Imputation Using the Other Pair Member

Peter Frechtel [1], Victoria Scott [1], Amy Couzens [1]
Andrew Moore [1], Jonaki Bose [2]
[1]RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709
[2] Substance Abuse and Mental Health Services Administration, 1 Choke Cherry Road, Rockville, MD 20857

Abstract
The National Survey on Drug Use and Health (NSDUH) provides national, state, and substate data on substance use and mental health in the civilian, non-institutionalized population age 12 and older. In the NSDUH, zero, one, or two people are selected from each selected household. "Pairs" account for about 60% of the annual sample, over 90% of which are members of the same family. The responses to some NSDUH questions have high positive correlations between pair members, especially the questions about family-level characteristics. The current imputation method, predictive mean neighborhood (PMN), does not exploit this correlation, even when the value for one pair member is missing and the value for the other pair member is not missing. This presentation discusses (1) a method for identifying variables for which the other pair member's response may be a better choice for imputation than the PMN-assigned value; and (2) for these variables, a method for identifying exact conditions under which the other pair member's response may be used.

Key Words: imputation, editing, NSDUH

1. Imputation Using the Other Pair Member

The National Survey on Drug Use and Health (NSDUH) provides national, state, and substate data on substance use and mental health in the civilian, non-institutionalized population age 12 and older. The survey is sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA), U.S. Department of Health and Human Services, and is planned and managed by SAMHSA's Center for Behavioral Health Statistics and Quality (CBHSQ). Data collection and analysis are conducted under contract with RTI International, Research Triangle Park, North Carolina.

In the NSDUH, zero, one, or two people are selected from each selected household. A "pair" occurs when two people are selected from a selected household, and both are unit respondents. Pairs enhance the analytic capability of the NSDUH (Chromy & Singh 2001). In the 2009 NSDUH, 58.8 percent of the unit respondents[1] were members of a responding pair. The pair relationship can be parent-child, sibling-sibling, spouse-spouse, or some other relationship. Many variables for which imputation is performed, e.g., health insurance coverage, can be expected to have high positive correlation between pair respondent members. However, the information about the other pair member is used only in editing of variables related to the household roster. This study evaluates whether assigning the value of the other pair member would be a better imputation method than

[1] A case is defined as a unit respondent if data were provided on lifetime use of cigarettes and at least nine other substances.

the current method, predictive mean neighborhood imputation (PMN) (Singh, Grau, & Folsom 2002).

The goal of this exercise is to assess the feasibility of using information from one pair member to assist with imputation of missing data for the other pair member. This includes choosing candidate variables for which this method would be appropriate, and determining whether the benefits of using this method outweigh the costs of development and implementation.

1.1 Choosing Candidate Variables

Certain variables are obvious candidates for this method. Some of the questions in the NSDUH ask for household-level information, such as the household roster. The responses the pair members give to these questions should almost always agree. Other than measurement error, the only reason for disagreement would be a change in the household composition between the times when the questions are answered. Some NSDUH questions, such as the majority of the ones in the income module, ask for information about the family in the household. If the pair members are members of the same family, then the responses are more likely to agree than if the pair members are not members of the same family.

In general, good candidate variables for this approach meet the following conditions:

- When both pair members are item respondents for the variable, their values almost always agree.
- There are enough missing values where the other pair member responded to justify the cost of developing and implementing the extra imputation steps.
- For item nonrespondents paired with item respondents, the values imputed by PMN often differ from the other pair member's response, and it is logical to assume that the responses would be the same (i.e., a better imputation method than PMN appears readily available for these item nonrespondents).

Table 1 below provides some information on agreement rates between two responding pairs, extent of nonresponse and agreement rates after the implementation of current imputation methods for variables that lend themselves better to this different imputation method. "Eligible nonrespondents" are those who (1) should have a valid response for this variable, (2) are a member of a pair, and (3) are paired with an item respondent for the variable who also should have a valid response. The demographic, household roster, income, and health insurance groups have an average agreement of 75 percent or higher between pair responses. Within the health insurance and household roster groups, an average of almost 50 percent of nonrespondents is eligible for the proposed method. Using PMN results in imputing a value that disagrees with the pair member for approximately 40 percent of eligible nonrespondents in the demographic and income groups and for almost 50 percent in the household roster group.

Table 1: Summary of Agreement Rates by Variable Group

Variable Group	Average Percentage of Pair Agreement	Average Percentage of Eligible Nonrespondents	Average Percentage of Pair Agreement after PMN
Demographics	75.7	26.8	61.5
Health Insurance	87.2	46.5	68.3
Income	82.4	38.4	58.7
Household Roster	94.6	45.4	52.3

Source: SAMHSA, Center for Behavioural Health and Quality, National Survey on Drug Use and Health, 2009.

Most of the demographic, household roster, income, and health insurance variables are possible candidates for this method because of their high pair agreement rates, relatively high percentages of eligible nonrespondents, and relatively high percentages of pair disagreement after the implementation of the PMN imputation method.

1.2 A Closer Look at Candidate Variables

As expected, the demographic, household roster, income, and health insurance variables appear to be possible candidates. Of these groups, the income and health insurance variables were selected for closer examination for the following reasons:

- The income and health insurance variables have more missing data, compared to the demographic and household roster variables.[2]
- All of the demographic variables are at the person level, not the household level or the family-in-household level, so the direct assignment of the value of the other pair member may result in lower quality imputed data. It may be better to use the other pair member as an influence on the final imputed value, not as the sole determinant of the final imputed value, perhaps by using the other pair member's value in the prediction model, for example. This approach was not examined in this study.
- All of the questions in the health insurance module ask about the respondent, not the family in the household. Still, since many health insurance plans provide coverage at the family level, the pair members are expected to agree more often when they are members of the same family. Since these questions ask about current coverage by health insurance, the number of days between the responses may be a factor: coverage status may change between the times the responses were given.
- All of the questions in the income module ask about the preceding calendar year, so the number of days between the responses of the pair members theoretically should not be a factor. Most of the questions in the income module ask about the family in the household, so as long as the pair members are members of the same family, the only theoretical source of disagreement is measurement error.[3]

[2] Item response rates are available in Appendix A of Ault et al. (2010). For the income variables examined in this study, weighted item response rates ranged from 90.3 to 99.8 percent. For the health insurance variables examined in this study, weighted item response rates ranged from 99.3 to 99.8 percent.

[3] Perhaps the differences between responses can be used as a simple assessment of measurement error for these items. The differences can be used directly for income, and for the household roster, the differences can be used if the responses were given on the same day (or within a reasonable number of days).

When taking a closer look at the income and health insurance variables, three factors were considered:

- **Family Pair Indicator:** As stated above, most of the income questions were asked at the family-in-household level, and although the health insurance questions were asked at the respondent level, family members might be expected to agree much of the time in their responses. There is an imputation-revised variable called IRPRREL that identifies the pair type. This variable was collapsed into a dichotomous variable for further analysis: either the pair members were clearly in the same family (parent-child, sibling-sibling, spouse-spouse, or grandparent-grandchild) or not. In the 2009 NSDUH, 85.6 percent of the responding pairs were clearly members of the same family according to IRPRREL.
- **Number of Days between Responses:** Date stamps are available for each module for each respondent, so the number of days between the responses can be calculated. As stated above, for the income questions, the number of days between responses should theoretically not affect responses. However, it may increase the likelihood of measurement error. For the health insurance questions, the number of days between responses may be important because the likelihood that there is a change in current health insurance coverage presumably would increase as the number of days between responses increases. In 2009, the responses were entered on the same day 65.5 percent of the time, within 7 days 86.2 percent of the time, within 14 days 92.2 percent of the time, and within 30 days 96.7 percent of the time.
- **Whether the Same Person Answered Both Questions:** For many respondents, the income and health insurance questions were answered by a proxy. This proxy had to be a member of the respondent's family and had to be at least 18 years old. Sometimes, the proxy was the same person as the other pair member. In these cases, the pair members might be expected to agree practically all of the time. An assessment was done of (1) how frequently the responses agree when the proxy and the other pair member are the same person, and (2) how frequently one response is missing and the other is not. This assessment led to the following conclusions:

    - In many cases, it is not easy to determine whether the proxy and the other pair member are the same person. A careful review of the household roster for each pair member is required. An algorithm was designed to identify the most obvious cases. For example, if the pair type was parent-child, the father answered the questions for the child, the parent reported being male, and the parent answered the questions himself, then it was assumed that the father answered both questions.
    - When the same person gives both responses, the responses agree practically all of the time, and it is very rare that one response is missing and the other is not.
    - Because of the difficulty of determining whether the same person gave both responses, and because of the limited number of cases where one response is missing and the other is not, this factor was dropped from consideration.

The two sections below describe the results of logistic regression models involving the income and health insurance variables. The dependent variable is an indicator of agreement between pair members, and the two independent variables are the first two factors mentioned above: the pair type and the number of days between interviews.

1.2.1 Income

There are nine edited income variables at the level of the family in the household. Seven of them are dichotomous, mostly pertaining to the receipt of income from different sources. The other two variables include a variable on the months on welfare and one categorical variable with different levels of family income. Separate logistic regression models were fit for each of the nine variables. As expected, for all nine models, the family pair indicator was a statistically significant covariate ($\alpha = 0.05$) and the predicted probabilities of agreement were higher when the pair members were in the same family. Also as expected, the predicted probability of agreement decreased with the increase in number of days between interviews for all nine variables. For seven of the nine variables, the regression coefficient associated with the number of days between interviews was significantly different from zero ($\alpha = 0.05$).[4] Table 2 lists the income variables used in the models, the percentages of family pairs and other pairs whose responses agree, and whether the regression coefficients were statistically significant.

The proportion of pairs that agree is lower for welfare months and for total family income (finer categories). This is mostly because the other variables are dichotomous, while these two are ordinal with several levels. The indicator of agreement does not account for the magnitude of the disagreement. For example, the months-on-welfare variable has integer values from 1 to 12: if one pair member reports 10 months on welfare in the prior year and the other reports 9, the pair is still in disagreement, just as if one pair member reports 1 month on welfare and the other reports 12 months on welfare. The finer-categories-of-income variable has 29 levels.

Table 2: Summary of Logistic Regression Results, Income Variables

Variable	Number of Pairs Used in Analysis	Actual Percentage of Pairs that Agree		Statistical Significance of Covariates	
		Family	Other	Family Pair Indicator	Number of Days between Interviews
Social Security	19,855	96.05	93.08	Yes	Yes
Supplemental Security Income	19,718	96.86	94.81	Yes	Yes
Welfare Payments	19,924	97.96	96.44	Yes	No
Welfare Services	19,986	96.81	95.41	Yes	Yes
Wages	20,101	94.96	85.05	Yes	No
Food Stamps	20,056	96.50	90.82	Yes	Yes
Welfare Months	854	80.13	66.67	Yes	Yes
Total Family Income (dichotomous)	19,049	95.05	78.17	Yes	Yes
Total Family Income (finer categories)	17,355	68.12	18.67	Yes	Yes

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Although the regression coefficient on the number of days between interviews was statistically significant for most income variables and negative for all income variables, statistical significance is easy to achieve when the sample size is so large. For all income variables except welfare months, the sample size includes several thousand pairs. It is possible that the results are statistically significant but not important in a practical sense:

[4] To check whether the results were affected by a few disagreeing pairs with a large number of days between interviews, the models were re-fit using an ordinal version of this dependent variable (0-7, 8-30, 31 or more). Results were similar.

the regression coefficient may still be close to zero, causing the predicted probability of agreement to decline slowly as the number of days between interviews increases. To assess this for the family pairs, predicted probabilities of agreement were calculated for fixed values of the number of days between interviews. These results are displayed in Table 3.

As shown in Table 3, for most variables, the predicted probability of agreement decreases slowly with the number of days between interviews. Recall that in 2009, the number of days between interviews was less than or equal to 30 for about 97 percent of the pairs. When the number of days between interviews is 30, the predicted probability of agreement is still greater than 90 percent for all variables except welfare months and finer categories income.

Table 3: Predicted Probabilities of Agreement for Family Pairs, as a Function of the Number of Days between Interviews, Income Variables

	Predicted Probability of Agreement for Various Values of the Number of Days between Interviews (%)						
Variable	**0**	**5**	**10**	**20**	**30**	**50**	**70**
Social Security	96.23	96.00	95.75	95.20	94.60	93.16	91.37
Supplemental Security Income	96.99	96.82	96.63	96.22	95.77	94.70	93.39
Welfare Payments	98.03	97.94	97.85	97.67	97.46	97.01	96.47
Welfare Services	96.86	96.79	96.71	96.56	96.40	96.05	95.68
Wages	94.99	94.95	94.90	94.81	94.71	94.52	94.32
Food Stamps	96.75	96.45	96.11	95.36	94.46	92.16	89.02
Welfare Months	81.44	78.40	75.01	67.26	58.43	39.69	23.56
Total Family Income (dichotomous)	95.48	94.97	94.40	93.08	91.48	87.25	81.34
Total Family Income (finer categories)	70.95	66.62	62.00	52.15	42.13	24.52	12.66

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Given that (1) the income questions ask about the prior year so that the response is not theoretically dependent on the exact date in the current year when the response was given, (2) the income variables discussed in this section store data at the level of the family in the household, and (3) for most of the variables discussed in this section, the predicted probability of agreement decreases slowly as the number of days between interviews increases, it appears that the use of the other pair member's value in imputation for these nine income variables as long as the pair members are in the same family would improve the quality of imputation. Based on this analysis it does not appear to be necessary to consider the number of days between interviews. As shown in Table 4, following this method of imputation would have reduced the amount of PMN imputation required by up to 40 percent for these variables in 2009.

Table 4: Under Recommendation, Proportion of Item Nonrespondents Handled Using the Other Pair Member, Income Variables

Variable	Number of Item Nonrespondents				Percentage Handled Using Other Pair Member
	Total	In Pairs	(and) Paired with an Item Respondent	(and) In Family Pair	
Social Security	660	364	310	255	38.64
Supplemental Security Income	935	524	424	352	37.65
Welfare Payments	492	293	243	190	38.62
Welfare Services	371	222	190	150	40.43
Wages	193	101	81	65	33.68
Food Stamps	262	147	125	97	37.02
Welfare Months	218	71	45	37	16.97
Total Family Income (dichotomous)	2,557	1,430	856	689	26.95
Total Family Income (finer categories)	6,624	3,836	1,828	1,511	22.81

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

1.2.2 Health Insurance

There are eight edited health insurance variables that undergo imputation. Most of them denote the type of health insurance a person has or whether they have specific types of health insurance. Logistic regression models were fit for these eight variables using the same methodology that was used for the income variables. Table 5 summarizes the results of the models. The family pair indicator was a significant covariate for seven of the eight models, and the number of days between interviews was a significant covariate for five of the eight models. The predicted probabilities of agreement were higher when the pair members were in the same family for eight of the nine variables. As for income, the predicted probability of agreement decreased with the number of days between interviews for all variables.

Table 6 shows the predicted probabilities as a function of the number of days between interviews for family pairs only. For all variables, the predicted probability of agreement decreases slowly as the number of days between interviews increases. This suggests that coverage status does not tend to change very often; assuming no measurement error, it tends to stay constant over the short term.

Table 5: Summary of Logistic Regression Results, Health Insurance Variables

Variable	Number of Pairs Used in Analysis	Percentage of Pairs that Agree		Statistical Significance of Covariates	
		Family	Other	Family Pair Indicator	Number of Days between Interviews
Overall Health Insurance, 1999 Method	19,894	83.08	69.18	Yes	Yes
Overall Health Insurance, 2001 Method	19,882	84.85	72.75	Yes	Yes
Private Health Insurance, Consistent with Pre-1999 Surveys	19,932	84.80	70.24	Yes	Yes
Medicaid/CHIP	19,899	86.92	84.15	Yes	No
Medicare	20,052	96.41	96.83	No	No
Military Health Care (CHAMPUS, TRICARE, CHAMPVA, VA)	20,070	97.88	95.79	Yes	No
Private Health Insurance, as Defined by Constituent Variables Method	19,932	84.80	70.24	Yes	Yes
Other Health Insurance	2,069	92.38	87.18	Yes	Yes

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Table 6: Predicted Probabilities of Agreement for Family Pairs, as a Function of the Number of Days between Interviews, Health Insurance Variables

Variable	Predicted Probability of Agreement for Various Values of the Number of Days between Interviews (%)						
	0	5	10	20	30	50	70
Overall Health Insurance, 1999 Method	83.43	82.93	82.41	81.34	80.21	77.82	75.22
Overall Health Insurance, 2001 Method	85.20	84.71	84.21	83.16	82.06	79.70	77.12
Private Health Insurance, Consistent with Pre-1999 Surveys	85.21	84.64	84.05	82.82	81.51	78.66	75.51
Medicaid/CHIP	87.06	86.85	86.64	86.20	85.75	84.82	83.84
Medicare	96.46	96.39	96.32	96.18	96.04	95.73	95.39
Military Health Care (CHAMPUS, TRICARE, CHAMPVA, VA)	97.88	97.87	97.87	97.85	97.83	97.80	97.76
Private Health Insurance, as Defined by Constituent Variables Method	85.21	84.64	84.05	82.82	81.51	78.66	75.51
Other Health Insurance	92.78	92.32	91.82	90.75	89.55	86.74	83.31

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

As stated above, the health insurance questions in the NSDUH differ from the income questions in two ways: (1) they are asked at the respondent level, not the family level; and (2) they ask about current coverage, not coverage over some fixed time interval. Because of this, the use of the other pair member's value in imputation is harder to justify theoretically, even when the responses were given at the same time. Even when the

responses were given on the same day, there is a nontrivial proportion of disagreeing responses. For example, for family pairs, the predicted probability of agreement is below 90 percent for five of the eight variables, even when the number of days between interviews is zero. Just to verify that the model's predictions were reasonable when the two interviews were conducted on the same day, the actual proportions of agreement when there were no days between interviews was compared with the predicted probabilities. The two measures were similar, as shown in Table 7.

Table 7: Comparison of Proportion of Agreement to Predicted Probability of Agreement for Family Pairs, for Interviews Conducted on the Same Day, Health Insurance Variables

	Respondent Pairs with No Days between Interviews			
Variable	**Total**	**Number of Pairs that Agree**	**Percentage of Pairs that Agree**	**Predicted Probability of Agreement, No Days between Interviews**
Overall Health Insurance, 1999 Method	15,082	12,618	83.66	83.43
Overall Health Insurance, 2001 Method	15,076	12,868	85.35	85.20
Private Health Insurance, Consistent with Pre-1999 Surveys	15,115	12,900	85.35	85.21
Medicaid/CHIP	15,082	13,143	87.14	87.06
Medicare	15,191	14,660	96.50	96.46
Military Health Care (CHAMPUS, TRICARE, CHAMPVA, VA)	15,202	14,884	97.91	97.88
Private Health Insurance, as Defined by Constituent Variables Method	15,115	12,900	85.35	85.21
Other Health Insurance	1,436	1,330	92.62	92.78

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, National Survey on Drug Use and Health, 2009.

Because there are two clear reasons why pair members can disagree, even if they are members of the same family, this imputation method does not appear to be suitable for the health insurance variables. Perhaps a more detailed investigation of the frequency of agreement between family members for certain types of health insurance would lead to the conclusion that the other-pair-member approach could be justified under specific conditions.

1.3 Summary and Recommendations

The purpose of this study was to assess the feasibility of an alternative approach to imputation: a simple assignment of the value of the other pair member. In certain situations, this approach seems preferable to PMN, since it is both simpler and more accurate.

A crude assessment of feasibility applied to all variables that undergo imputation suggested that the income and health insurance variables were the best candidates for this method. There are nontrivial numbers of missing values for most of these variables, and

there are adequate numbers of cases to which the other-pair-member method would apply: item nonrespondents that were paired with item respondents. For these cases, the proportion of pairs whose responses agreed after PMN imputation was usually considerably lower than the proportion of responding pairs whose responses agreed, suggesting that the other-pair-member approach was preferable. Most of the income questions ask about the family in the household, and when both pair members respond, they very often agree. Although the health insurance questions ask about the respondent, not the respondent's family in the household, pair members often agree when they are members of the same family.

For the income and health insurance variables, a more refined analysis was done to identify exact conditions under which the use of the other pair member was preferable to PMN. Two factors were considered: (1) the type of pair (definitely in the same family, or not), and (2) the number of days between the presentation of the questions to the pair members. After consideration of these factors using logistic regression models, the following conclusions may be drawn:

- For the nine income variables that store data at the level of the family in the household, the use of the other pair member's value in imputation appears to improve the quality of imputed data as long as the pair members are in the same family. This would also reduce the amount of PMN imputation required for these variables by more than 30 percent.
- For the eight health insurance variables that undergo imputation, the use of the other pair member's value in imputation does not appear to improve data quality, because the deterministic nature of this method is inappropriate given the rate of disagreement between responding pair members.

Some reasonable next steps include the following:

- **Take a closer look at the household roster variables.** The household roster variables that undergo imputation store data at the household and family-in-household level. That alone makes them good candidates for the other-pair-member approach. Like the health insurance questions, the household roster questions ask about the current situation instead of the situation at some fixed period like the income questions; thus, the responses are somewhat dependent on the date the questions were presented. Perhaps an approach similar to that suggested for the health insurance variables would be reasonable. For the questions about the household that are not dependent on familial relationships, perhaps the pair type would not be an important factor. The main reason the household roster variables were not considered in this chapter was that the level of missingness is low.
- **For the health insurance variables, search for specific situations where using the other pair member's response in imputation is appropriate.** Some of the types of health insurance covered by the NSDUH may be at the family level, even though the questions are asked only of the respondent. It might be useful to consult with a subject matter expert and to complete more refined data analyses to determine when the other-pair-member approach is best.
- **Consider using the other pair member's response in the prediction models.** For many variables, including the health insurance and demographic variables, the other pair member's response may be useful as supporting information but not as the sole

determinant of the imputed value. Methods for integrating this information into the prediction models could be explored.

Acknowledgements

The presentation is sponsored by RTI International's Statistics and Epidemiology Unit, with the research included in the presentation stemming from ongoing methodological work conducted under the contract for the National Survey on Drug Use and Health (NSDUH). The NSDUH is funded by the Substance Abuse and Mental Health Services Administration (SAMHSA), Center for Behavioral Health Statistics and Quality under contract no. 283-2008-00004C and project no. 0211838.

The views expressed in this presentation do not necessarily reflect the official position or policies of SAMHSA or the U.S. Department of Health and Human Services; nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

References

Ault, K., Carpenter, L, Frechtel, P., Laufenberg, J., Martin, P., McNerney, V., & Moore, A. (2010). 2009 National Survey on Drug Use and Health: Editing and imputation evaluation report (prepared for the Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality, under Contract No. 283-2004-00022, Deliverable No. 15, RTI/0209009.583). Research Triangle Park, NC: RTI International.

Chromy, J.R., and Singh, A.C. (2001). Estimation for person-pair drug related characteristics in the presence of pair multiplicities and extreme sampling weights for the NHSDA. ASA Proc. Surv. Res. Meth. Sec.

Singh, A. Grau, E., & Folsom, R. (2002). Predictive mean neighborhood imputation for NHSDA substance use data. In Redesigning an ongoing national household survey: Methodological issues. DHHS Publication No.SMA 03-3768, edited by J. Gfroerer, J. Eyerman and J. Chromy. Rockville, MD: Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality.