Supplemental Material

A Computation Via MCMC

The posterior sampling from the models presented previously were done via Markov chain Monte Carlo (MCMC), more specifically a Gibbs-type sampler, with some steps implemented via the Metroplis-Hastings algorithms ([19]). The exact details are tedious, so here we outline only the major steps, via specifying the full conditional distributions corresponding to Gibbs steps. We also do so only for the CR-probit model, since the multivariate probit model is easier.

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

- Sample from $P(Z|\mu_a, \beta_b, \beta_l, \Sigma_a, \Sigma_b, \Sigma_l, \alpha, \tilde{\alpha}, W, \tilde{W})$, a truncated multivariate normal. Draw each triple $Z = (Z_{a,j}, Z_{b,j}, Z_{l,j})$ one by one given the rest, where $Z_{a,j}, Z_{b,j}$, and $Z_{l,j}$ denote the latent variables for $S_{a,j}$, $S_{b,j}$, and ξ_j respectively. Since $(Z_{a,j}, Z_{b,j}, Z_{l,j})$ are jointly independent given the rest, $(Z_{a,j}, Z_{b,j}, Z_{l,j})$ is distributed as independent truncated normal. Here j indicates different services.

- Sample from $P(\mu_a, \beta_b, \beta_l|Z, \Sigma_a, \Sigma_b, \Sigma_l, \alpha, \tilde{\alpha}, W, \tilde{W})$, a multivariate normal distribution. The mean and covariance can be computed by a linear regression routine combined with the prior distribution we specified in Section 2.4.

- Sample from $P(\Sigma_a, \Sigma_b, \Sigma_l|\beta, Z, \alpha, \tilde{\alpha}, W, \tilde{W})$. Notice that conditional on the remaining parameters, Σ_a, Σ_b, and Σ_l, are inverted Wishart random matrices. Hence, we can, for instance, draw matrix Σ_b from Inv-Wish$(S, df_b + n)$, where $S = (Z_b - \beta_b^\top X - W_c)(Z_b - \beta_b^\top X - W_c)^\top + \bar{\Sigma}_b$ and n is the sample size.

- Sample from $P(\alpha, \tilde{\alpha}|Z, \mu_a, \beta_b, \beta_l, \Sigma_a, \Sigma_b, \Sigma_l, W, \tilde{W})$, a multivariate normal distribution. This is similar to the sampling of $P(\beta|Z, \Sigma_{11}, \rho, \alpha, W)$.

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

We ran 10 independent Markov chains. The starting points of the latent variables are independent truncated $N(0, 1)$, truncated to obey the sign restrictions from our observed service indicators. Each chain contains $10,000$ iterations, which took about 30 hours to run on a 3.0 GHz computer. The last iteration from each chain is used as one imputation.

The iteration number $10,000$ was decided based on Gelman and Rubin's $\hat{R}$ (see [11]) and graphical diagnostic checks. The Gelman-Rubin statistic $\hat{R}$ was computed for the regression coefficients, β_b's and β_l's. For Group A, there are 884 variables to be monitored; and for Group B and Group C, 806 each. By discarding the first 2000 iterations and taking a 20-skeleton (that is, a chain of length $10000/20 = 500$), all the $\hat{R}$'s were below the common cut-off 1.1 for Group A. For Group B, 6 out of 806 are between 1.1 and 1.12; for group C, 1 out of 806 is between 1.1 and 1.2; the rest are all below 1.1. We took a 20-skeleton was due to lack of storage and memory in the computer used. In usual applications of MCMC, it would be a waste to throw away so many draws. In our setting, however, the goal was to create only 10 imputations for all subsequent analyses. We therefore wanted to make sure that they are of as good quality as possible in the sense of being genuinely independent draws from the posited posterior distribution.

B Checking for Sub-Populations

Here we present several exploratory analyses to illustrate the complications in checking imputation quality. In Table 4 and Table 6, we compared the imputed rates and the observed rates under the new design across different Latino ethnicity groups. Here we compare them in sub-populations obtained via a few stratifying variables. In particular, we examine gender, insurance type, and major depression, among which gender and insurance type are predictors included in the imputation model, but major depression (MDE) is not included. Figure 5 shows the comparison in male and female groups for lifetime service use. Visually, the results seem to be acceptable, with the bars for the imputed rates much closer to the bars for the rates under the new group than under the old one, and the difference in patterns resemble those in Table 4. However, this provides minimal comfort, since gender is one of the predictors in our imputation model.

For a public data set, a potential analysis may include variables that are 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 not included in the imputation model. Figure 6 shows the observed rates and imputed rates for both the MDE positive and negative cases. Though MDE is not part of the model, the comparative results are quite similar to that of Figure 5, 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 one wishes to, is that those variables included in the model will

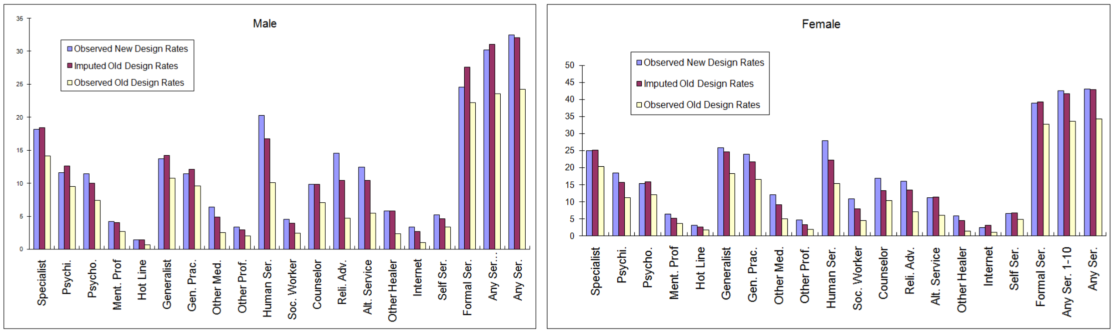

Figure 5: Lifetime Service by Gender

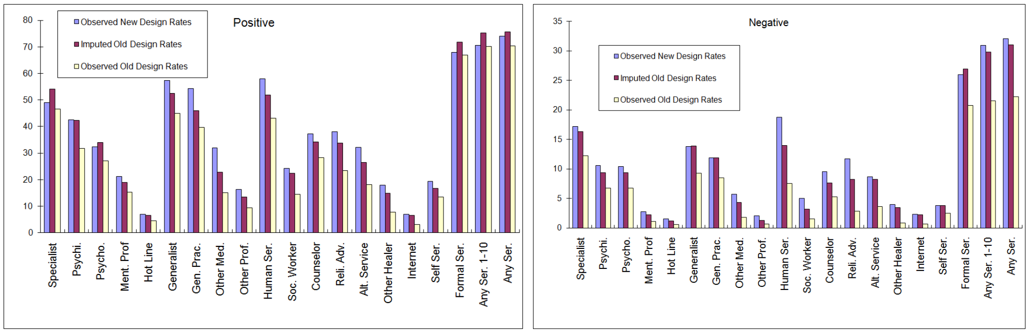

Figure 6: Lifetime Service by Major Depression

capture the essence of all important covariates for the outcome variable that is being imputed.

So far, so good. But unfortunately there is no guarantee that there are no large "discrepancies" for sub-populations that are stratified even according to a variable that is included in the model, let alone for those that are not included. For instance, we discover this when comparing the rates by different insurance types. NLAAS documented six types of insurance, that is, *not insured, private insurance through employer*, *private by purchased insurance*, *Medicare*, *Medicaid*, and *other insurance*. Figure 7 shows the comparison of the imputed service rates and the observed new group

	Effective Sample Size			Sample Size		
	Total	New	Old	Total	New	Old
Not Insured	610	163	447	1082	283	799
Private Through Employer	1006	234	776	2349	588	1761
Private Purchased	107	25	83	259	61	198
Medicare	158	50	123	508	113	395
Medicaid	318	82	237	515	125	390
Other Ins.	67	19	48	151	36	115

Table 8: Sample Size and Effective Sample Sizes of the Insurance Groups.

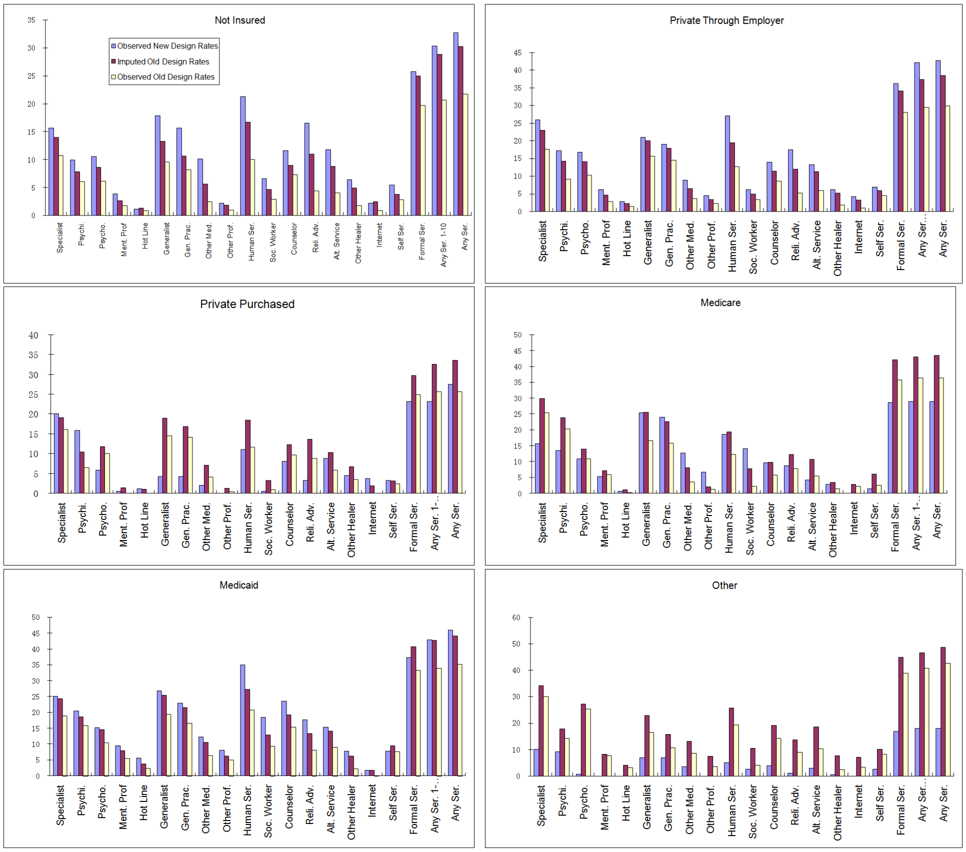

Figure 7: Lifetime Service by Insurance Categories

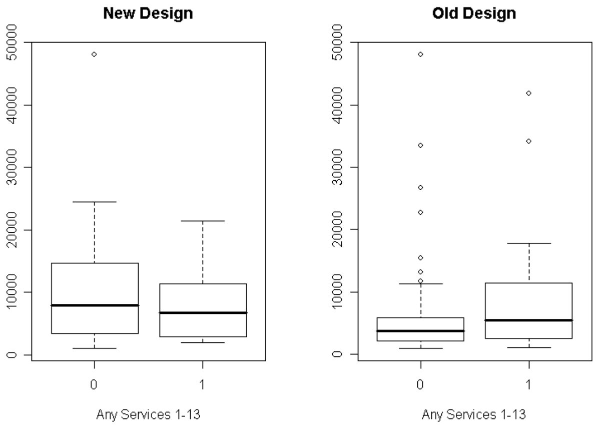

Figure 8: Weights versus Service in the “Other” Insurance Category

	Weighted Averages	
	New Design	Old Design
Number of Disorders	0.37	0.92
Any Disorder	0.20	0.22
Anxiety Disorder	0.11	0.18
Substance Disorder	0	0.17
	Correlations with Survey Weights	
	New Design	Old Design
Formal Services	-0.07	0.17
Any Services 1-10	-0.11	0.17
Any Services 1-13	-0.11	0.16

Table 9: Some Statistics in Other Insurance Group.

	Simple Mean			Weighted Mean		
	New	Old	Old-New	New	Old	Old-New
Formal Service (1-8)	0.19	0.30	0.11	0.17	0.39	0.22
Any Service (1-10)	0.22	0.31	0.09	0.18	0.41	0.23
Any Service (1-13)	0.22	0.34	0.12	0.18	0.43	0.25

Table 10: Observed Rates for Other Insurance Group.

rates by insurance types. We see that, out of the six sub-populations, the comparative results from the (highlighted) groups of *Private Purchased, Medicare*, and *Others*, some of the imputed rates are substantially higher than the observed rates from the new-design group. The *Others* group is most extreme, with some imputed rates more than double the observed rates from the corresponding sub-group under the new design.

So what is wrong? Is this an evidence of the gross failure of our imputation model? To answer these questions, let us first look into a few more facts. First, Table 8 gives the sample sizes and effective sample sizes of all six insurance groups, and we notice 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, with the ratio of maximum to minimum exceeding 1,000, it is important to calculate the effective sample size. We use the common approximation for this purpose (see [14, 15]):

$$n_{eff} \approx \frac{n}{1 + s_W^2/\bar{W}^2}, \qquad (8)$$

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 8, the effective sample sizes are significantly smaller than the nominal sample sizes, making the instability problems with small sub-populations particularly serious. In particular, the Medicaid group has a similar sample size as the Medicare group (515 vs 508), but its effective sample size is more than twice that of the Medicare group (318 vs 158). This further indicates the impact of (real) sample sizes, because the substantial over-imputation occurs for the Medicare group but not for the Medicaid group.

Perhaps more disturbingly, we observe that whenever our model "over-imputes", the corresponding observed rates from the traditional design group are higher than those of the new group. 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 under-report the service uses. Is it a contradiction to our assumption of under-reporting by the oldgroup when the observed rates from the old-design group are higher than those from the new group?

Not necessarily. The comparison to the new group reveals under-reporting of the old group only when the two groups are comparable to start with. We rely 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 group starts with a very high service rate compared to a new group, then even with the under-reporting, it can still end up with significantly higher self-reported rates, causing the false impression that under-reporting may not have occurred.

Indeed, there is good evidence for this possibility. When we look into some diagnostic depression variables included in our model (number of disorders, any disorder, anxiety disorder, substance disorder), we discover that the old group has much higher values on these important predictors. Table 9 shows the sample weighted averages of these variables for the *other* group. We see that the old group has much higher disorder rates and more disorders per person than the new group In the most extreme case, there are no reported cases of substance disorders in the new group, a good reminder of the large variability due to small (effective) sample sizes, but a 17% estimated rate from the old group. Such variables are known to be highly correlated with psychiatric service use, and therefore a serious unbalance between them can lead to substantial discrepancies in the service rates. 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 1.64 (0.23) for substance disorder and 2.13 (0.11) 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 9 shows that the correlation between the survey weights and three “any rates” are much higher in the old group than those in the new group, even with different signs! Second, Figure 8 shows box plots of the survey weights against the binary “any services” variable. There are two positive cases in the old group that have very large weights. These two facts together also contribute to the phenomenon of the higher self-reported rates from the old group. To see this more clearly, Table 10 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 a quantitative indication as to how much of the problem is due to weights and how much is due to the difference in the background variables. For this case, the examination above leads us to believe that both problems are very significant. In general, how to conduct such an examination in a systematical way and how to disentangle them is an important but exceedingly challenging issue.

C Lower Bound on $Var(\bar{Z} - \bar{Y}_M)$ and its Estimate

Let $\mathcal{F}$ denote all the observed data, and hence $\bar{Z}$ is $\mathcal{F}$ measurable. The usual “EVE law” then allows us to decompose

$$Var(\bar{Z} - \bar{Y}_M) = E(Var(\bar{Z} - \bar{Y}_M|\mathcal{F})) + Var(E(\bar{Z} - \bar{Y}_M|\mathcal{F})). \tag{9}$$

The first term on the right-hand side of (9) is the variance due to finite imputation and therefore it can be estimated by B_M/M. The second term can be interpreted as the variance due to sampling variations in covariates. Let $\bar{Y}_\infty = E(\bar{Y}_M|\mathcal{F})$ be the posterior mean (fitted value) of the oldrate. The second term then becomes

$$Var(E(\bar{Z} - \bar{Y}_M|\mathcal{F})) = Var(\bar{Z} - \bar{Y}_\infty). \tag{10}$$

Letting $\mathcal{H}$ be all the covariates and the remaining 12 service variables, we can then apply the EVE law to (10) and conclude that,

$$\begin{aligned} Var(\bar{Z} - \bar{Y}_\infty) &= E(Var(\bar{Z} - \bar{Y}_\infty|\mathcal{H})) + Var(E(\bar{Z} - \bar{Y}_\infty|\mathcal{H})) \\ &\geq Var(E(\bar{Z} - \bar{Y}_\infty|\mathcal{H})). \end{aligned}$$

To proceed further, let $\bar{Y}$ be the service rate of the oldgroup without under-reporting, that is, if the ordering of question has no effect. Under the null hypothesis that our imputation model is adequate, the bias of the imputation model should be of order $o_p(1/\sqrt{n})$, that is,

$$E(\bar{Y}_\infty|\mathcal{H}) = E(\bar{Y}|\mathcal{H}) + o_p(\frac{1}{\sqrt{n}}). \tag{11}$$

This assumption requires that the bias is of a smaller order than that of the standard error. It holds trivially in the case of standard linear prediction, that is, when $\bar{Y}_\infty$ is the prediction from a linear regression (with a constant prior on the regression coefficient) and X is the covariate matrix. In such a case, $E(\bar{Y}_\infty|\mathcal{H}) = E(X\hat{\beta}|\mathcal{H}) = X\beta = E(Y|\mathcal{H})$, so (11) holds without the error term. In the case of a generalized linear model with informative prior, the posterior mean is no longer an unbiased estimator. Nevertheless, invoking the philosophy of hypothesis testing, we can simply take (11) as our null hypothesis (or a consequence of it) that our imputation model is a "good" model. Under this null, we can further approximate

$$E(\bar{Z} - \bar{Y}_\infty|\mathcal{H}) = E(\bar{Z} - \bar{Y}|\mathcal{H}) + o_p(\frac{1}{\sqrt{n}}),$$

where n is the total sample size. Since the old and new groups are randomized, $E(\bar{Z} - \bar{Y}|\mathcal{H}) = O_p(1/\sqrt{n})$ by Central Limit Theorem. Therefore, we have

$$Var(\bar{Z} - \bar{Y}_\infty) \geq Var(E(\bar{Z} - \bar{Y}|\mathcal{H})) + o_p(\frac{1}{n}) = \frac{V^{new}}{n^{new}} + \frac{V^{old}}{n^{old}} + o_p(\frac{1}{n}). \tag{12}$$

The equality in (12) is true because $E(\bar{Z}|\mathcal{H})$ depends only on the new design covariates and $E(\bar{Y}|\mathcal{H})$ depends only on the old covariates. Since the new and old design assignments are randomized, the corresponding covariates are independent and hence $E(\bar{Z}|\mathcal{H})$ and $E(\bar{Y}|\mathcal{H})$ are independent too. To estimate V^{new}, we fit a generalized linear model:

$$E(S|\tilde{X}) = G^{-1}(\tilde{X}\beta), \tag{13}$$

where $\tilde{X}$ includes all the covariates used in the imputation as well as the remaining 12 service variables. Let $\hat{\beta}$ be the estimate of β based on the new design data only. The sample variance of the fitted values,

$$\hat{V}^{new} = \frac{1}{n-1}\sum_{j=1}^{n}\left(G^{-1}(\tilde{X}_j\hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n}G^{-1}(\tilde{X}_i\hat{\beta})\right)^2 \tag{14}$$

then serves as our estimate of V^{new}.

However, estimating V^{old} is trickier, because $\mathcal{H}$ includes all the covariates and the other service variables not under investigation. For the old-design group, the service indicators are only partially observed, therefore it is not straightforward to obtain a consistent estimator for V^{old}. Note however,

$$V^{old} = Var(E(S^{old}|\mathcal{H})) \geq Var(E(S^{old}|X)) \equiv V^{old}_{low}, \tag{15}$$

where X is the matrix of covariates alone (services not included). Therefore, we fit a similar model as (13) but only include the covariates in matrix X and use (14) to obtain $\hat{V}^{old}_{low}$.