and makes the most efficient use of the matching information, thereby minimizing the false positive rate associated with any number of presumed matches.

INTRODUCTION

This work is part of a larger project concerned with the survival and health costs of very low birthweight infants. These infants account for 1.5 percent of births, nearly 40 percent of infant deaths, and a disproportionate share of health care costs. Medicaid is the principal payor for their health care. Because no single data set provides information on costs and outcomes for a large sample of VLBW infants, linking vital statistics and Medicaid data is a crucial step in this research.

The ability to link records from existing data sets to form new data sets offers great benefits to researchers. Unfortunately, when two data sets do not share a unique identifier, record linkage can be a challenge. This paper describes a procedure used to link Medicaid claims data to California vital statistics records for very low birthweight (VLBW) live births and fetal deaths. The linkage involved about 53,000 of these infants (under 1,500 grams) in 1980 to 1987 and 1.46 million Medicaid claims for delivery/birth related hospital admissions during the same period. Consequently, it was essential to develop an efficient computer algorithm for discriminating between matches and non-matches.

Record linkage between the California vital statistics and Medicaid claims files is complicated by the need to use imperfect linking variables such as name, birthdate, and zip code of residence. None of the variables common to both files is a unique identifier (like Social Security number), and each is subject to error. We used record linkage theory to compute scores that measure the likelihood of a match, i.e., that two records correspond to the same delivery. The scores appropriately weight the various pieces of evidence for or against a match.

DATA

The data consist of two files. The vital statistics file includes one record for each live birth or fetal death in California for the period 1980 to 1987. The Medicaid claims file includes data for a subset of claims likely to involve child birth or fetal death. The claims were subset using an extensive list of approximately 145 ICD9 codes related to pregnancy, delivery and conditions of early infancy. Each file is a combination of data from several sources that were merged using unique identifiers to obtain information for each birth or claim.

Table 1 lists the potential linking variables available on the two files. While birth certificate records contain a fixed set of information about the mother and child, the items in the Medicaid file vary depending on the person for whom the claim was filed. We merged the Medicaid claims and Enrollment data in order to obtain additional beneficiary information. Medicaid deliveries may generate a claim for the mother only, one for the child only, or separate claims for each. Medicaid claims for the child are frequently filed under the mother's Medicaid number for the first month of life. Therefore, the information on the claim can pertain exclusively to the mother when the mother is both beneficiary and recipient of the services, for the child and the mother when the claim is for the child but filed with the mother's Medicaid number or exclusively for the child when the child is both recipient and beneficiary.

Table 1 POTENTIAL LINKING VARIABLES

<table>
<thead>
<tr>
<th>Vital Statistics File</th>
<th>Medicaid File (data source)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital</td>
<td>Hospital (C)</td>
</tr>
<tr>
<td>Last name-mother</td>
<td>Last name-beneficiary (E)</td>
</tr>
<tr>
<td>Last name-child</td>
<td>/</td>
</tr>
<tr>
<td>First name-child</td>
<td>First name-beneficiary (E)</td>
</tr>
<tr>
<td>Age of mother (years)</td>
<td>Birth date-mother &amp;/or child (C &amp; E)</td>
</tr>
<tr>
<td>Delivery date</td>
<td>Dates-admission/discharge(C)</td>
</tr>
<tr>
<td>Zip code-mother</td>
<td>Zip code-beneficiary (E)</td>
</tr>
</tbody>
</table>

Source of data in Medicaid file:
C = Medicaid Claims file
E = Medicaid Enrollment file
EVIDENCE REGARDING MATCHES

The actual record linkage procedure needs to combine evidence from several linking variables to determine whether a pair of records match. In addition, the procedure needs to account appropriately for the evidence provided by each linking variable and integrate the pieces of evidence, giving proper weight to each one. The procedure must address several issues.

We cannot require perfect agreement on all the linking variables without missing a substantial fraction of matches. A pair of matching records does not guarantee perfect agreement on all of the linking variables because there are errors and inconsistencies in the data.

It is inefficient to simply count the number of agreements because not all agreements provide the same amount of evidence for a match, nor do all disagreements between records provide the same amount of evidence that a match is incorrect.

For a given linking variable, the amount of evidence provided by an agreement varies depending upon the actual value of the variable. Agreement on an unusual name provides stronger evidence for a match than does agreement on a relatively common name.

ANOTHER RECORD LINKAGE METHOD

Record linkage is a common problem in health and other application areas. Using variables similar to ours, McCullough (1988) linked California vital statistics records to Medicaid claims for the year 1984. His procedure involved 97 iterations. The agreement criterion for one or more of the linking variables was relaxed with each successive iteration thus increasing the chance of false positives.

DERIVATION OF THE SCORE

Record linkage theory provides a method for dealing with each of the above issues. Our procedure follows the theory first developed by Newcombe et al. (1959) and Fellegi and Sunter (1969). Applications of this method to the U.S. Census are described by Jaro (1989). Although, in theory, x and y should be vectors containing values for all the linking variables, we will consider only one linking variable at a time. In practice, we compute a separate score for each linking variable and sum the scores to form a measure of the cumulative evidence.

Record linkage theory produces a score based on the conditional probability of a match given the values of the linking variable on the two records. The score, S, measures how much the log odds of the probability that the records match changes after observing the values of the linking variable. The score is

\[ S = \log \frac{P(x \text{ and } y \text{ given a match})}{P(x, y \text{ given linked at random})} \]

\[ = \log \frac{P(x,y \text{ given a match})}{P(x)P(y)} \]

since the probability of the two independent events occurring is the product of the probabilities for the two events alone.

Because the necessary probabilities cannot be estimated well for the joint distribution of all the linking variables, component scores for each linking variable, computed from empirical frequencies, are summed to produce a total score.

The formula produces several desirable properties for the scores: (1) more points are given for agreement on a specific variable, like birth date, than a vague one, like age; (2) more points are given for agreement on an unusual value, i.e., an unusual name like Westhafer would receive more points than a common name like Smith; and (3) there is a penalty for disagreement based on the frequency of disagreement among true matches. Ideally, all values above some threshold would indicate matches while lower values would indicate non-matches.

FORMULAS FOR A SINGLE COMPONENT SCORE

The values agree. The score is

\[ S = -\log[P(y)] = \log[1/P(y)] \]

where P(y) is the marginal probability of y. S is always positive because probabilities are always less than 1. As desired, the value of S is largest when P(y) is very small, i.e., when agreement is very unlikely to occur by chance.

The values disagree. The score is

\[ S = \log[P(\text{Values disagree, given a match})] \]

which is always negative. Among pairs that disagree, the score does not vary with the actual values of x and y. However, the value would vary among linking variables, with a more negative score occurring when there is a very small probability that a matching pair disagrees. Computation of the score requires estimating the probability that matching records disagree on the value of the linking variable.

The values are close. The score is

\[ S = -\log[P(\text{Value for a random record close to } x)] \]

\[ + \log[P(\text{Values are close, given a match})] \]

where P(\text{value for a random record close to } x) is the sum of P(z) for all z close to , but not equal to, x. The score consists of both a positive part (reward) for coming close and a negative part (penalty) for not agreeing exactly.

IMPLEMENTATION

The programs for the following work were written in SAS™ version 6.07 and executed on a Sun™ SPARCstation™ 2 using a UNIX™ operating system. Several interesting SAS techniques are demonstrated as part of the procedure description: (1) the use of PROC FORMAT for table look-up; (2) direct access of SAS data sets using the POINT option; (3) generation of probabilities using PROC FREQ.
Computing Rewards for Agreements

The Vital Statistics data for the study population of very low birthweight infants for 1980-1987 were used to compute probabilities and the corresponding rewards for agreement for the linking variables. The probabilities were determined using PROC FREQ as illustrated in Figure 1 for one of the study variables, the last name of the child. The output from this step is shown in Figure 2. The vital statistics data were appended with the probabilities, and the corresponding scores were computed; this was accomplished using SORT/MERGE techniques and is illustrated in Figure 3. The results are shown in Figure 4.

Fig. 1. Computing the Probability Last Name of Child

PROC FREQ DATA=VLBW;        VLBW VITAL STATISTICS (80-87)
TABLES CLST16 /OUT=OUT.LPROB NOPRINT MISSING;

Fig. 2. Probabilities - Last Name of Child

<table>
<thead>
<tr>
<th>OBS</th>
<th>CLST16</th>
<th>COUNT</th>
<th>PERCENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAGARDMORGAN</td>
<td>2</td>
<td>0.00350</td>
<td></td>
</tr>
<tr>
<td>AALPOELKING</td>
<td>1</td>
<td>0.00175</td>
<td></td>
</tr>
<tr>
<td>AARON</td>
<td>3</td>
<td>0.00525</td>
<td></td>
</tr>
<tr>
<td>AARONBENNETT</td>
<td>1</td>
<td>0.00175</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Appending the Probabilities and Computing the Score

PROC SORT DATA=VLBW&YR;       VLBW DATA FOR &YR;
BY CLST16;
DATA LVLBW&YR;                  VLBW DATA WITH PROB.;
MERGE VLBW&YR (IN=INVLBW)       KEEP=CLST16 PERCENT;
OUT=LPROB (KEEP=LPROB CLST16 PERCENT);   BY CLST16;
P_LST=PERCENT*0.01;            Probability;
S_LST=LOG(1/P_LST);           Score;
IF INVLBW THEN OUTPUT LVLBW&YR;

Fig. 4. Probability and Score- Last Name of Child

<table>
<thead>
<tr>
<th>OBS</th>
<th>CLST16</th>
<th>P_LST</th>
<th>S_LST</th>
</tr>
</thead>
<tbody>
<tr>
<td>MITCHELL</td>
<td>.0013269</td>
<td>6.6227</td>
<td></td>
</tr>
<tr>
<td>JACKSON</td>
<td>.0006221</td>
<td>5.6207</td>
<td></td>
</tr>
<tr>
<td>COBURN</td>
<td>.000175</td>
<td>10.9534</td>
<td></td>
</tr>
<tr>
<td>GRIFFIN</td>
<td>.0006074</td>
<td>7.5861</td>
<td></td>
</tr>
</tbody>
</table>

Penalties for Disagreement

When two records disagree on one of the linking variables, the score for that component should be negative to account for the evidence against a match. Computing the appropriate penalty requires knowing the probability that a match will produce various types of disagreements. To estimate these probabilities for a particular linking variable, we formed samples of apparent matches, based on information from other variables. The idea was to loosen one criterion at a time and analyze agreement on that criterion for cases where all the other linking variables provided strong evidence in favor of a match. We illustrate the procedure for mother's age. To study the agreement status among matches, we formed a sample of 1510 records in 1984 that agreed exactly on hospital, zip code, last name and exact birth date (or birth date within hospitalization period if birth date was not in the Medicaid file). Mother's age had the following distribution:

- Agreed: 92.5 percent
- Off by 1 year: 5.4 percent
- Off by 2 years: 0.5 percent
- Off by 3 years or more: 1.6 percent

Because disagreements of one year seem to be common, we reclassified those cases as "close." Those cases received both a penalty of 2.9 points (log(0.054)=2.9) and a reward that was slightly less than if the observed ages had agreed. The reward is smaller because there are two ways to disagree by one year, compared with only one way to agree exactly. The reward is reduced by 0.7 because the log(1/2)=−0.7. Disagreements of 2 or more years received a pure penalty of 3.8, (log(0.021)=−3.8).

Missing Data

The absence of the value for a linking variable in one file or the other reduces our ability to correctly link records to that observation. However, missing data should not affect the score for that component because the comparison does not provide any evidence for or against a match. Thus, when a variable was missing from either file, we assigned a score of zero—no reward and no penalty.

Accounting for Dependence among Linking Variables.

Because zip code relates very strongly to hospital, there is much overlap between the evidence provided by those two linking variables. To account for the dependence, we computed empirical frequencies of zip code within hospital. Use of these conditional probabilities eliminates the correlation between component scores for hospital and zip code. Thus, we avoid double counting information about area of residence.

Comparison of Vital Statistics and Medicaid Records

Claims and birth certificates were linked within hospital and year. Eliminating inter-hospital comparisons is equivalent to assigning an arbitrarily large penalty for disagreement on hospital. All possible record comparisons were made and a score computed for each. Each birth certificate was matched to every inpatient claim having the same provider. Despite the reduction in explicit comparisons the program compared about 11 million pairs and required about 1.5 hours of cpu time per year.

A table of the first and last observation numbers corresponding to a given inpatient provider identifier in the Medicaid claims file was constructed using PROC FORMAT. The creation of this table is illustrated in Figure 5 with the resulting output in Figure 6.
Application Development

Fig. 5. Table of the First and Last Observation Numbers Corresponding to a Given Provider

```
DATA MKFRMT (KEEP=START END LABEL TYPE FMTNAME);
    SET MCAIO.TOTlNP&YR;
    BY BCPROV;
    LENGTH BEGOBS ENDOBS $6 LABEL $12;
    RETAIN BEGOBS ENDOBS;
    IF FIRST.BCPROV THEN BEGOBS=.;
    IF LAST.BCPROV THEN DO;
        FMTNAME='"G&YR.BEG";
        TYPE::'C';
        START=BCPROV;
        END =BCPROV;
        ENDOBS=_N_;  
        LABEL=8EGOBS
    END;
    OUTPUT MKFRMT;
END;
```

- MEDICAID DATA
- PROVIDER ID

**Fig. 6. First and Last Observation Numbers Corresponding to a Given Provider**

<table>
<thead>
<tr>
<th>START</th>
<th>END</th>
<th>LABEL</th>
<th>VER 6.06 23.JUL92</th>
</tr>
</thead>
<tbody>
<tr>
<td>10001</td>
<td>1</td>
<td>1</td>
<td>63</td>
</tr>
<tr>
<td>10002</td>
<td>1</td>
<td>64</td>
<td>65</td>
</tr>
<tr>
<td>10003</td>
<td>1</td>
<td>66</td>
<td>2344</td>
</tr>
<tr>
<td>10005</td>
<td>1</td>
<td>2345</td>
<td>2393</td>
</tr>
<tr>
<td>10006</td>
<td>1</td>
<td>2364</td>
<td>3980</td>
</tr>
</tbody>
</table>

**Accessing the Medicaid Claims Data Using PROC FORMAT**

We used the table described in Figure 6 to determine the beginning and ending observation numbers corresponding to a given inpatient provider identifier. This is illustrated in Figure 7. The file of inpatient claims data was then accessed directly using the point option. This is illustrated in Figure 8.

**Fig. 7. Determining the First and Last Observation Numbers for a Given Provider Using PROC FORMAT**

```
DATA MATCH;
    SET OUT.VLBW&YR; BY H_CODE;
    RETAIN BEGOBS ENDOBS FLG_MATCH REF_ID;
    LENGTH BEGEND $12;
    DETERMINE FIRST.LAST REC. NO. IN MEDICAID INPATIENT FILE CORRESPONDING TO PROVIDER ID
```

**Fig. 8. Accessing the Claims Data Using the Point Option**

```
READ MEDICAID INPATIENT DATA AS DIRECT ACCESS FILE
READ OBSERVATIONS FOR SPECIFIED PROVIDER - H_CODE
```

**Fig. 9. Initializing the Match Flags**

```
M_BIRTH=0;
M_LOS=0;
M_LAST=0;
M_SUR=0;
M_TLST=0;
M_TSUR=0;
M_MAGE=0;
M_MAG=0;
M_ZIP=0;
M_FRST=0;
```

** Computing the Score **

Care had to be taken that extra points weren't assigned for redundant information, i.e., if the mother's surname and child's last name were the same, points were given only for the best match on the last name. Credit was given for the best match in each of five categories: (1) the birth date on the birth certificate matched either of the birth dates on the claim or fell within the admission/discharge period of the claim; (2) either the child's last name or mother's surname on the birth certificate matched the last name on the claim in its entirety, using the phonetic spelling, the first five characters or the first three characters; (3) the zip codes matched; (4) the mother's age from the birth certificate matched or differed by 1 from the age determined from either the claim or enrollment birth date; (5) the first names on the birth certificate and claim matched. Figures 9 and 10 show the code used to initialize and set the match flags.

**Fig. 9. Initializing the Match Flags**

```
M_BIRTH=0;
M_LOS=0;
M_LAST=0;
M_SUR=0;
M_TLST=0;
M_TSUR=0;
M_MAGE=0;
M_MAG=0;
M_ZIP=0;
M_FRST=0;
```
**Fig. 10. Setting the Match Flags**

```
************MATCH ON BIRTHDATE OR ADMISSION PERIOD************;
IF (BCBRTH=BDTH) THEN B_BIRTH=1;
ELSE IF (BDTH=ENRLBDTH) THEN B_BIRTH=1;
ELSE IF (SAS_ADM<=SAS_FCBR<=SAS_DIS) THEN M_LOS=1;

************CREDIT ONLY FOR BEST MATCH ON NAME ************;
IF (LAST=C_LAST) THEN M_LST=1;
ELSE IF (LAST=M_SURNAM) THEN M_SUR=1;
ELSE IF (LAST5=CLST5) THEN M_LST=1;
ELSE IF (LAST5=M_SUR5) THEN M_SUR=1;
ELSE IF (TLAST=LC_LAST) THEN M_LST=1;
ELSE IF (TLAST=DM_SUR) THEN M_SUR=1;
ELSE IF (LAST3=CLST3) THEN M_LST=1;
ELSE IF (LAST3=M_SUR3) THEN M_SUR=1;

IF (ZIP::ZIP_CODE) THEN M_ZIP=1;
IF (MC_AGE=M_AGE) THEN M_AGE=1;
THEN M_MAG=1;
IF (MC_AGE=M_AGE) THEN M_MAG=1;
THEN M_MAG1=1;

************COMPUTE PENALTIES FOR DISAGREEMENT************;
PENALTY=0;
IF (M_MAG1=1) THEN PENALTY=PENALTY+2.8; MOTHERS AGE OFF 1 YEAR;
IF (M_MAG=0 & M_MAG1=0) THEN PENALTY=PENALTY+3.8; DISAGREEMENT-MOTHERS AGE;
IF (M_MAG=1) THEN PENALTY=PENALTY+4.7; NAME DIFFER;
IF (M_MAG=1) THEN PENALTY=PENALTY+5.0; BIRTH/LOS;

************COMPUTE FINAL SCORE************;
TSCORE=SAS; S=SCOR - PENALTY;

The distribution of maximum scores, shown in Figure 13, is clearly bimodal with a trough in neighborhood of 16 points. This suggests that a cutoff around 16 would correctly classify at least 90 percent of the cases.

**Fig. 11. Computing the Score**

```
************COMPUTE THE SCORE ************;
ARRAY SCOR(14) S_BIRTH S_LST S_SUR S_ZIP S_MAGE
S_MAG1 S_LST5 S_SUR5 S_TLST S_TSR5 S_LST3 S_SUR3 S_FRST;
ARRAY MGRAPH(14) M_LOS M_BIRTH M_LST M_SUR M_ZIP M_MAGE
M_MAG1 M_LST5 M_SUR5 M_TLST M_TSR5 M_LST3 M_SUR3 M_FRST;
SCOR=S_HSP;
DO i=1 TO 14;
SCOR=SCOR+SCOR(i)*MTCH(i);
END;
```

A number of problems arise, however, in deciding which claim, if any, matches a particular record in the vital statistics file: (1) a value needs to be set for the cutoff; (2) two or more claims might link to the same VLBW birth; (3) two or more VLBW births might link to the same claim; (4) a claim that links to a VLBW birth may actually match a non-VLBW birth; (5) the score does not incorporate human judgment regarding whether the names really do agree.

In an effort to match as many very low birthweight infants as possible, we chose a cutoff of 10, well below the trough. We deleted clearly inferior links--multiple links to the same claim where one of the links was clearly

**WUSS 1993 Proceedings**
superior. We deleted linked pairs when a non-VLBW birth record provided a clearly superior match to the same claim. This step, which we call the "backmatch," searched for claims that really match non-VLBW births and therefore cannot match VLBW births. We then raised the cutoff to its final value, 14 points, and manually reviewed questionable cases that had not been resolved in the previous steps.

**HOW WELL DID THE SCORE DISCRIMINATE MATCHES FROM NONMATCHES?**

Figure 13 strongly suggests that the score discriminates well between matches and nonmatches except, perhaps, for a narrow gray area around 16 points. To investigate more fully, we used results from the backmatch to estimate the frequency with which a non-VLBW claim that would score 18 points or higher, as a function of the score for the original link. To avoid linking to non-VLBW twins, we restricted the sample to links for single births. If a link is correct, it is very unlikely that any non-VLBW birth record would produce a score exceeding 18.0 points.

The results are as expected (Figure 14). When a VLBW link had scored below 13 points, the odds favored finding a higher scoring non-VLBW birth record during the backmatch. However, that probability quickly declined as the score for the original link rose toward 16 points. For scores above 18 there was practically no chance of finding a convincing non-VLBW link. It is clear that the false positive rate for links scoring above 16 would be much lower than the rate for links scoring below 14. Even so, the figure illustrates the value of the backmatch in deleting false positives with a score between 14 and 18.

**CONCLUSION**

The score based on record linkage theory worked very effectively in our application. It immediately sorted over 96 percent of birth records into two groups: (1) those which almost certainly match the corresponding Medicaid claim (maximum score > 18) and (2) those for which a match seems very unlikely (maximum score < 14). The "backmatch" and other checks further reduced the number of cases falling in the grey zone.

This theory-based procedure offers several advantages over ad hoc iterative methods. It automatically weights different linking variables based on the relative evidence that they provide. It also takes into account the observed value of the linking variable when two records agree; iterative procedures are unlikely to do so. The procedure provides a basis for defining "close" and using that information appropriately. On the practical side, the score simplifies the entire exercise, potentially reducing computer time by an order of magnitude or more. In addition, record linkage theory can dramatically reduce the number of subjective decisions, speed up the process and improve the accuracy.

**REFERENCES**


SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries.

Sun is a trademark of Sun Microsystems, Inc.

UNIX is a registered trademark of UNIX System Laboratories, Inc.

SPARCstation is a trademark of SPARC International, Inc., licensed exclusively to Sun Microsystems, Inc.

™ indicates USA registration.

Other brand and product names are registered trademarks of their respective companies.