Statistics Can Lie But Can Also Correct for Lies: Reducing Response Bias in NLAAS via Bayesian Imputation

Jingchen Liu, Xiao-Li Meng, Chih-nan Chen, and Margarita Alegria

Working Draft: January 4, 2008

Abstract

National Latino and Asian American Study (NLAAS) is a multi-million-dollar survey of psychiatric epidemiology, the most comprehensive survey of this kind and its data were made public in July 2007. A unique feature of NLAAS is its embedded experiments for estimating effect of alternative orderings of interview questions. The findings from the experiment are not completely unexpected, but nevertheless astonishing. Compared to the survey results from the widely used traditional ordering, the self-reported psychiatric service-use rates are often doubled or even tripled under a more sensible ordering introduced by NLAAS. These findings partially answer some perplexing questions in the literature, e.g., why the self-reported rates of using religion services were typically much lower than results from other empirical evidences? But at the same time, they also impose some grand challenges, for example, how can one assess racial disparities when different races were surveyed with different survey instruments (e.g., the exiting data on the White populations were collected using the traditional questionnaire ordering) that are now known to induce substantial differences? The project documented in this paper is part of the effort in addressing these questions, by creating models for imputing the correct responses had the respondents under the traditional survey had not taken advantage of the skip patterns to reduce interview time, which resulted in increased rates of untruthful negative responses over the course of the interview. The imputation modeling task is particularly challenging because of the complexity of the questionnaire, the small sample sizes for subgroups of interests, the existence of high-order interactions among variables, and above all, the need of providing sensible imputation for whatever subpopulation a future user might be interested in studying. This paper is intended to serve three purposes: (1) to provide a published record of the key steps and strategies adopted in creating the released multiple imputation for NLAAS, (2) to alert the potential users of the limitations of the imputed data, and (3) to provide a vivid demonstration of the type of challenges and opportunities typically encountered in modern applied statistics.

1 Trivial Ordering But Serious Bias

1.1 National Latino and Asian American Study

The National Latino and Asian Study (NLAAS) is a complex interview-based survey of household residents, ages 18 or older, in the non-institutionalized population of the coterminous United States. A basic task of NLAAS is to report the prevalence of psychiatric disorders and service uses, where psychiatric disorders are based on diagnosis and service uses are based on self-reporting. The
sample consists of 2554 Latinos, 2095 of Asians, and 215 of whites. Overall, there are more than 5000 variables, measured or constructed based on raw measures, available for users; the data were just made public in July of 2007. Details about NLAAS can be found in [2, 14].

Survey responses are known to be influenced by many factors, including the seemingly innocent ordering of the questions. A very substantial response bias induced by ordering is observed in NLAAS for the respondents’ self-reported psychiatric service uses. The ordering effect was detected because NLAAS has two sets of questionnaire designs, the traditional design and a new design, which share identical set of questions, but they are arranged in different orders for the service use part.

There are thirteen types of psychiatric services in NLAAS, as listed in Table 1. For each service, there is a “stem question” asking if the respondent ever had this service before, for both lifetime and during the past 12 months. Together with the stem question, there are between five to ten follow-up questions asking more details about a self-reported service use, such as when did the respondent have the service the first time and the last time, how long was the treatment, etc. The follow-up questions were obviously skipped by the interviewer if the respondent responded negatively to the stem question. This logically correct skipping pattern, however, has an unintended interaction with the ordering of the questions.

1.2 So What Is Wrong with the Traditional Design?

The traditional (common) design adopts a sequential ordering. After each stem question, if the response is positive, follow-up questions are asked immediately; otherwise, the follow-up questions are skipped and the next stem question is asked. In addition, the traditional design also arranged the whole module of service questions after a series of diagnostic questions for identifying psychiatric disorders. And to make the matter worse, similar service-use questions, including the follow-up questions, were asked within each diagnostic session (more discussion of this later). This implies that service questions typically comes 30 minutes after the interview starts, as illustrated in the left column of Figure 1, adopted from [8], and by then the respondents had ample opportunity to realize the unintended benefit of the skipping pattern.

This sequential design has been used in common practice with a long history; see, for instance, [23]. That NLAAS has an embed experiment was due to the suspicions by its investigators (one of us, Alegria is a Co-PI for NLAAS) that subjects who are given the traditional design might be more likely to underreport the actual service use for (at least) two reasons. First, because the follow-up questions are asked immediately after each stem question, respondent can learn quickly from the previous service question format and underreport to avoid follow-up questions to shorten the interview period. The interview for such detailed survey tends to be very long; for NLAAS, the average interview time is about 2.5 hours, with the maximum reported interview time being 48 hours (though it is not clear what such an extreme figure really means!). Secondly, the stem questions are asked after the psychiatric diagnostic questions (which itself provides ample opportunity for learning the skipping pattern). Respondents tend to react more negatively when they run out of patience, especially when being asked about activities that are subject to memory decay. That is, why should I subject myself with so many follow-up questions when I’m not 100% sure if I ever had a particular service in the first place?
1.3 An Embed Experiment within NLAAS

To confirm such an ordering effect and assess its impact, NLAAS included an experiment. 75% of the subjects were randomly assigned to the traditional *sequential design* as described, while 25% were assigned to a new *parallel design*. The new design moved all the stem question far ahead, before all the diagnostic questions, but leaves additional follow-up questions after the diagnostic questions as illustrated in the right column of Figure 1. In the new design, follow-up questions are asked long after all stem questions are completed. Also, stem questions in the new design come earlier than those in the traditional design. That is, at the time of answering the stem questions, the interview was at the relatively early stage, and the respondent had no opportunity to learn that a “yes” answer would subject to him/herself with a large number of follow-up questions.

The 75-25 splitting of the sample, instead of the more natural 50-50 splitting, was more or less out of concerns of maintaining comparability with the literature, where essentially all data were collected using the sequential ordering. That is, if something goes wrong with the new design, we would still have 75% usable data. Unfortunately, the end results, as we shall see shortly, are that the sequential design is subject to serious underreporting, thereby in effect we only have about 25% usable data as far as the service use measurement goes.

This can be most easily seen in Table 1, which compares the (weighted) samples averages for the self-reported life-time service uses (and later in Table 5 for the past-12-month service uses).
The estimates from the new design is uniformly higher than those from the traditional design for all 13 services. The order of the questions in Table 1 is the same as they were in the actual survey. However, the last three service questions (i.e., from “Hot Line” and on) were asked only once during the stem-questions session, under both designs, irrespective of the type of disorders a respondent might have suffered. In contrast, for the first 10 service questions, they were also asked within each disorder category (e.g., Depression, Panic disorder, Social Phobia, etc.) in addition to the stem-question session. Therefore, for the first 10 questions, under the traditional designs, respondents have a more repetitive task for each service question (since they were asked within each disorder section), a fact that may have induced respondents’ negative responses more than questions 11-13 had. For the new design, all the stem questions on the service use were first asked before all the disorder questions and then followed up later in disorder diagnostic sections. Since a respondent is classified to have a particular service, say, psychiatrist service, if the respondent reported positively to that question under at least one of the disorder categories or to the stem question, under the new design the learning of use skipping patterns during the disorder diagnostic sessions would have little effect on the self-reported service uses because it can only take place after the completion of all stem questions.

The phenomena of underreporting under the traditional design persisted in each sub-population when the overall population was stratified according to ethnicity, as studied in detail in [8]. However, non-service variables show no significant difference between the two design groups at all; Table 2 demonstrates this for a randomly selected group of variables. It is therefore logical, or as confident as one can possibly be with comparisons of this sort, to conclude that the significant discrepancy in the self-reported service use rates (with \( p \)-value < 0.01, which is robust to different model assumptions, as discussed in [8]), is a direct result of the different orderings of the questions.

<table>
<thead>
<tr>
<th>New Design</th>
<th>Old Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Psychiatrist</td>
<td>14.9%</td>
</tr>
<tr>
<td>2. General Practitioner</td>
<td>17.6%</td>
</tr>
<tr>
<td>3. Other Medical Doctor</td>
<td>9.2%</td>
</tr>
<tr>
<td>4. Psychologist</td>
<td>13.4%</td>
</tr>
<tr>
<td>5. Social worker</td>
<td>7.6%</td>
</tr>
<tr>
<td>6. Counselor</td>
<td>13.2%</td>
</tr>
<tr>
<td>7. Other Mental Health Prof</td>
<td>5.3%</td>
</tr>
<tr>
<td>8. Nurse, Occupational Therapist,</td>
<td>4.0%</td>
</tr>
<tr>
<td>9. Religious/Spiritual Advisor</td>
<td>15.3%</td>
</tr>
<tr>
<td>10. Other Healer</td>
<td>5.9%</td>
</tr>
<tr>
<td>11. Hot Line</td>
<td>2.3%</td>
</tr>
<tr>
<td>12. Internet Group or Chat Room</td>
<td>2.9%</td>
</tr>
<tr>
<td>13. Self Help Service</td>
<td>5.9%</td>
</tr>
</tbody>
</table>

Table 1: Comparing Self-reported Lifetime Service Uses

1.4 Our Challenging Task

Serving as a data set for public use, such response bias in NLAAS will lead to misleading results for most potential analysis involving service uses. One strategy to deal with such a problem is to create
multiple imputation to impute the unobserved responses of those given the traditional design had they were given the new design. Multiple imputation is a handy tool for dealing with incomplete data via complete data procedures, see [24], [11], [10]. Further results on how to create and analyze multiple imputations can be found in [4], [19], [20], [15]. Our central task, therefore, was to create multiply imputed data sets for the service uses under traditional design based on the data from the new design, which is viewed as (approximately) bias free. As demonstrated in the literature cited above, Bayesian prediction is a principled approach to accomplish this task, but to yield sensible results, the modeling task, as well as the associated computational task, are often tremendous. The rest of this paper summarize our effort in these regards.

Specifically, Section 2 documents our basic model assumptions, including prior specification, and the corresponding imputation results for life-time service use. Section 3 reported a model improvement effort over the initial model to address high-order interactions, as well as extensions for the last-12-month service uses. Section 4 gives the key steps of the MCMC methods we used for fitting the model and creating multiple imputations. Section 5 discusses the tough issue of assessing the quality of imputation, and concludes.

## 2 Preliminary Model Specifications

### 2.1 Modeling Response Behavior

Considering the context discussed in Section 1, our most basic assumptions are:

1. Respondents from the new design group responded truthfully;
2. Respondents from the traditional design group responded truthfully, if the respondent indeed did not have service.

One could of course argue how do we know that respondents from the new design were not over-reporting? Strictly speaking, we don’t, and there is no information in the data to verify Assumption 1. However, there was no compelling reasons or conceivable incentives for over-reporting under the new parallel design, unlike the strong incentive for under-reporting under the traditional design for reasons listed in Section 1.2. Any applied statistical project requires some common sense
judgements, and Assumption 1 is ours for the current project. However, even for those who do not agree with this assumption, our imputations can be viewed as predicting the self-reported service rates for those from the traditional design groups had they were given the new design. That is, our imputation results are useful for adjusting one group’s responses according to the other group, without referencing to which group is correct. (Our modeling and computational strategies below are evidently applicable if we reverse the roles of the two groups, and indeed the task would be somewhat easier because we would have 75% of data to build the predictive model, instead of only 25%).

The rationale behind Assumption 2 is also common sense. Under the traditional design, there is really no incentive to knowingly provide false positive response, because that can only prolong the interview. In addition, falsifying a positive response to a stem question will require the respondent to falsify answers to an array of follow-up questions, which is not a trivial task without actual service experience.

Given these two assumptions, our basic sampling model is as the following. For each respondent, we adopt following notations:

- $y$: the self-reported service use status, 1 for having service and 0 otherwise;
- $S$: the true service use status, 1 for having service and 0 otherwise.
- $\xi$: the response behavior of those people from the traditional design group who have service use (i.e. $S = 1$), 1 for responding truthfully and 0 otherwise;
- $I$: the group design indicator, 1 for traditional design and 0 for new design.

Using these notation, we have $y = S$ when $I = 0$, but $y = S\xi$ when $I = 1$. Note here that for our modeling purposes $\xi$ can be defined arbitrarily when $S = 0$ because $y = 0$ whenever $S = 0$. Therefore, for simplicity we can assume $S$ and $\xi$ are independent, and both follow the Bernoulli distribution:

$$S \sim \text{Bernoulli}(p), \quad \xi \sim \text{Bernoulli}(r).$$

This is the very first step of our modeling effort, as the whole modeling process will need to take into account covariates, survey design and weights, prior specification, etc, as shall be detailed shortly. Here we just use this simplest model to illustrate how the information from the new-design group is used by the likelihood to estimate the rate for the traditional-design group.

Let’s consider a simple case where we have $n$ independent samples, $n_0$ of them with $I = 0$ and $n_1$ with $I = 1$. Define $m_0 = \sum_i y_i(1 - I_i)$ and $m_1 = \sum_i y_iI_i$, which are the numbers of positive responses in group $j = 0, 1$ respectively. In the absence of any covariates, the likelihood function would be given by

$$L(p, r) \propto p^{m_0}(1 - p)^{n_0 - m_0}(pr)^{m_1}(1 - pr)^{n_1 - m_1}. \quad (1)$$

The maximum likelihood estimate (MLE) provides us some insights because it takes the following form:

- When $\frac{m_0}{n_0} > \frac{m_1}{n_1}$, $\hat{p} = \frac{m_0}{n_0}$ and $\hat{r} = \frac{m_1}{m_0}/\frac{n_1}{n_0}$;
- When $\frac{m_0}{n_0} \leq \frac{m_1}{n_1}$, $\hat{p} = \frac{m_0 + m_1}{n_0 + n_1}$ and $\hat{r} = 1$. 

6
Intuitively, when the new design group (Group 0) has higher observed service rate, then the rate from the order design Group 1 provides little information for $p$ under our model assumption, because it suffers from underreporting. However, if Group 1 has higher observed rate, then under our model assumption the MLE approach will have no choice but to declare that the truthful response rate, $r$, in group 1 to be 100% (i.e., no one lied), therefore the two groups will be pooled together in estimating $p$. The latter situation occurs within some small sample stratum, which imposes an intriguing issue for our model, as we shall discuss in Section 5, but typically we encounter the former situation.

2.2 Multivariate Probit Regression with Clustering

The imputation task would be trivial—and in fact meaningless—if our goal is to just “fix” the overall service use rates for the traditional design group. The simple model given in Section 2.1 would do the job. The ideal goal here, however, is to correct the rate for any subpopulation that might be of interest to a potential analyst of NLAAS. This turns out to be a far exceedingly difficult, indeed impossible, task for NLAAS (or any similar survey), because NLAAS literally has more than 5000 variables but only has 4864 subjects. In principle we should use all variables for reasons discussed in [20], but this is simply impossible given the complexity of the survey and the limitation of data and our resources and schedule constraints; the schedule constraint is the one that most theoretical statisticians might have hard time to appreciate, but by contract with the funding agency the data must be made available to the general public by a pre-agreed date. Therefore, we have to compromise by using predictive variables that are noticeably correlated with the service use and those that are likely to be used more frequently in analysis involving service use.

Coming up with such a list of variables is a very difficult task by itself. It was a long iterative process, based on both statistical analysis and discussions (sometimes even negotiations) with researchers in the substantive fields, as well as considerations of model’s identifiability and computational constraints. The chosen variables are

- **Categorical variables**: marital status, insurance status, working status, region in the country, ethnicity, immigration status, gender, various psychiatric disorder diagnostic including any affective disorder (lifetime and last year), any substance disorder (lifetime and last year), any anxiety disorder (lifetime and last year), and any disorder (last year).

- **Continuous variables**: logarithm of annual income, total number of psychiatric disorders, social status, age, k10 distress (a psychiatric disorder measure), and logarithm of survey weights.

Note that it was necessary to treat some category variables, such as total number of psychiatric disorders, as continuous variable, in order to reduce the number of the predictive variables (recall a categorical variable requires multiple dummy variables to represent its levels) so that some other variables can be accommodated. The total number of variables (including dummy variables) for our imputation model was 39.

To incorporate these variables in imputation together with the dependence among those 13 service uses, we need a suitable model for multi-variate categorical data. There are many such models. We initially chose multivariate probit model mainly because of its computational and interpretational simplicity due to its normal latent structure. The model also makes the sampling from the corresponding posterior predictive distribution easier (but not easy, as we will report late), especially when the number of covariates is large.
Specifically, let $S^{(j)}$ indicate the $j$th service use and $\xi^{(j)}$ be the corresponding response behavior indicator, $j = 1, \ldots, 13$. The multivariate probit regression model links these 26 indicator variables to a 26-dimension latent vector $Z = (Z^{(1)}, \ldots, Z^{(26)})^\top$ such that $S^{(j)} = \mathbf{1}\{Z^{(j)} > 0\}$ and $\xi^{(j)} = \mathbf{1}\{Z^{(j+13)} > 0\}$, where $\mathbf{1}\{A\}$ is the usual set indicator function. The latent vector is then assumed to follow a multiple regression model

$$Z_i \sim N(\beta^\top X_i + W_c, \Sigma), \tag{2}$$

where $i$ indexes the respondent, $c_i$ indexes the survey cluster the $i$th respondent belongs to, $X_i$ is the $k \times 1$ covariate vector, and $\beta$ is the $k \times 26$ regression parameter matrix. As discussed above, here $k$ is larger than the actual number of variables, denoted by $\tilde{k}$, in the model, because each categorical variable requires multiple dummy variables to represent its levels; in the current setting, we have $\tilde{k} = 19$ but $k = 39$. The clustering variable $W_c = (W_c^{(1)}, \ldots, W_c^{(26)})^\top$ is assumed to have independent normal components: $W_c^{(j)} \sim N(0, (\alpha^j)^2)$, $j = 1, \ldots, 26$ and for all $c$'s, and the covariance matrix is assumed to have the form

$$\Sigma = \left( \begin{array}{cc} \Sigma_{11} & 0 \\ 0 & \Sigma_\rho \end{array} \right),$$

where $\Sigma_{11}$ is an arbitrary $13 \times 13$ correlation matrix (since probit model assumes unit variance), but $\Sigma_\rho$ is a correlation matrix with a single common correlation $\rho$. That is, the response behavior latent variable, $\{Z^{(j)}\}_{j=1}^{26}$, are allowed to be correlated with each other but with a common correlation, and they are independent of the service use latent variable, $\{Z^{(j)}\}_{j=1}^{13}$. The common correlation assumption on $\{Z^{(j)}\}_{j=1}^{13}$ helps to reduce the number of parameters and hence model (and computational) complexity; in any case, information about individual correlations among these $Z$'s are too weak to yield stable estimates. The use of $W$ helps to model the cluster effect due to survey design, by allowing respondents in cluster $c$ to share the same $W_c$.

### 2.3 Prior Specifications

As part of the model, we need to specify our prior distribution for $(\beta, \Sigma_{11}, \rho, \alpha)$. Assuming that they are a priori independent, an assumption common for this type of parameters, we only need to specify the marginal distributions for these four sets of parameters. First, we used a rather strong prior for $\rho$, that is, $\rho \sim \text{Beta}(9, 4)$, as shown as the red curve in Figure 2. This prior puts the middle 50% mass of $\rho$ between roughly (0.6, 0.8), reflecting our strong belief that someone who tends to underreport on one service question also tends to underreport on the rest. The prior for $\Sigma_{11}$ is the correlation matrix resulted from an inverse Wishart matrix, with identity precision matrix and 16 degrees of freedom (d.f.), where the choice of $d = 16$ was somewhat arbitrary but small enough to reflect our desire to be non-informative (recall the matrix is 13 by 13). We denote such distribution by,

$$p(\Sigma_{11}) \sim \text{Inv-Wish-Cor}(I, 16). \tag{3}$$

For more details about inverted Wishart distribution, see [3] and [10]. It turns out that the posterior distribution of $\rho$ and $\Sigma_{11}$ are not very sensitive to the choice of prior distribution, because data provides reasonably strong information on them. For example, as we see from Figure 2, the posterior for $\rho$, the black curve, essentially lives in an area that has very little prior mass. The evidence of a
positive correlation \( \rho \) is strong, but it’s magnitude (around 0.15) is much less than what we a priori believed. This is in some sense the most useful part of a Bayesian inference, learning something from data that is very different from our a priori knowledge or believe.

Figure 2: Prior Vs Posterior of \( \rho \)

In comparison, the prior distribution of \( \beta = (\beta_1, \ldots, \beta_{26}) \) need to be more carefully chosen, where \( \beta_i \)'s are \( k \) by 1 vectors (recall \( k \) is the number of variables in \( X \), including dummy variables). In particular, a constant prior distribution on \( \beta \) will result in an improper posterior distribution. Take a simple case discussed in Section 2.1 for example, where there is only one service use and no background variables. The posterior distribution with a constant prior is

\[
p(\beta_1, \beta_2|y) \propto \Phi_{\beta_1}^{m_0} \cdot (1 - \Phi_{\beta_1})^{n_0-m_0} \cdot (\Phi_{\beta_1} \Phi_{\beta_2})^{m_1} \cdot (1 - \Phi_{\beta_1} \Phi_{\beta_2})^{n_1-m_1},
\]

where \( \beta_1 \) is the mean of the service use latent variable, \( \beta_2 \) is the mean of the response behavior latent variable, \( \Phi(t) \equiv \Phi(t) \) is the cdf for \( N(0, 1) \), and \( m_i \) and \( n_i \), \( i = 0, 1 \), are as defined in (1). Note that as \( \beta_2 \to \infty \), the right hand side of (4) converges to \( \Phi_{\beta_1}^{m_0} \cdot (1 - \Phi_{\beta_1})^{n_0-m_0} \cdot (\Phi_{\beta_1})^{m_1} \cdot (1 - \Phi_{\beta_1})^{n_1-m_1} > 0 \). Thus, the posterior distribution is improper.

For this simple situation, if we want to be “non-informative”, we can put a uniform prior on the service-use rate \( p = \Phi(\beta_1) \) and the truthful-response probability \( r = \Phi(\beta_2) \) respectively. We can easily achieve this by assuming \( (\beta_1, \beta_2)^T \sim N(0, I_2) \). However, for more complicated situations, such natural prior distribution does not always exist. In particular, when the covariate matrix \( X = (X_1, \ldots, X_n) \) has more than one distinctive column, it is not always feasible to find a distribution on \( \beta_j \) such that \( \beta_j^T X_i \sim N(0, 1) \) for all \( i = 1, \ldots, n \).

Nevertheless, we can find an approximate solution, and in fact a better solution by taking into account real prior knowledge on the magnitude of the service use rates. To proceed, we first assume
that all the continuous covariates in $X$ have been standardized (across the sample $i = 1, \ldots, n$) to have sample mean 0 and sample variance 1. For any dummy variable, it is standardized by a scaler multiplier such that the sum of squares across the entire sample is $n$, the sample size. Note that these standardizations of $X$ are for convenience and will not change the nature of the model. For each $\beta_j$ we choose an independent multivariate normal prior distribution with covariance $\tilde{\Sigma} = \frac{n}{k+1} (XX^\top)^{-1}$. The prior mean is chosen as the following:

$$\beta_j \sim \begin{cases} N(\mu_s, \tilde{\Sigma}), & j \leq 13; \\ N(0, \tilde{\Sigma}), & j > 13, \end{cases}$$

where $\mu_s^\top = (-0.8, 0, \ldots, 0)$ is $k$ by 1 and only the first element—corresponding to intercept—is not 0. The value -0.8 is chosen to approximately match our prior mass allocation with psychologists’ general knowledge. In particularly, under this choice, the median service rate is about 22% and the inter-quartile range is about 37%, in the absence of any covariances. (We also tried the prior with all prior means set to zero, and did not find any significant changes in our imputation results.) For the response behavior, we choose zero prior mean so the resulting prior on the response probability is approximately uniform, reflecting our weak prior knowledge.

The prior distribution of $\alpha^{(j)}$ is set to be $N(0, 1)$. Note that the sign of $\alpha^{(j)}$ is not identifiable. But this does not affect our imputation because only $(\alpha^{(j)})^2$ enters the model. It is for computational convenience and speed that we let $\alpha$ live on the whole real line instead of $\mathbb{R}^+$, as discussed in [9].

### 2.4 Initial Results and A Hidden Defect

Under model specified above, Table 3 shows the average of 10 imputations for the Latino samples. The bold faced services are combined services. For instance, “specialist” means whether a respondent ever used at least one of the “psychologist”, “psychiatrist”, “other mental health professionals”, or “hot line”. For each ethnicity group, three columns are provided. The first column is the observed service rates from the new design group (“New”). The second column has the imputed average service rate for the traditional design group using our multi-variate probit model (“Imp”), and the third column is the observed service rate from the traditional design group (“Old”).

Ideally, the rates in the second column should be similar to those in the first column. Of course we do not expect a perfect match, nor would that be desirable as that would be a reflection of over-fitting by our model. Indeed, even if the new and traditional designs do not induce systematic differences in responses, we would expect to see differences between the 25% samples and the 75% samples due to sampling variations. What we are looking for here, therefore, are systematic patterns of differences between the first and second columns. There are no obvious such differences for the individual service rates. However, we observe a disturbing pattern for the last three rows, namely, the three most aggregated “any rates” the table shows, “Formal Service (1-8)”, “Any Services (1-10)”, and “Any Services (1-13)”. We see the imputed rates (“Imp”) for those rows are uniformly higher than the observed rates in the new design group by $1 - 5\%$. In the next Section we explore the source for this systematic bias and report our effort in remedying it.

3 Further Modeling Effort Via the Continuation-Ratio Strategy

3.1 The need of modeling high-order interactions

Since our multivariate probit model performed reasonably in imputing individual service rates, it is natural to suspect that the observed systematic over-imputation of the “any rates” is due to the failure of the multivariate probit model in capturing interactions. Indeed, since for any Bernoulli variables \{S^{(1)}, \ldots, S^{(m)}\},

\[
E \left[ \max_{j=1}^{m} S^{(j)} \right] = \sum_{j=1}^{m} E \left[ S^{(j)} \right] - \sum_{1 \leq i < j \leq m} E \left[ S^{(i)} S^{(j)} \right] + \ldots + (-1)^{m} E \left[ \prod_{j=1}^{m} S^{(j)} \right],
\]

we see in order to correctly capture the combined rate of \(m\) services, which is represented by \(\max_{j=1}^{m} S^{(j)}\), our model needs to capture high-order interaction among the \(S^{(j)}\)'s. This need immediately points to a fundamental defect of multi-variate probit model. Since it relies on the underlying multi-variate normal distribution to capture interactions, it essentially only has enough free parameters to model two-way interactions – a \(m \times m\) correlation matrix can only model \(m(m-1)/2\)
pair-wise correlations. In contrast, an \( m \)-way contingency table with all binary factors potentially can have as many as \( 2^m - 1 \) free parameters. Consequently, while multi-variate probit model offers both modeling and computational simplicity, as well as intuitive interpretation, it is not a sensible model when there are significant interactions higher than two-way interactions.

We therefore needed to seek an extension of the multi-variate probit model to accommodate higher order of interactions. There are, of course, many models of this sort, such as log-linear models. We, however, needed an extension that essentially can maintain the computational simplicity of the multi-variate probit model for Bayesian inference, because of various recourses constraints we faced (e.g., a test run often takes days on the available computers and we needed to do many of them due to revisions of database, modifications of variables, refinements of priors, etc). In this quest, we initially were motivated by the hierarchical structure of psychiatric services. That is, the individual services belong to some general types of services, for example, the four types correspond to the first four bold faced variables in Table 3. It is therefore reasonable to postulate a “two-stage” indicators for using a particular service, that is, in order use a specific service, say, psychologist, one has to first be in the category of needing seeing a “Specialist” and then choose or be assigned to seeing a psychologist.

Specifically, suppose the 13 services can be divided into \( L \) types. Let \( S_l \) be the indicator for \( l \)th type, \( l = 1, \ldots, L \). Within each type, suppose there are \( J_l \) services. Then for any service belong to the \( l \)th type, we can express our service indicator as

\[
S(j) = S_l \cdot S_{l(j)},
\]

where \( S_{l(j)} \) is the \( j \)th service within \( l \)th type, and \( \{S_{l(j)}, j = 1, \ldots, J_l\} \) and \( S_l \) are assumed to be independent. We note that the expression in (6) is a special case of the continuation ratio (CR) model set up, that is, the probability of the binary outcome is modeled as a product of a sequence of binary probabilities (hence continuation ratio); this is a common strategy in model censored survival data, see, for example, [18], [7], [22], [1], [13], [5], [6], and [21].

For NLAAS, we have four natural groups, as determined by the usual categorization in practice. That is, we have generalist (max\( \{S(2), S(3), S(8)\} \)), specialist (max\( \{S(1), S(4), S(7), S(11)\} \)), human services (max\( \{S(5), S(6), S(9)\} \)) and other services (max\( \{S(10), S(12), S(13)\} \)). Our initial hope was that such a grouping will capture most of the higher order interactions, as we imagined that there would be a positive correlation among the types (e.g., a generalist may recommend a patient to see a specialist) and possibly a negative correlation within the same type (e.g., once a patient sees a specialist, s/he might have a less need to see another one). Somewhat surprisingly, this extended model did not lead to noticeable correction for the over imputation for the three "any rates". This suggested to us that the latent structure given by (6) needs to be further extended in order to capture the complex high-order interactions.

### 3.2 A Continuation-Ratio Multi-probit model

To further extend the model, we drop the restriction on sharing the same-type indicator and adopt the more general product-form of the continuation ratio model by letting

\[
S(j) = S_a(j) \cdot S_b(j)
\]

where \( \{S_a(j), j = 1, \ldots, 13\} \) and \( \{S_b(j), j = 1, \ldots, 13\} \) are jointly independent. While such a relaxation will lead to better fit to the observed data because of the increased number of parameters
in the resulting model, we must be mindful about its interpretation and the potential issues of over-fitting and non-identifiability. As for the interpretation, it is less appealing than that of (6), but nevertheless one can consider (7) as an attempt to let the data to decide which “type” a service belongs to (e.g., the grouping in Table 3 is not written in stone; for example, it is not that clear if “Hot Line” should be grouped with the “Specialists” or “Human Services” or even “Alternative Service”). That is, if we view one of the two $S$’s as a “type” indicator, then one can imagine that a posterior inference might indicate a block correlation structure where a subgroup of the same $S$’s are highly correlated with each other, indicating that these service tend to be “grouped” together – indeed, in the extreme case where all indicators in the same “group” are the same, then we are back to the special case of (6).

The above interpretation would make sense, however, only when we have ways to identify which of $S$’s in (7) can be associated with type and which with the more specific individual services. As it stands by itself, $S_a$ and $S_b$ are completely symmetric, and hence we have a fundamental non-identifiability problem. To deal with this problem, we assumed that the type-indicator $S_a$ is common to all respondents, that is, its distribution does not depend on the covariates $X$. This of course is a very strong assumption and likely it is not true. However, given the non-identifiability issue of the model without such a strong assumption, we could justify this choice as our attempt to model common correlations among the service uses via the hidden “type” variables, rather than capturing how the “type” probabilities vary with the covariates. This latter dependence is indistinguishable, based on our observed data, from the dependence on covariates of the conditional probability of using a specific service within each type, and hence we may as well leave it to $S_b$.

Specifically, as with the probit-model setting in Section 2, we model each of $S_a$ and $S_b$ via 13-dimensional multi-variate normal variables $Z_a$ and $Z_b$, respectively, by letting $S_{a/b}^{(j)} = 1 \{ Z_{a/b}^{(j)} \geq 0 \}$ for all $m = a, b$ and $j = 1, \ldots, 13$. The latent variables $Z_a$ and $Z_b$ are assumed to be independent, and have the following distributions:

$$
Z_a \sim N(\mu_a, \Sigma_a) \\
Z_b \sim N(\beta_X^\top X + W, \Sigma_b)
$$

where both $\Sigma_a$ and $\Sigma_b$ are unconstrained correlation matrices. The model for the response indicator $\xi$ is identical to that of (2).

The model as specified may still suffer non-identifiability because the design matrix $X$ in the regression model for $Z_b$ allows a constant (i.e., intercept) term, which would be in confounding with the $\mu_a$ term for $Z_a$. We resolved this problem by putting a relatively strong prior for $\mu_a$:

$$
\mu_a \sim N(2, I_{13}/200)
$$

The choice of the prior mean “2” is a bit arbitrary but guided by our desire that the “type” indicator $S_a$ should be not too far from the that for the standard probit model, which is equivalent to setting $S_a \equiv 1$ in (7). In any case, the significant part of the new model is the introduction of $\Sigma_a$, which leads to a quite enriched model for high-order interactions.

To see this empirically, we fit both the original probit model and the CR-probit model, and compare the posterior predictive means of the three combined services, as in Figure 3. The left panel is for the life-time use, and the right panel is for the past 12-month service uses, a topic we shall discuss in the next sub-section. As the left panel of Figure 3 shows, the imputed rate from the new CR-probit model (red) is always closer to that of the observed rate from the new design group (black) than the imputation from the original probit model (green); the observed rates from the old
design group (blue) of course, is always much lower. More importantly, we no longer observe the uniform over-imputation by the green bars in the top plot for the life-time use. For the right pane, the green bar did not uniformly over-impute to start with (an interesting issue we will discuss in the next sub-section), but nevertheless the red bars are still closer to the black bars in majority cases.

We want to emphasize that the CR-probit model does not only change the imputed rates for the combined services, but also rates for individual services. But as a comparison of Table 3 with Table 4, we see the changes are not as significant or systematic as the changes on the combined rates.

![Figure 3: New Model Vs Old Model](image)

3.3 A CR model for the Last 12 Month Service Use

In addition to the lifetime services, NLAAS also collected data on the last-12-month service, for which similar underreporting is observed for the traditional-design group, as shown in Table 5.

14
Table 4: Percentage rates of Latino lifetime service use. New: observed new design rates. Imp: imputed old design rates. Old: observed old design rates.

When setting up the last-12-month service model, we need to consider the following natural logical constraint: someone who did not have a service in lifetime should not have such service in the last 12 month either, a logic that is respected in the observed data. Therefore, for each service, we have a bivariate random variable \((S, S_T)\) that can only take three possible values: \(\{1, 1\}, (1, 0), (0, 0)\), where \(S_T\) is the true last-12-month service use. This bivariate variable can be modeled as \((S, S_T) = (S, S\tilde{S})\), where \(S\) and \(\tilde{S}\) are independent Bernoulli random variables, with \(\tilde{S}\) being the indicator for the true last-12-month service use given the lifetime service use \(S = 1\). In other words, our joint model for the self-reported life-time uses \(y\) and the past 12-month use \(y_T\) will be formulated in two stages, first the marginal model for the life-time use and then the conditional model of the past 12-month give the life-time use. This is a natural CR modeling strategy given the NLAAS questionnaire structure, where in the interview if the response to the lifetime service questions is negative, then the last year question is skipped and the reported value is set to be 0. Thus the reported last-12-month service use does not provide any additional information about the true lifetime service in additional given the self-reported lifetime service.

Under our most basic model assumptions as listed in the beginning of the Section 2.1, for the traditional design group, if the observed service use is \((y, y_T) = (1, 1)\), then \((S, S_T) = (1, 1)\). If \((y, y_T) = (0, 0)\), then \(\tilde{S}\) is missing, because the respondent was never asked about the last year service use. When \((y, y_T) = (1, 0)\), for which the respondent was asked about her/his last-12-month service use, s/he could either underreport or really does not have service use in the last
12 months. In the last case, we introduce a new respondence behavior indicator for the last-12-month service use, $\tilde{\xi}$, such that it is independent of $(S, \xi, \tilde{S})$ and that the reported last-12-month service use is expressed as $y_T = S \xi \tilde{S} \tilde{\xi}$. That is, $\tilde{\xi}$ is a direct analog to $\xi$ in the lifetime service use model. Furthermore, since the likelihood for $(S, \xi)$ and $(\tilde{S}, \tilde{\xi})$ factors under our two-stage modeling, they are a posteriori independent under independent prior. This simplicity allows us to use independent Markov chains, as we shall discuss in detail in the Section 4, to sample from the posterior distributions and thereby reducing computational burden.

The conditional model for the last-12-month service uses is a complete analogue of the life-time model, except for that we use multi-probit model for $\tilde{S}$ instead of the CR-probit model. The key reason for this is that we did not encounter an over-imputation problem for the combined rates with the past 12-month services as with the life-time uses when using the original probit model. Given the simplicity of the multi-variate probit model, we choose to stay with it. As a result, the red bars in the right plot of Figure 3 cannot be uniformly closer to the black bars than the green bars, because the red bars represented the imputation under the improved model but the improvement here is through the life-time service part only (recall $S_T = S \tilde{S}$). So the fact that the imputed rates from the life-time use was uniformly smaller under the CR-probit model means that the corresponding past-12-month rates, which are a product of the unchanged conditional rates and the life-time use rates have to be uniformly smaller than that from the original probit model as well. Consequently, since the over-imputation for the combined rates occurred for most items but not all, for those rates that over-imputation under the original model did not occur, the imputed rates under the CR-model will be further away from the observed rates under the new design group. But the overall improvement is evident in Figure 3, and as before the changes of the individual rates are not significant. Table 6 show the results for the last-12-month use, corresponding to the life-time service use as given in Table 4.

We, however, believe it is worthwhile to comment on the interesting phenomenon that the high-order interactions (higher than two ways) appear to be less pronounced for the self-reported past-12-month uses than the life-time service use. One possible explanation is that memory decay has played an important role for inducing a higher order interaction. We can imagine that in the presence of an incentive for underreporting, if a respondent could not be sure of whether a particularly service was actually rendered and hence reported negatively, it is more likely for the respondent to more conveniently associate the same “negative” memory to other services possibly occurred in a distant past than in the most recent history. If this speculation is true, then the “over-imputation” by the original multi-variate probit model for the combined rates might actually be compensating—accidently of course—for the artificially high-order interaction induced by the interaction of memory decay and incentive for underreporting. Of course, without further data on measuring the memory decade, this will remain as a speculation.

4 Computation Via MCMC

4.1 An Outline of Our Algorithm

The posterior sampling from the models presented in the previous were done via Markov chain Monte Carlo (MCMC), more specifically a Gibbs-type sampler, with some steps implemented via the Metropolis-Hasting algorithms. The exact details are a bit tedious, so here we only outline the major steps, via specifying the full conditional distributions corresponding to Gibbs steps. We also
New Design | Old Design
--- | ---
Psychiatrist | 32.9% | 22.9%
Other Medical Doctor | 59.5% | 31.4%
Psychologist | 28.6% | 17.4%
Social worker | 43.7% | 17.6%
Counselor | 21.9% | 21.3%
Other Mental Health Prof | 50.6% | 45.0%
Nurse, Occupational Therapist, | 56.1% | 17.8%
Religous/Spiritual Advisor | 41.6% | 29.8%
Hot Line | 19.3% | 18.8%
Other Healer | 53.8% | 41.1%
Internet Group or Chat Room | 64.0% | 31.8%
Self Help Service | 22.7% | 26.9%

†all numbers are in $10^{-2}$ scale.

Table 5: Comparing Last Year Service Use: the rate is the percentage of people reported having service last year among those who reported having service use in lifetime.

only do so for the CR-probit model, since the initial probit model is easier.

According to the model specifications in Section 3.2, the conditional distributions and sampling techniques are as follows:

- Sample from $P(Z|\beta, \Sigma_{11}, \rho, \alpha, W)$, a truncated multivariate normal. Draw each triple $(Z_a^{(j)}, Z_b^{(j)}, Z_\xi^{(j)})$ one by one given the rest, where $Z_a^{(j)}$, $Z_b^{(j)}$, and $Z_\xi^{(j)}$ denote the latent variables for $S_a^{(j)}$, $S_b^{(j)}$, and $\xi^{(j)}$ respectively. Since $(Z_a^{(j)}, Z_b^{(j)}, Z_\xi^{(j)})$ are jointly independent given the rest, $(Z_a^{(j)}, Z_b^{(j)}, Z_\xi^{(j)})$ is distributed as independent truncated normal. Here $j$ indicates different services.

- Sample from $P(\beta|Z, \Sigma_{11}, \rho, \alpha, W)$, multivariate normal distribution. The mean and covariance can be computed by following linear regression routine combined with the prior distribution we specified in Section 2.3.

- $P(\Sigma_{11}|\beta, Z, \alpha, W)$, draw matrix $\Gamma$ from $Inv\text{-}Wish(S, 13)$, where $S = (Z - \beta^T X)(Z - \beta^T X)^T$. Treat $\Gamma$ as a variance-covariance matrix, and $\Sigma_{11}$ is the corresponding correlation matrix.

- Sample $P(\alpha|\beta, \Sigma_{11}, \rho, Z, W_k)$, multivariate normal distribution. This is similar to the sampling of $P(\beta|Z, \Sigma_{11}, \rho, \alpha, W)$.

- Sample $P(W|\beta, \Sigma_{11}, \rho, Z, \alpha)$, multivariate normal distribution.

- Updating $\rho$ by directly Metropolis-Hasting step is very inefficient. Therefore we propose the following solution. Let $\Sigma_1 = \rho 11^T$ and $\Sigma_2 = (1 - \rho)I$, and define $\eta$ as

$$
\eta_i = \mu + \sqrt{\rho} \begin{pmatrix} \tilde{Z}_i \\ \vdots \\ \tilde{Z}_i \end{pmatrix}, \quad Z_{i,\xi} | \eta_i \sim N(\eta, \Sigma_2),
$$
<table>
<thead>
<tr>
<th>Specialty</th>
<th>Puerto Rican</th>
<th>Cuban</th>
<th>Mexican</th>
<th>Other Latino</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>Imp</td>
<td>Old</td>
<td>New</td>
</tr>
<tr>
<td>Specialist (1,4,7,11)</td>
<td>11.4</td>
<td>12.9</td>
<td>7.8</td>
<td>5.7</td>
</tr>
<tr>
<td>1. Psychiatrist</td>
<td>10.0</td>
<td>7.9</td>
<td>5.1</td>
<td>5.1</td>
</tr>
<tr>
<td>4. Psychologist</td>
<td>3.9</td>
<td>5.5</td>
<td>2.5</td>
<td>2.7</td>
</tr>
<tr>
<td>7. Other M.H. Prof.</td>
<td>3.4</td>
<td>3.6</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>11. Hot Line</td>
<td>0.7</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Generalist (2,3,8)</td>
<td>21.3</td>
<td>18.8</td>
<td>6.2</td>
<td>14.2</td>
</tr>
<tr>
<td>2,3. Other Med. Doc.</td>
<td>21.3</td>
<td>18.2</td>
<td>6.4</td>
<td>13.3</td>
</tr>
<tr>
<td>7. Other M.H. Prof.</td>
<td>4.8</td>
<td>3.2</td>
<td>0.3</td>
<td>2.3</td>
</tr>
<tr>
<td>8. Other Professionals.</td>
<td>16.5</td>
<td>14.0</td>
<td>4.6</td>
<td>7.0</td>
</tr>
<tr>
<td>Human Services (5,6,9)</td>
<td>6.2</td>
<td>5.3</td>
<td>1.8</td>
<td>3.4</td>
</tr>
<tr>
<td>5. Social Worker</td>
<td>6.8</td>
<td>6.3</td>
<td>3.2</td>
<td>4.8</td>
</tr>
<tr>
<td>6. Counselor</td>
<td>8.6</td>
<td>7.6</td>
<td>1.7</td>
<td>3.5</td>
</tr>
<tr>
<td>9. Religious Advisor</td>
<td>11.5</td>
<td>8.7</td>
<td>3.3</td>
<td>2.5</td>
</tr>
<tr>
<td>Alt Services (10,12,13)</td>
<td>7.9</td>
<td>4.0</td>
<td>0.4</td>
<td>2.5</td>
</tr>
<tr>
<td>10. Other Services</td>
<td>1.1</td>
<td>3.2</td>
<td>1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>12. Internet</td>
<td>2.6</td>
<td>3.1</td>
<td>1.6</td>
<td>0.0</td>
</tr>
<tr>
<td>13. Self Services</td>
<td>25.9</td>
<td>26.7</td>
<td>14.0</td>
<td>18.0</td>
</tr>
<tr>
<td>Formal Services (1-8)</td>
<td>30.8</td>
<td>30.2</td>
<td>14.9</td>
<td>18.6</td>
</tr>
<tr>
<td>Any Services (1-10)</td>
<td>31.9</td>
<td>31.6</td>
<td>16.2</td>
<td>18.6</td>
</tr>
</tbody>
</table>

Table 6: Percentage rates of Latino last year service use. New: observed new design rates. Imp: imputed old design rates. Old: observed old design rates.

where \( \tilde{Z}_i \) is univariate standard normal distribution. Note that under this setting, \( Z_{i,\xi} \) is multivariate normal with mean \( \mu_i = \beta^T X_i + W_{c_i} \), and variance \( \Sigma = \Sigma_1 + \Sigma_2 \). Observing that moving \( \eta \) and \( \rho \) together is much more efficient than updating \( \rho \) alone, we adopted the following strategy. First, we draw \( \eta_i | (Z_{i,\xi}, \rho, \beta) \), which is normal distribution; denote \( \eta = (\eta_1, ..., \eta_n) \). Second, we update \( (\rho, \eta) \) by Metropolis-Hasting algorithm. Let \( (\rho, \eta) \) be the current state, and propose \( (\rho', \eta') = \lambda(\rho, \eta) \), where \( \lambda \sim \chi^2_{30}/30 \). If \( \rho > 1 \), reject \( (\rho', \eta') \); otherwise, accept \( (\rho', \eta') \) with probability \( \min \{ p(\rho', \eta' | \beta, \Sigma_{11}, \alpha, W, Z) d(\lambda^{-1}) \} \), where \( d(\cdot) \) is the density of \( \chi^2_{30}/30 \) and \( p(\cdot, \cdot | \beta, \Sigma_{11}, \alpha, W, Z) \) is the conditional density of \( (\rho, \eta) \), which is a multivariate Gaussian density.

### 4.2 Convergence Diagnostics and Computation Time

In order to provide 10 imputed data sets, we ran 10 independent Markov chains, as outlined above. The starting points of the latent variables are independent truncated Gaussian, truncated to obey the sign restrictions from our observed service indicators, with mean zero and variance one. Each chain contains 10,000 iterations, which took about 30 hours to on a 3.0 GHz computer. The last state from each chain as used as one imputation.

The iteration number 10,000 was decided based on Gelman and Rubin’s \( \hat{R} \) [12] and graphical diagnostic checks (e.g., comparing the sampling distributions of some low dimensional quantities for which we have independent ways to approximate their posterior distributions). For the Gelman-
Rubin statistic \( \hat{R} \), we computed it for the imputed rates of 13 services across 9 different ethnicity groups, resulting 117 \( \hat{R} \)'s in total. By discarding the first 500 iterations and taking a 500-skeleton (that is, a chain of length 10000/500 = 20), 114 \( \hat{R} \)'s out of the 117 variables were below the common cut-off 1.1 and the other 3 were greater than 1.1 but less than 1.12. The reason that we take 500-skeleton instead of the whole chain was due to lack of storage of memory in the computer we used. Note that in usual applications of MCMC, it would be a complete waste to through away so many draws. In our setting, however, our goal was to only create 10 imputations, and once they are released, they will become permanent and will affect all subsequent analysis. We therefore want to make sure that they are of as good quality as possible in the sense of being genuine independent draws from the posited posterior distribution.

5 Imputation Quality Checking

5.1 The need and difficulties of assessing imputation quality

Once multiple imputations are created, one natural question is how good are they? This question in general is hard to answer definitively, since one usually does not have a “Golden Standard” to check against. In our current content, however, we do have the observed rates from the new design group, and since our goal was to impute the old design group to make it “looks like” the new group, obviously we do not want to see that the imputed rates to be very different from the corresponding observed rates from the new-design group. If that happens, we may suspect that there is a serious failure of the imputation model or errors in the computation (e.g., the MCMC failed to converge).

With this in mind, one might be tempted to cast this as a usual two-group comparison problem, and perform some two-sample tests to see if there is a significant difference. It turns out that the problem is far complicated than it might been seen at the first for the following reasons. First, since our goal here is to ensure the validity of the subsequent analysis on whatever sub-populations of interests to a potential user of the data, ideally we need to ensure the quality of imputation for any sub-population. This is simply an unavailable goal because the number of possible sub-populations is astronomical with 5000 variables in place. A more workable goal is to develop a “data-mining” strategy that would allow us to comb through the imputed data sets to flag those sub-populations where there may be a serious discrepancy between the observed rates from the new design group and the imputed rates. This would allow one to examine these sub-populations in detail to identify the potential causes of the discrepancy, or minimally set a warning for any potential users who want to study these sub-populations. Unfortunately, as far as we are aware of, this is largely an unexplored territorial, with few existing systematic approaches we can apply directly. Our own attempt to develop such a method has been largely hindered by the lack of a workable numerical criterion to measure the discrepancy, for reasons given below.

Second, we need to have a standard on what are the statistically acceptable discrepancies and what are not, even if we have a way to comb through the imputed data sets effectively. One might find it is odd for us to emphasize this, because surely we all understand the importance of distinguishing biases from variances. After all, even if everyone reported truthfully and the survey protocol was followed perfectly, we will still see “discrepancies” of the sample estimates between two randomly divided sub groups of the population. But the reason we emphasize this is that, unlike the case where two rates being compared are from two independent samples, in our current case the imputed rate is highly correlated with the observed rate from the new-design group because the imputation was done precisely by using the data from the new-design group. However, there
does not seem to be an easy way to estimate reliably this correlation because the dependence of imputed rates is a highly complex function of the observed-data from the new group. The usual strategies of using asymptotics or bootstrap have their serious limitation in this context. As we shall demonstrate, problems often occur with small sub-populations, which often means that the sample sizes are too small to justify asymptotics. Furthermore, to derive asymptotics will require the knowledge of the imputation model, and therefore it is typically not feasible for a user to employ any such method to check the quality of the imputation for the sub-population of his/her interests. The bootstrap is simply too costly to use in general, because it would require the re-imputation after each bootstrap sample, in order to capture the real variations.

Third, perhaps the most problematic problem at all, is that the discrepancies one observe include all “defects” in the original randomization. While randomization ensure balanced background variables between the two groups overall, as illustrated in Table 2, for some small sub-populations, it is entirely possible that the two groups are very unbalanced, just by chance. When this occurs, as we shall illustrate below, it can lead to serious discrepancies between the observed rates from the new group and the imputed rate for the old group. It would be very misleading, however, to attribute this as an failure of the imputation model. On the other hand, we know no imputation model can be perfect, and any its potential defect indeed will likely be more easily revealed for smaller sub-populations because there is not enough data to “over-power” the defects in our model (e.g., a prior specification). It is then an exceedingly challenging issue to disentangle the disparities caused by the failure of the randomization and that by the failure of the imputation model within small sub-populations.

We currently do not have effective strategies to deal with these problems, but nevertheless we made an attempt to at least examine these issues, as we report below, in hoping to stimulate more research in these areas.

5.2 Checking for sub-populations

In Table 4 and Table 6, we compared the imputed rates and the observed rates across different Latino ethnicity groups. In this section, we further explore our imputation results by comparing the rates in strataums according to different variables. In particular, we exam gender, insurance type, and major depression, among which gender and insurance type are predictors included in the imputation model, but major depression (MDE) was not included. Figure 4 shows the comparison in male and female groups. The results look similar to those by ethnicity groups. This is as expected since gender is one of the indicators.

Serving as a public data set, a potential analysis may include variables that were not used in our imputation model. Therefore, it is more important to make the same type of comparisons for sub-populations that are formed by stratifying on variables that were not included the imputation model. Figure 5 shows the observed rates and imputed rates for both the MDE positive and negative cases. Though MDE was not part of the model, the comparative results are quite similar to that of Figure 4, with no obvious patterns of over or under imputations. We believe that our imputation model did well for this stratification because it already took into account a significant number of predictors that are highly correlated with MDE. Indeed, the hope with any imputation model, which typically cannot include all variables as we want to, is that those variables included in the model would capture the essences of all important covariates for the outcome variable that is being imputed.

There is no guanantee, however, that there is no large “discrepancies” for sub-populations that
are even stratified according to a variable that is included in the model, let alone for those that are not included. For instance, we “discovered” this when comparing the rates by different insurance types. NLAAS documented six types of insurance:, that is, not insured, private through employer, private purchased, medicare, medicaid, and other insurance. Figure 6 shows the comparison of the imputed service rates and the observed new design group rates by insurance types. We see that out of the six sub-populations, the comparative results from the group of Private Purchased, Medicare, and Others, some of the imputed rates are substantially higher than the observed rates from the new-design group, with the other group being most extreme, with some imputed rates more than double the observed rates from the new-design sub-group.

So what went wrong? Is this an evidence of the failure of our imputation model? To answer these questions, let us first look into a few more obvious facts. First, Table 7 gives the sample size and effective sample sizes of all size insurance groups, and we noticed that the three most problematic groups correspond to the three smallest samples sizes or effective sample sizes. We emphasize here that because the variations in survey weights from NLAAS are very large, it is important to calculate the effective sample size, which we used the common approximation.

\[ n_{eff} \approx \frac{n}{1 + \frac{S_W^2}{\bar{W}^2}} \]

where \( n \) is the nominal sample size, and \( \bar{W} \) and \( S_W^2 \) are the sample mean and sample variance (from
the sub-population of interest) of the survey weights. As we see from Table 7, the effective sample sizes are significantly smaller than the nominal sample sizes, making the instability problems with small sub-populations particularly serious.

Perhaps more disturbingly, we observe that whenever our model “over imputes”, the corresponding observed rates from the traditional design rates are higher than those of the observed new design rates. This is expected since our model setup is such that the imputed rates are designed to be higher than those from the traditional-design groups, which we assume underreport the service uses. So is it a contradiction to our model assumption of under-reporting by the old-design group when the observed rates from the old-design group is higher than that from the new-design group?

Not necessarily. The comparison to the new-design group reveals underreporting of the old-design group only when the two groups are comparable to start with. We relied on the randomization to achieve this comparability. However, there is no guarantee especially on small sub-populations that the randomization has worked perfectly (even assuming the survey protocols have been followed perfectly, which is never the case in practice). If the old-design group started with a very high service rate compared to a new-design group, then even with the under-reporting, it can still ends up with significantly higher self-reported rates, causing the false impression that under-report may not have occurred.

Indeed, there is strong evidence that this is what has happened here. When we look into some diagnostic depression variables included in our model (number of disorders, any disorder, anxiety disorder, substance disorder), we discovered that the old-design group has much higher values on these important predictors. Table 8 shows the sample weighted averages of these variables for the

Figure 5: Lifetime Service by Major Depression
other group. We see that old design group has a much higher disorder rates and more disorders per person than the new design. In the most extreme case, there was no reported cases of substance disorders in the new design group, a good reminder the large variability due to small (effective) sample sizes, but 17% estimated rate from the old-design group. Such variables are known to be highly correlated with the psychiatric service uses. As an illustration, we ran a logistic regression for any service $\sim$ anxiety disorder + substance disorder. The estimated regression coefficients and the standard errors are 1.64413 (0.23144) for substance disorder and 2.13228 (0.11465) for anxiety disorders, which are highly significant.

Furthermore, the two groups exhibit noticeably different patterns of sample weights, which also contribute to the differences. First, Table 8 shows that the correlation between the survey weights and three "any rates" are much higher in the old design group than those in the new design group. Second, Figure 7 shows box plot of the survey weights against any services. There are two positive responses in the old design group that has very large weights. These two facts together also contributes to the phenomenon of the higher self-reported rates from the old-design group. To see this more clearly, Table 9 compares the weighted and unweighted sample means of the three "any rates" for the other group, where the differences are doubled with the weighted version. This simple comparison between weighted and unweighted averages allows us to have quantitative indication as how much of the problem is due to weight and how much is due to the difference in the background variables. For this case, the examination above led us to believe that both problems are significant.

In general, how to conduct such an examination in a systematical way and how to disentangle them is an important but exceedingly challenging problems.
Figure 7: Weights Vs Service in the “Other” Insurance Category

References


<table>
<thead>
<tr>
<th>Insurance Group</th>
<th>Total</th>
<th>New</th>
<th>Old</th>
<th>Total</th>
<th>New</th>
<th>Old</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Insured</td>
<td>610</td>
<td>163</td>
<td>447</td>
<td>1082</td>
<td>283</td>
<td>799</td>
</tr>
<tr>
<td>Private Through Employer</td>
<td>1006</td>
<td>234</td>
<td>776</td>
<td>2349</td>
<td>588</td>
<td>1761</td>
</tr>
<tr>
<td>Private Purchased</td>
<td>107</td>
<td>25</td>
<td>83</td>
<td>259</td>
<td>61</td>
<td>198</td>
</tr>
<tr>
<td>Medicare</td>
<td>158</td>
<td>50</td>
<td>123</td>
<td>508</td>
<td>113</td>
<td>395</td>
</tr>
<tr>
<td>Medicaid</td>
<td>318</td>
<td>82</td>
<td>237</td>
<td>515</td>
<td>125</td>
<td>390</td>
</tr>
<tr>
<td>Other Ins.</td>
<td>67</td>
<td>19</td>
<td>48</td>
<td>151</td>
<td>36</td>
<td>115</td>
</tr>
</tbody>
</table>

Table 7: Sample Size and Effective Sample Sizes of the Insurance Groups.

<table>
<thead>
<tr>
<th>Weighted Averages</th>
<th>New Design</th>
<th>Old Design</th>
<th>Correlations with Survey Weights</th>
<th>New Design</th>
<th>Old Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Disorders</td>
<td>0.37</td>
<td>0.92</td>
<td>Formal Services</td>
<td>-0.07</td>
<td>0.17</td>
</tr>
<tr>
<td>Any Disorder</td>
<td>0.20</td>
<td>0.22</td>
<td>Any Services 1-10</td>
<td>-0.11</td>
<td>0.17</td>
</tr>
<tr>
<td>Anxiety Disorder</td>
<td>0.11</td>
<td>0.18</td>
<td>Any Services 1-13</td>
<td>-0.11</td>
<td>0.16</td>
</tr>
<tr>
<td>Substance Disorder</td>
<td>0</td>
<td>0.17</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Some Statistics in Other Insurance Group.


25
<table>
<thead>
<tr>
<th></th>
<th>Simple Mean</th>
<th>Weighted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New Old</td>
<td>Old New</td>
</tr>
<tr>
<td>Formal Service (1-8)</td>
<td>0.19 0.30</td>
<td>0.11 0.39</td>
</tr>
<tr>
<td>Any Service (1-10)</td>
<td>0.22 0.31</td>
<td>0.09 0.43</td>
</tr>
<tr>
<td>Any Service (1-13)</td>
<td>0.22 0.34</td>
<td>0.12 0.43</td>
</tr>
</tbody>
</table>

Table 9: Observed rates for Other Insurance Group


26