SUPPLEMENTARY MATERIAL

In this section we discuss the multiple imputation (MI) procedure described in Section 3 and used to generate the results displayed in Table 1. We first show that the MI procedure is in general proper in the sense of Rubin (1987). We also give insights as to why MI and the semiparametric maximum likelihood (SPMLE) estimators had different efficiency.

| MI procedure

Let Y be the outcome, X be the covariate of interest, X^* its error-prone counterpart, and Z be the error-free covariate. We have the following setup: $X \sim \mathrm{N}\left(g(Z), \sigma_x^2\right)$, where $g(Z)$ is a function of Z; $X^* = X + U$, where $U \sim \mathrm{N}\left(0, \sigma_u^2(Z)\right)$, and $\sigma_u^2(Z)$ is a function of Z; $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 X^* + \epsilon$, where $\epsilon \sim \mathrm{N}\left(0, \sigma^2\right)$. Then,

$$\begin{pmatrix} X \\ X^* \end{pmatrix} \Big| Z \sim \mathrm{N}\left(\begin{pmatrix} g \\ g \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_u^2 \end{pmatrix}\right),$$

and $X \,|\, X^*, Z \sim \mathrm{N}\left(\lambda X^* + (1-\lambda)g, \lambda\sigma_u^2\right)$, where $\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$. A critical step in multiple imputation (MI) is to sample X from the conditional distribution $p(X \,|\, X^*, Z, Y)$, whose form is derived below:

$$\begin{aligned}
& p(X \,|\, X^*, Z, Y) \\
\propto & f(Y \,|\, X, Z, X^*) p(X \,|\, X^*, Z) \\
\propto & \exp\left\{-\frac{1}{2\sigma^2}\left(Y - \beta_0 - \beta_1 X - \beta_2 Z - \beta_3 X^*\right)^2\right\} \exp\left[\frac{1}{2\lambda\sigma_u^2}\left\{X - \lambda X^* - (1-\lambda)g\right\}^2\right] \\
\propto & \exp\left(-\frac{\lambda\sigma_u^2\beta_1^2 + \sigma^2}{2\lambda\sigma^2\sigma_u^2}\left[X - \frac{\lambda\sigma_u^2\beta_1\left(Y - \beta_0 - \beta_2 Z - \beta_3 X^*\right) + \sigma^2\left\{\lambda X^* + (1-\lambda)g\right\}}{\lambda\sigma_u^2\beta_1^2 + \sigma^2}\right]^2\right).
\end{aligned}$$

That is,

$$X \,|\, X^*, Z, Y \sim \mathrm{N}\left(\frac{\lambda\sigma_u^2\beta_1\left(Y - \beta_0 - \beta_2 Z - \beta_3 X^*\right) + \sigma^2\left\{\lambda X^* + (1-\lambda)g\right\}}{\lambda\sigma_u^2\beta_1^2 + \sigma^2}, \frac{\lambda\sigma^2\sigma_u^2}{\lambda\sigma_u^2\beta_1^2 + \sigma^2}\right)$$

To estimate $p(X \,|\, X^*, Z, Y)$, one may fit a linear regression of $X \sim X^* + Z + Y$ on the validation sample. This is proper if and only if $g(Z)$ is a linear function of Z and $\sigma_u^2(Z)$ is constant. Otherwise, the model is misspecified and the corresponding MI will be improper. Scenarios 1-3 from Section 5 are therefore proper, but improper in scenario 4. However, we show below that our simulations led to unbiased estimates with coverage close to the nominal value, for all four settings.

β_1	SRS	SSRS	ODS	RS	WRS	SFS	SRS	SSRS	ODS	RS	WRS	SFS
	Scenario 1, $\sigma_U^2 = (0.5, 0.5)$						Scenario 2, $\sigma_U^2 = (1, 1)$					
0	0.000	0.000	0.001	0.001	0.001	0.001	0.000	0.001	0.002	0.001	0.001	0.001
0.5	0.499	0.499	0.500	0.500	0.500	0.501	0.499	0.499	0.500	0.501	0.501	0.501
1	1.000	1.000	1.001	1.001	1.001	1.001	0.999	0.999	1.000	0.999	0.999	1.000
	Scenario 3, $\sigma_U^2 = (3, 3)$						Scenario 4, $\sigma_U^2 = (0.5, 1)$					
0	0.000	-0.001	0.002	0.001	0.001	0.000	-0.001	0.000	0.001	0.001	0.001	0.001
0.5	0.498	0.498	0.500	0.501	0.502	0.501	0.500	0.500	0.501	0.502	0.508	0.494
1	1.000	0.999	1.001	1.000	1.001	1.000	0.999	0.999	1.001	1.001	1.008	0.990

Abbreviations: SRS: Simple random sampling; SSRS: stratified simple random sampling; ODS: Outcome-dependent sampling; RS.

TABLE S.1 Point estimates for β_1 of the MI estimator under SRS, SSRS, and extreme-tail sampling of ODS, RS, WRS, and SFS.

β_1	SRS	SSRS	ODS	RS	WRS	SFS	SRS	SSRS	ODS	RS	WRS	SFS
	Scenario 1, $\sigma_U^2 = (0.5, 0.5)$						Scenario 2, $\sigma_U^2 = (1, 1)$					
0	0.962	0.962	0.960	0.958	0.958	0.956	0.946	0.946	0.946	0.960	0.961	0.954
0.5	0.960	0.961	0.962	0.968	0.968	0.960	0.942	0.936	0.962	0.958	0.959	0.962
1	0.952	0.946	0.960	0.960	0.959	0.948	0.956	0.952	0.932	0.938	0.940	0.938
	Scenario 3, $\sigma_U^2 = (3, 3)$						Scenario 4, $\sigma_U^2 = (0.5, 1)$					
0	0.946	0.950	0.926	0.944	0.943	0.940	0.940	0.952	0.940	0.950	0.952	0.938
0.5	0.938	0.954	0.946	0.942	0.942	0.944	0.940	0.936	0.948	0.958	0.962	0.942
1	0.950	0.954	0.928	0.934	0.934	0.928	0.944	0.950	0.948	0.944	0.947	0.935

Abbreviations: SRS: Simple random sampling; SSRS: stratified simple random sampling; ODS: Outcome-dependent sampling; RS.

TABLE S.2 Coverage for β_1 of the MI estimator under SRS, SSRS, and extreme-tail sampling of ODS, RS, WRS, and SFS.

| Efficiency of MI and SPMLE

Now, suppose $g(Z) = a + bZ$ for some constants a and b, and $\sigma_u^2(Z)$ is constant, we have

$$
\begin{aligned}
&\mathrm{E}\left(X|X^*, Z, Y\right) \\
&= \frac{\lambda\sigma_u^2\beta_1\left(Y - \beta_0 - \beta_2 Z - \beta_3 X^*\right) + \sigma^2\left\{\lambda X^* + (1-\lambda)g\right\}}{\lambda\sigma_u^2\beta_1^2 + \sigma^2} \\
&= \frac{\sigma^2(1-\lambda)a - \lambda\sigma_u^2\beta_1\beta_0}{\lambda\sigma_u^2\beta_1^2 + \sigma^2} + \frac{\sigma^2\lambda - \lambda\sigma_u^2\beta_1\beta_3}{\lambda\sigma_u^2\beta_1^2 + \sigma^2}X^* + \frac{\sigma^2(1-\lambda)b - \lambda\sigma_u^2\beta_1\beta_2}{\lambda\sigma_u^2\beta_1^2 + \sigma^2}Z + \frac{\lambda\sigma_u^2\beta_1}{\lambda\sigma_u^2\beta_1^2 + \sigma^2}Y.
\end{aligned}
$$

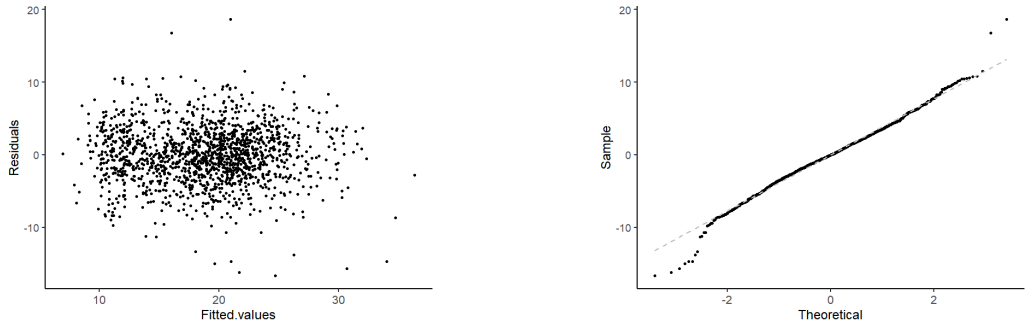
Notice that β_3 only affects the coefficient of X^*. Thus, when $f(Y|X, Z, X^*) = f(Y|X, Z)$, i.e., $\beta_3 = 0$ which is the assumption made in this manuscript (see equation (1)), removing Y from the imputation model is incorrect. The

unconstrained MI approach encompasses a larger classes of models than the SPMLE, leading thus to less efficient estimates. Now, if $\beta_1 = \beta_3 = 0$, we have that

$$\mathrm{E}\left(X \mid X^*, Z, Y\right) = (1-\lambda)a + \lambda X^* + (1-\lambda)bZ.$$

Notice that there is no need to include Y in the imputation model. The unconstrained MI will, therefore, result in higher loss in efficiency compared to SPMLE when $\beta_1 = 0$.

| Residual plots for case study



(a) (b)

FIGURE S.1 Residual (a) and quantile-quantile (b) plots for the linear regression using the fully validated data.