# STATISTICAL MATCHING TECHNIQUES FOR A HOUSEHOLD HEALTH INSURANCE SURVEY

Michael Sinclair, Frank Potter, and Barbara Lepidus Carlson
Mathematica Policy Research, P.O. Box 2393, Princeton, NJ 08543-2393

**Keywords: Statistical Matching, Imputation, Health Surveys, Community Tracking Study**

## 1. Introduction

The household component of the Community Tracking Study (CTS) conducted for the Center for Studying Health System Change (HSC) collected data on a variety of health care issues, including whether the members of a household were covered by private health insurance and, if so, the characteristics of the health plan. To obtain more detailed information on the household-reported private health plans and to validate the information from the household respondent, the Followback Survey collected data from the insuring entities (insurance carriers) on characteristics of specific health plan products that they offer. In general, these data were then linked to the household survey data to prepare a complete analytic data file consisting of both household- reported and entity-reported characteristics of the health plans. For some household sample members, a partial or "soft" linkage was established between an entity and the household members given we could not identify the specific product that they had. This paper describes the statistical matching procedures we used to resolve these linkages. Overall, the methods are estimated to have achieved an exact-match rate of 63.5 percent.

As a starting framework, the CTS Household Survey and the corresponding Followback survey sample are based on a complex 60-site clustered sample design. From the household survey, the respondents (about 60,000 people, 32,000 households) reported 22,211 private family-specific health care policies; we attempted to obtain policy attributes from the insuring entity on all policies. We define a policy as consisting of a unique relationship between a private health plan and the set of household members that the plan covers. We successfully linked 11,651 household-reported policies to entity-reported product data. For another 4,318 policies, we linked the policy to an entity but not to a specific product. In these cases, the policy was tentatively matched to two or more of the entity's products (that would be available to the person in their geographical area) and we chose one to be the final linkage based on the statistical matching procedures described in this paper. The remaining 6,242 policies (28.1 percent) could not be linked to any entity or product. To compensate for these policies without product data, we adjusted the survey weights for each person matched to the 15,969 policies using a propensity model adjustment procedure. We followed this by post-stratification to the original CTS population distribution to create the final analysis database.

## 2. Methodology

The solution to a record linkage problem depends on the data available for linking purposes. In the basic setup, a primary set of data (denoted as file A) must be linked to another set of data (denoted as file B). In our case, file A contains the data from the household survey interviews, and file B contains the health plan product information from the entity interviews. In record linkage problems, a set of variables common to each file, such as a person's name, address, or other unique identifying information, traditionally facilitates the linkage process. Therefore, a researcher can simply develop an algorithm to compare the data in these common fields across the two files and then, based on the degree of similarity in the fields, select a final match.

In our case, we had few, if any, common data items to facilitate the matching process. Although both the household respondent and entity reported on five basic characteristics of the health plan, prior analysis among the successful matches showed that the consistency between these items was low (Cunningham et al., 2001). As a result, we could not rely completely on these variables to select the best linkage. Therefore, we adopted a modeling-based procedure suggested by Singh et al. (1983).

This method uses an auxiliary data file of known linkages to develop a statistical matching procedure for assigning linkages to another set of files. Because our successfully linked cases provide the appropriate data for modeling the linkage process, this procedure was ideally suited to our needs. The approach is conducted in four developmental stages. First, a key set of attributes from file B (the product file) is selected that appears to most accurately describe the differences among the records on that file. Second, from the auxiliary file, which contains a set of file A (the household or family characteristics) and file B linked data, a series of standard regression or logistic regression models are developed to predict each of the selected items on file B from the items on file A. Third, the model coefficients are used to obtain predicted values for the selected file B items for each of the unlinked file A records. Fourth, the predicted values for each file A record are compared

with the values on the file B records. The file B record with the closest set of values is selected as the final link. The data from the linked B record replaces all of the missing product information on the file A record. In our case, the matching task is heavily "blocked" in that the possible set of matches for each soft-matched A file record is limited to those the entity offered in the site.

Following the steps above, we began by selecting a set of the product attributes that appeared to have the greatest discriminatory power among the four self-reported product types (Health Maintenance Organization, (HMO), Point-of-Service (POS), Preferred Provider Organization (PPO), and Fee-for-service or indemnity (FFS)). We describe this selection process in Section 3. We used the linked data to develop a logistic regression model for each product attribute to predict the attribute from the household-reported information. We discuss the results of these modeling procedures in Section 4. We used the resulting models to obtain predicted values for the attributes for the soft-matched policies. We compared the predicted values with the actual values to select one of the products as the final link. Finally, as a refinement and validation step, we prepared two "mock" soft-matched sets of records from the successfully matched cases to simulate the matching process. The linkage procedure and the preparation of the mock files are discussed in Sections 5 and 6, respectively.

## 3. Selection of Product Attributes to Use in the Matching Process

The accuracy of Singh's matching procedure relies on two assumptions: (1) the variables selected from file B to serve as the primary matching variable fully capture the differences among these records, and (2) the primary matching variables can be predicted accurately from the file A information. We therefore selected a set of the product variables that would best meet these criteria.

We selected nine product attributes to use in the statistical matching process. Since we had asked the entities to classify each product into four categories (HMO, POS, PPO and FFS), we decided to use this outcome as a primary matching variable and then to see which of the product attributes best predicted this classification scheme. We conducted cross-tabulations between each questionnaire item and the self-reported product type. The attributes that had differential response patterns across the product types became candidates for the matching variables. To support these findings, we also conducted a series of stepwise discriminate analysis procedures to identify the entity reported variables that together best predicted the entity self-reported product type. Listed below is a general summary of the product attributes collected during the survey, which we evaluated in this analysis.

1. B6, Whether Plan Covers Out-of-Network Physicians
2. B10, Whether Plan Requires a Primary-Care Physician (PCP)
3. B8, Whether a Member Can Self-Refer to In-Network Specialists.
4. B13, Whether the Plan Required a Coinsurance or Copayment for Routine Physician Visits, and If So, How Much.
5. B14, the Level of the Plan Deductible
6. C4, Payment Method Used for PCPs
7. Whether the Plan is Associated With a Network of Physicians (based on two questionnaire items)

We conducted the discriminant analysis procedure on the 4,663 product interviews, using the self-reported product type (four categories) as the dependent variable. For this task, we converted the continuous data items associated with the level of coinsurance or copayment (item B13) and the deductible (item B14) into series of range indicators. We also transformed the response categories on the payment methods for PCPs (item C4) into three categorical indicators. We conducted an initial model using range indicators for the level of coinsurance or copayment. A second model identified only whether a coinsurance percentage or a copayment amount was required by the plan.

The results suggested that household survey items B6, B10, the C4 categories, the values in B13 and B8 described the entity self-reported product type. We therefore included items B6, B8, B10, and C4 among the list of matching variables. Comparing the models also indicated that the coinsurance/copayment status, B13, seemed to capture the majority of the explanatory power in the coinsurance and copayment levels. Therefore, to limit the matching variables to a manageable number, we used only the coinsurance/copayment status variable. Although the variable for deductible level (item B14) showed some predictive ability, we could not accurately predict the deductible level from the CTS household information, so we excluded it from the matching list.

As a final step in preparing our list of matching variables, we examined the constructed variable that indicated whether the product had a network. This variable had been coded from the self-reported product type and whether a list or directory of physicians was associated with product. Therefore, did not include it in the stepwise discriminate

analysis procedures, because it would have distorted the results for other variables. We did include network status among the matching variables, because the item is a direct by-product of the self-reported product type.

## 4. Modeling the Product Attributes on the Successfully Linked Cases

We prepared a series of weighted logistic regression models to predict each of the nine matching product variables using the household survey policy-level variables for the successfully linked policies. We could then use these models to obtain predicted values for the matching variables on soft-matched policies. As the first step in the modeling process, we prepared a set of weights to apply to the successfully linked data during the modeling process. These weights consisted of the product of a policy weight, which was equal to the sum of the person survey weights for the people who were members of the plan, as well as an adjustment factor to account for any differences in the profiles of the soft-matched and successfully linked records. Since the relationships between the matching variables and the household survey policy-level variables could be different between the successfully linked and soft-matched policies, we used these adjustments in combination with the survey weights to compensate for the fact that only the successfully linked policy data was used in preparing these models.

Review of the successfully linked and soft-matched data showed that the rate of soft matches varied by site. Furthermore, the percentage of policies reported to an HMO in the household survey varied between the soft and successfully linked cases. The other demographic and socioeconomic factors did not seem to explain the variation in the proportion of the successfully linked cases. We computed a nonresponse adjustment to the survey weights for the successful linkages based on 120 weighting cells defined by site of residence (60 sites), in combination with reported HMO membership.

As the second major step, we reduced the set of household survey policy variables to those that appeared to have some influence on the entity self-reported product type. We generated frequency distributions for each policy variable by the self-reported product type and eliminated variables from the list that showed similar patterns among all four product categories.

We computed a weighted logistic regression model for each matching attribute using the household policy variables to predict the outcome of each product attribute. These models were developed using a combination of stepwise and nonstepwise procedures, setting a significance level for the model selection process liberally at 0.15 to be sure all of the potential predictors were included in the model. In most cases, we started with a full model containing all predictors. After reviewing the output from the full-model and the stepwise procedures, we eliminated variables that were not significant based on a chi-square test of significance**.** Table 1 presents the key predictors (based on the Chi-square test of significance) in each model and the model r-square values.

## 5. Selecting the Linkages Among the Soft Matches

For each of the 4,318 soft-matched cases, the data collection and editing process appended from two to nine potential products to each policy to yield 11,040 potential soft-matched products. Table 2 shows a frequency distribution of the number of potential products linked to each policy. The majority (63.8 percent) of the policies had only two choices. We selected one of the soft-matched products as the final product for a policy.

We computed predicted values for each of the nine product attributes for the 4,318 soft-matched policies, using the coefficients from the model. We then compared the predicted values of the nine attributes with the actual values among the linked products. For each possible link, we computed the absolute difference between predicted and actual value. This computation produced nine "gap" measures for each potential product link. Since the predicted value was the estimated probability of having the trait, the gap measures had the form of either (a) the absolute difference between a value of zero (not having the trait) and the predicted probability, or (b) the absolute difference between a value of one (having the trait) and the predicted probability.

To test various matching procedures and to estimate the accuracy of the process, we created a "mock" or simulated file of the soft matches that were based only on the successfully linked data for which the correct linkage was known. To select the linkages we prepared a logistic regression model that predicted the probability of a match on the basis of the nine absolute gap measures. We then used the coefficients from this model to compute the probability of a match for each soft-matched product and selected the product that had the highest estimated probability of a match as the final link. For 75 of the 4,318 soft-matched policies, the predicted probability of a match was the same for two or more of the choices with the highest probabilities of a

match. For these cases, we selected one of the products at random.

## 6. Validation of the Techniques

Our objective was to prepare a mock file(s) containing known matches and fabricated nonmatches to best design the matching procedures and to estimate the accuracy in the final approach selected. We wanted the mock file to mimic the distribution of choice patterns on the file of potential soft matches. In particular, we wanted this mock file to meet two objectives: (1) to have the same distribution of the number of choices for each policy, and (2) to have the same distribution of entity-reported product-type combinations. We also created a second mock file that simply represented the mix of known linkages and fabricated nonmatches prior to receiving the adjustments required to make the file mimic the properties of the actual soft-matched file. This second mock file is referred to as the initial mock file as the final mock file was created from it after a series of adjustment steps.

We developed a set of nonmatches based on the same process that generated the soft-matched choices. A choice of products is available for each soft-matched policy because entities offered multiple products in the sites. We therefore were able to generate a similar set of choices for each successfully linked policy by creating, for each, a list of the products the entity offered at the site. We identified one or more additional product offerings for 10,058 of the 11,651 successfully linked policies, creating 36,694 potential links.

The 10,058 successfully linked policies contained a higher proportion of self-reported HMO and POS plans than did the full set of 11,651 successfully linked policies. Because they represented a slightly skewed set of the successfully linked cases, we selected a sample of the HMO and POS policies to remove from the list. This step created a final set of policies that had the same proportion of policies in each of the four product types as in the original set of 11,651. After the reduction, the mock file contained 8,941 successfully linked policies with 32,616 potential (and actual) links. This file became the initial mock file and in essence reflected what would have been contained in soft-matched file for the successfully linked cases had such a file been created.

To meet the first criteria in preparing the final mock file, we compared the distribution of the number of choices on the soft-matched cases with the distribution in the mock file containing 32,616 linkages. The initial mock file contained a substantially larger proportion of policies with three or more choices than did the soft-matched file. To correct this disparity, we used a combination of two sampling procedures on the initial mock file: (1) deleting a random selection of policies and all the linkages associated with these, and (2) deleting one or more potential product links from each policy.[1] After this step, the revised mock file contained 6,068 policies representing a total of 15,425 choices.

To achieve the second objective for the final mock file, we prepared a weighting-class-based weight to correct for differences in the choice patterns between the revised mock file and the actual soft-matched file. To compute the weights, we tabulated the proportion of cases on the actual file and on the mock file with a given number of choices that had a particular set of choice combinations (for example, one each of HMO, POS, and PPO). We used these two values to compute a weight equal to the ratio of the proportion in the soft-matched file divided by the proportion in the mock file.

We evaluated three matching methods on both mock files (initial and final) to examine a total of six matching procedures. We decided to use both mock files given the extensive level of transformations conducted to prepare the final file. In the first of the three matching techniques, referred to as a scoring method, we computed a score for each product choice on the basis of the weighted average of the absolute gaps. We designed the "gap" weights to represent the relative ability of each gap measure to identify the correct linkage. To measure this ability, we used a logistic regression analysis to model the actual match status as function of the gap measure. From the analysis, we obtained the Wald chi-square test statistics for testing the influence of each gap measure on the prediction. We could then normalize these values to sum to one to reflect the relative contribution of each gap measure in identifying a correct match. Finally, we selected the policy with the smallest score value as the link.

The second and third techniques were also based on the same logistic regression model but used the model coefficients directly. For the second approach, we applied the unstandardized model coefficients to the gaps, taking into account the exponential structure of the model, to provide an estimated probability of a match. We then selected the product with the highest estimated probability.

[1]We could have deleted a sufficient number of choices from each policy to meet the distributional requirements, but we believed that deleting several choices from some policies would distort the pattern of choices.

The third approach used the standardized coefficients instead.

Overall, the results for the different matching methods were similar across the two mock files. However, among the cases in which the correct link was an FFS policy, the methods based on the initial successfully linked mock file produced an average gain of about four percent in the percentage of cases assigned correctly. The predicted probability method using the unstandardized coefficients assigned produced a slightly higher rate of correct linkages across product types. Consequently, for our final strategy, we used the unstandardized coefficients from the initial successfully linked data model to predict a probability of a match.

Table 3 presents the accuracy rates for the selected statistical matching procedure as measured from the final mock file. Different estimates of accuracy rates are produced for two grouping of the records: (1) the entire file, which reflects the overall rates; and (2) as classified by the four entity self-reported product-type categories (based on the product category associated with the correct linkage) which reflects the accuracy rate within each product type. Table 3 also shows two types of accuracy rate estimates. The first rate indicates the percentage of cases in which the correct linking record was selected among the choices (referred to as an exact match). The second indicates the percentage of cases in which the selected choice was of the same product type as the correct link. For each type of accuracy rate, we also computed the corresponding percentage of improvement in the link rate relative to a random pick based on the average number of choices (AC) in each group (equal to the [match rate – 1/AC] divided by [1-1/AC]). The results show that the statistical linking procedures obtained an overall exact match rate of 63.5 percent, and a 67.3 percent match rate with a product of the same type. These rates reflect respective percentage improvements of 49.8, and 55.0, percent relative to a random selection methodology. Within each product type, the HMO products had the highest exact match rate (71.7 percent), and FFS products had the lowest rate (42.3 percent).

In summary, while the individual steps/components in this process had moderate predictive power, the combined process produced a reasonable level of accuracy. For future studies, we are identifying the obstacles in establishing these linkages and attempting to find solutions for improving the match rate so that the procedures described herein are applied to a smaller proportion of the analytical records. In particular, we are examining the use of an extensive employer-based data collection effort similar to that conducted for the Medical Expenditure Panel Survey (MEPS) conducted by the Agency for Healthcare Research and Quality.

**References:**

Singh, A.C., Mantel, H.J., Kinack, M.D., and Rowe, G. (1993), "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption" Survey Methodology, 19, 59-79.

Cunningham, P., Denk, C., and Sinclair, M.(2001), "Do Consumers Know How Their Health Plan Works?" Health Affairs, March/April 2001, 20:2, 159-166.

**TABLE 1. LOGISTIC REGRESSION RESULTS TO PREDICT EACH PRODUCT ATTRIBUTE FROM THE HOUSEHOLD SURVEY POLICY DATA**

(R-square values presented on first row of table)

| HMO | B6 Covers Out-of-Network | B8 Covers Spec. Without Referral | B10 Plan Requires PCP | B13 Co-insurance | C4_1 Payment FFS | C4_2 Payment Disc FFS | C4_4 Payment Capitation | NET |
|---|---|---|---|---|---|---|---|---|
| 0.2969 | 0.1652 | 0.0897 | 0.3051 | 0.1439 | 0.1134 | 0.1015 | 0.2169 | 0.1240 |
| Is plan an HMO? | Is plan an HMO? | Is plan an HMO? | Sign-up with PCP required | Is plan an HMO? | Sign-up with PCP required | Site Area | Sign-up with PCP required | Sign-up with PCP required |
| Pay for specialist w/o referral | Pay for specialist w/o referral | Pay for specialist w/o referral | Site Area | Plan has a network | Pay for specialist w/o referral | Is plan an HMO | Site Area | Plan has a network |
| Need referral to see specialist | Usual Place of Care HMO | Usual Place of Care HMO | Need referral to see specialist | Sign-up with PCP required | Site Area | Need referral to see specialist | Need referral to see specialist | Employer Type |
| Usual Place of Care HMO | Sign-up with PCP required | Site Area | Is plan an HMO? | Site Area | Usual Place of Care HMO | Usual Place of Care HMO | Is plan an HMO? | Income Level |
| Site Area | Need referral to see specialist | Income Level | Pay for specialist w/o referral | Usual Place of Care HMO | Employer Type | Need referral to see specialist | Pay for specialist w/o referral | Usual Place of Care is HMO |
| Sign-up with PCP required | Site Area | Usual Place of care is doctor's office | Education Level | Pay for specialist w/o referral | Gender | Income Level | Education | Pay for specialist w/o referral |
| Race | Usual place of care is hospital | Sign-up with PCP required | Years in HMO | Need referral to see specialist | Multiple Coverage | Employer Type | Plan has a network | Multiple Coverage |
| Multiple Coverage | Employer sponsored plan | Employer size is 250-499 | Previously had private insurance | Income Level | Education Level | Race | Age | Site Area |

**TABLE 2. DISTRIBUTION OF THE NUMBER OF POTENTIAL LINKS ASSOCIATED WITH EACH SOFT-MATCHED POLICY**

| Number of Potential Links | Records (Number) | Cases (Number) |
|---|---|---|
| 2 | 5,510 | 2,755 |
| 3 | 2,946 | 982 |
| 4 | 1,532 | 383 |
| 5 or more | 1,052 | 198 |
| | 11,040 | 4,318 |

**TABLE 3. ESTIMATED ACCURACY RATES IN THE STATISTICAL LINKING PROCEDURES**

| Self-Reported Product Type | Average Number of Choices | Exact Match | | Same Type | |
|---|---|---|---|---|---|
| | | Link Rate | Percentage Greater than Random | Linkage Rate | Percentage Greater than Random |
| ALL | 3.67 | 0.64 | 49.81 | 0.67 | 55.02 |
| HMO | 3.60 | 0.72 | 60.84 | 0.76 | 66.13 |
| POS | 3.64 | 0.52 | 34.40 | 0.55 | 38.21 |
| PPO | 3.89 | 0.67 | 54.86 | 0.71 | 61.55 |
| FFS | 3.38 | 0.42 | 18.07 | 0.44 | 20.77 |