Supplementary Material for Leveraging Prognostic Baseline Variables to Gain Precision in Randomized Trials

ELIZABETH COLANTUONI, MICHAEL ROSENBLUM

March 2015

Table of Contents **Page**

1 Case with Missing Outcomes

We now consider the case where the outcome variable Y may be missing for some participants. Let Δ denote the indicator of Y being observed, and assume the data in a trial consists of n independent, identically distributed realizations $(W_i, A_i, \Delta_i, \Delta_i Y_i)_{i=1}^n$ of the random vector $(W, A, \Delta, \Delta Y)$, each drawn from the (unknown) probability distribution P. The parameter

of interest is the same as in the case of no missing outcomes, i.e., ψ as defined in (3.1). Define the unadjusted estimator of ψ as

$$\frac{\sum_{i=1}^{n} \Delta_i Y_i A_i}{\sum_{i=1}^{n} \Delta_i A_i} - \frac{\sum_{i=1}^{n} \Delta_i Y_i (1 - A_i)}{\sum_{i=1}^{n} \Delta_i (1 - A_i)}.$$

We assume the outcome is missing at random [1], i.e., $P(\Delta = 1|W, A, Y) = P(\Delta = 1|W, A)$. Under this assumption, the average treatment effect ψ is still identified by the distribution of the observed data $(W, A, \Delta, \Delta Y)$ since we have for each $a \in \{0, 1\}$,

$$P(Y = 1|A = a) = E_W P(Y = 1|A = a, W) = E_W P(Y = 1|\Delta = 1, A = a, W), \qquad (1)$$

where E_W denotes expectation with respect to the marginal distribution of W under P; the first equality above follows from the randomization assumption that A and W are independent, and the second equality follows from the missing at random assumption.

All of the adjusted estimators we consider can be modified to handle missing outcomes, under the missing at random assumption. We show how to do this for PLEASE. The main change is that a logistic regression working model for $P(\Delta = 1|A, W)$ will be incorporated into the DR-WLS estimator and PLEASE (see Sections 4.1 and 4.2). Let $h(W, A, \eta)$ be a logistic regression working model for $P(\Delta = 1|A, W)$, which contains an intercept, a main term A, and is allowed to have additional, prespecified main terms and interaction terms, e.g., $h(W, A, \eta) = \text{expit}(\eta_0 + \eta_1 A + \eta_2 W_1 + \eta_3 W_1 A + \eta_4 W_2)$. Fit the working model $h(W, A, \eta)$ using maximum likelihood estimation to obtain the vector of estimated coefficients $\hat{\eta}$. The DR-WLS estimator of Section 4.1 is modified to be the following:

> For each study arm $a \in \{0, 1\}$, define a logistic regression working model $Q^{(a)}(\tilde{W}, \beta^{(a)})$ for $P(Y = 1|\Delta = 1, A = a, W)$, e.g., $Q^{(a)}(\tilde{W}, \beta^{(a)}) = \text{expit}(\beta^{(a)T}\tilde{W})$. Fit the model $Q^{(1)}(\tilde{W}, \beta^{(1)})$ for $P(Y = 1|\Delta = 1, A = 1, W)$ using weighted logistic regression with weights $\left\{h(W, 1, \hat{\eta}) g(\tilde{W}, \hat{\alpha})\right\}^{-1}$ and only using data from participants with both $\Delta = 1, A = 1$; similarly, fit the model $Q^{(0)}(\tilde{W}, \beta^{(0)})$ for $P(Y = 1|\Delta = 1, A = 0, W)$ using weighted logistic regression with weights $\left[h(W, 0, \hat{\eta}) \left\{1 - g(\tilde{W}, \hat{\alpha})\right\}\right]^{-1}$ and only using data from participants with both $\Delta = 1, A = 0$. For each study arm $a \in \{0, 1\}$, the initial estimator (called DR-WLS) for $E(Y|A = a)$ is computed using (5).

Next, the augmented model g_{aug} is fit as in the original Step 1 to obtain $g_{aug}(\tilde{W}, \tilde{\alpha}, \tilde{\gamma})$. The following procedure is added to Step 1, to augment the working model $h(W, A, \eta)$:

Define the following new variables:

$$\begin{aligned} u_1'(W,A) &= A\{Q^{(1)}(\tilde{W},\hat{\beta}^{(1)}) - \hat{\mu}_1\}\left\{h(W,1,\hat{\eta})\right\}^{-1}; \\ u_0'(W,A) &= (1-A)\{Q^{(0)}(\tilde{W},\hat{\beta}^{(0)}) - \hat{\mu}_0\}\left\{h(W,0,\hat{\eta})\right\}^{-1}. \end{aligned}$$

Fit the following augmented logistic regression model for $P(\Delta = 1|A,W)$:

$$h_{aug}(W,A,\eta,\nu) = \text{expit}\left\{\text{logit}(h(W,A,\eta)) + \nu_0 u_0'(W,A) + \nu_1 u_1'(W,A)\right\},$$

to obtain estimated coefficients $\tilde{\eta}, \tilde{\nu}$.

In Step 2, the DR-WLS estimator is computed using the augmented model fits for g_{aug} and h_{aug} from the modified Step 1. This completes the description of how PLEASE is modified to handle missing outcome data. Under the missing at random assumption, and if the working model h is correctly specified, then properties C-E hold for this estimator, and it is guaranteed to be at least as precise as the IPW estimator that uses the working models g and h in constructing its weights. Also, under the missing at random assumption, if the conditional probability of the outcome being missing $P(\Delta = 1|A,W)$ depends on W, then the unadjusted estimator is not guaranteed to be consistent; in contrast, the above estimator is consistent under the missing at random assumption if either of the following hold: the working model h is correctly specified, or the working models $Q^{(0)}, Q^{(1)}$ are correctly specified. In Section 7 of the Supplementary Material, we sketch the proof of the above claims.

We ran additional simulations comparing PLEASE to the unadjusted estimator in the context of the MISTIE II trial where we set outcome data to be missing at random. We considered scenarios 1 and 2, where the only change was to set $P(\Delta|A,W)$ such that approximately 15% of outcome values were missing (i.e., $\Delta = 0$). We first ran simulations where data is generated with missingness completely at random, denoted MCAR, where we set $P(\Delta|A,W)$ to not depend on A, W, and to equal 0.15. In the missing at random (MAR) case, we similarly set the marginal probability of missingess to approximately 15% and we generated:

$$P(\Delta = 0|A,W) = \text{logit}^{-1}(-3.57{+}0.23{\times}(NIHSS{-}22){\times}(1{-}A){+}1.5{\times}0.23{\times}(NIHSS{-}22){\times}A),$$

where $NIHSS$ is an element of the vector W. This results in a greater probability of a missing outcome for larger values of NIHSS, and creates an interaction by treatment arm (with greater probability of a missing outcome under $A = 1$ than $A = 0$, conditioned on W).

The resulting performance of PLEASE versus the unadjusted estimator is shown in Table 1. The measure of relative efficiency is the ratio of mean squared error (MSE) comparing that for the unadjusted estimator to the that for PLEASE. We use MSE instead of the ratio of variances since the unadjusted estimator is not consistent in some cases. Each case labeled zero average treatment effect corresponds to scenario 1, and each case labeled positive average treatment effect corresponds to scenario 2. (Scenarios 1 and 2 are defined in Section 5.1 of the main paper.) The only modifications we make are to introduce a missingness distribution $P(\Delta|A, W)$ as described above. In each case involving MCAR, there is little bias in either estimator, and similar efficiency gains for PLEASE as in the complete data case. In the MAR case, as expected, the unadjusted estimator has severe bias, while the PLEASE estimator (which uses correctly specified propensity score and missingness working models) has very small bias. The efficiency gains for the PLEASE estimator were large in the case of zero average treatment effect, and there was only a small gain for the case with a positive average treatment effect.

To determine if the PLEASE estimator would outperform the unadjusted estimator under MAR data in larger studies, we tripled the size of the simulated studies to n = 1236. We observed larger efficiency gains for both cases (zero and positive average treatment effect) than for the smaller sample size (n=412) simulation studies. This is because at the larger sample sizes, bias dominates variance in terms of the mean squared error (MSE); the unadjusted estimator, unlike PLEASE, can have substantial bias even at large sample sizes due to not adjusting for informative missingness.

2 Additional Covariates Used in Rotnitzky *et al.*$_{K=1}$

We describe our implementation of the estimator referred to as Rotnitzky *et al.*$_{K=1}$ in Section 4.3, which was applied in the simulation studies in Sections 5 and 6. It is a special case from the class of estimators in Rotnitzky *et al.* [2, Section 3]. In the context of a randomized trial, this class of estimators is defined by a model $h(a, \tau)$ for $E(Y|A = a)$, where $\tau = (\tau_0, \tau_1)$. We set h to be the saturated model $E(Y|A = a) = h(a, \tau) = (1-a)\tau_0 + a\tau_1$. Let $\tau^* = (E(Y|A = 0), E(Y|A = 1))$, and define the contrast $\phi_1(\tau^*) = \tau_1^* - \tau_0^*$, which equals ψ (defined in Section 3.1). We define the function $b(a)$ to be the column vector $b(a) = (1 - a, a)^T$.

The above definitions, along with the working models as defined in Section 4, uniquely determine an estimator in the class from [2, Section 3]. This estimator has a similar structure as the PLEASE estimator from Section 4.2, except for the modification described in Section 4.3 and that the following two additional covariates are added to the propensity score working

model in Step 1:

$$\begin{aligned} u_1^R(\tilde{W}) &= \{Q^{(1)}(\tilde{W}, \breve{\beta}^{(1)}) - \breve{\mu}_1\}/g(\tilde{W}, \hat{\alpha}); \\ u_0^R(\tilde{W}) &= \{Q^{(0)}(\tilde{W}, \breve{\beta}^{(0)}) - \breve{\mu}_0\}/\{1 - g(\tilde{W}, \hat{\alpha})\}, \end{aligned}$$

where $(\breve{\beta}^{(0)}, \breve{\beta}^{(1)})$ is the solution to the optimization problem in equation (12) of [2, Section 2.2], which depends on ϕ_1 defined above, and where we define

$$\begin{pmatrix} \breve{\mu}_0 \\ \breve{\mu}_1 \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \left[b(A_i) \frac{Y_i - Q^{(A_i)}(\tilde{W}, \breve{\beta}^{(A_i)})}{g(\tilde{W}, \hat{\alpha})^{A_i} \{1 - g(\tilde{W}, \hat{\alpha})\}^{1-A_i}} + \begin{pmatrix} Q^{(0)}(\tilde{W}, \breve{\beta}^{(0)}) \\ Q^{(1)}(\tilde{W}, \breve{\beta}^{(1)}) \end{pmatrix} \right].$$

Next, the analog of Step 2 from Section 4.2 is carried out.

(Please note: there is a notational difference in that [2] use β wherever we used τ; the reason we use τ is to avoid the notational conflict with β, which we already defined in Section 4.2 as the parameter in the outcome regression score working models.)

3 Calculation of q for simulation scenario 2

We separately consider each of the two trials: MISTIE II and PEARLS. First, define p_0 to be the observed probability that $Y = 1$ based on the data set in the trial (pooling all participants in that trial). Let RD denote the risk difference (i.e., average treatment effect) observed in the trial, based on the unadjusted estimator. For example, for the MISTIE II trial, $p_0 = 0.32$ and $RD = 0.12$. We set $q = RD/(1 - p_0)$, as explained in Section 5.1.

We detail the calculations required for simulation scenario 2, in which Y and W are dependent and there is a positive average treatment effect. For a given simulated study of size n, each simulated participant's values of the variables (W_i, Y_i) are drawn by resampling individuals from the real trial data set with replacement. Next, A is assigned randomly with probability 0.5, independent of (W_i, Y_i). The corresponding data generating distribution has average treatment effect equal to 0. Instead, we would like to simulate from a distribution where the average treatment effect equals p_0. To engineer this, for each participant who initially was assigned $A = 1$ and $Y = 0$, with probability q we randomly reassign their value of Y to be 1. For example, for the MISTIE II trial where $p_0 = 0.32$ and $RD = 0.12$, the solution to the above equation is $q = 0.18$.

4 Details of PEARLS simulation study

We provide details for how Y was simulated based on the PEARLS trial. This was done analogously as for the MISTIE II trial, as described in Section 5.1.

Simulations are conducted under the following four different types of data generating distributions (called scenarios):

Scenario 1: Y and W dependent; zero average treatment effect.

Scenario 2: Y and W dependent; positive average treatment effect.

Scenario 3: Y and W independent; zero average treatment effect.

Scenario 4: Y and W independent; positive average treatment effect.

In scenarios 1 and 2, the relative increase in R^2 was 3% and 2%, respectively. In scenario 1, after resampling (Y,W) from the observed PEARLS data, A is generated independent of (Y,W) with probability 1/2 of being 0 or 1. This induces zero average treatment effect. For scenario 2, we desired to replicate the same magnitude of average treatment effect observed in the PEARLS trial. After generating each simulated participant's data as in scenario 1, for each participant with $A = 0$ and $Y = 0$, we replaced the participant's outcome Y by 1, with probability $q = 0.07$ (independent of W). In scenario 3, W is generated by resampling with replacement from the empirical distribution in the PEARLS data, and A is assigned independent of W with probability 1/2 of being 0 or 1; Y is generated, independent of (W, A), from a Bernoulli distribution with probability $p = 0.18$, which is the observed proportion of successes in PEARLS ignoring treatment assignment (pooling all participants). In scenario 4, W is generated by resampling with replacement from the empirical distribution in the PEARLS data, and A is assigned independent of W with probability 1/2 of being 0 or 1; the conditional distribution of Y given $A = a, W$ is set to be Bernoulli with probability p_a of $Y = 1$, where p_a is the observed proportion of successes in each treatment group in the PEARLS trial ($p_0 = 0.21$ and $p_1 = 0.15$).

5 Proof that Properties B-D hold for PLEASE when the parameter is any smooth contrast between $E(Y|A = 1)$ and $E(Y|A = 0)$

Consider the case of no missing outcomes. It follows directly from the arguments in [2] that the PLEASE estimator has properties B-E from Section 3.3, when the parameter of interest is the difference between population means $E(Y|A = 1) - E(Y|A = 0)$. We consider the more general case when the parameter is any smooth contrast between $E(Y|A = 1)$ and

$E(Y|A=0)$ as defined in Section 8 of the main paper. Let f be any differentiable function from $\mathbb{R}^2$ to $\mathbb{R}$, and denote the corresponding contrast by

$$\psi_f = f(E(Y|A=1), E(Y|A=0)).$$

Let $f(\bar{\mu}_1, \bar{\mu}_0)$ denote the generalization of PLEASE to estimate ψ_f given in Section 8 of the main paper. Define the unadjusted estimator of ψ_f to be $f(m_1, m_0)$, where m_a is the sample mean among those assigned to arm a, for each $a \in \{0, 1\}$.

We next show that for any $P \in \mathcal{M}$, the asymptotic variance of the adjusted estimator $f(\bar{\mu}_1, \bar{\mu}_0)$ is at most that of the unadjusted estimator $f(m_1, m_0)$; this implies property B holds when the parameter of interest is ψ_f. The proof is a direct extension of the results of [3] and [2], based on applying the delta method.

[2] show that as sample size $n \to \infty$, the distribution of

$$\sqrt{n}\left\{(\bar{\mu}_1, \bar{\mu}_0) - (E(Y|A=1), E(Y|A=0))\right\}$$

converges to a bivariate normal distribution with mean $(0, 0)$ and covariance matrix Σ_{adj}, and the distribution of

$$\sqrt{n}\left\{(m_1, m_0) - (E(Y|A=1), E(Y|A=0))\right\}$$

converges to a bivariate normal distribution with mean $(0, 0)$ and covariance matrix Σ_{unadj}. [2], on page 447, prove that $\Sigma_{adj} - \Sigma_{unadj}$ is negative semi-definite. These results hold under regularity conditions given by [2, pp. 445].

Applying the delta method, we have

$$\sqrt{n}\left\{f(\bar{\mu}_1, \bar{\mu}_0) - f(E(Y|A=1), E(Y|A=0))\right\}$$

converges to a normal distribution with mean 0 and covariance $(\nabla f)^T \Sigma_{adj} \nabla f$, and

$$\sqrt{n}\left\{f(m_1, m_0) - f(E(Y|A=1), E(Y|A=0))\right\}$$

converges to a normal distribution with mean 0 and covariance $(\nabla f)^T \Sigma_{unadj} \nabla f$. Since $\Sigma_{adj} - \Sigma_{unadj}$ is negative semi-definite, we have

$$(\nabla f)^T \Sigma_{adj} \nabla f \leq (\nabla f)^T \Sigma_{adj} \nabla f,$$

that is, the asymptotic variance of the adjusted estimator $f(\bar{\mu}_1, \bar{\mu}_0)$ is smaller or equal to the

asymptotic variance of the unadjusted estimator $f(m_1, m_0)$. This shows property B holds when the parameter of interest is the contrast $\psi_f = f(E(Y|A=1), E(Y|A=0))$.

Property C follows from the analogous proof in [2], combined with the delta method as above. Property D is immediate since PLEASE does not require solving a non-convex optimization problem. Since the extension of PLEASE to estimate ψ_f is defined as $f(\bar{\mu}_1, \bar{\mu}_0)$, it is always in the range of the function f; this is the analog of property E.

6 Sensitivity of our simulation studies based on the MISTIE trial to exclusion of a highly prognostic baseline variable

To assess the sensitivity of our simulation studies to excluding a highly prognostic baseline variable, we excluded the NIHSS from our simulation studies based on the MISTIE trial. Table 2 displays the results. In scenario 1, the efficiency gains (compared to the unadjusted estimator) were roughly 13% for the IPW estimator and ranged from 15% to 27% for the other adjusted estimators. In scenario 2, efficiency gains were roughly 10% for the IPW estimator and ranged from 11% to 18% for the other adjusted estimators. These efficiency gains are substantially lower than when NIHSS is included (as shown in Table 2). This shows that leaving out a highly prognostic baseline variable can substantially reduce the efficiency gains of the adjusted estimators.

7 Sketch of proofs for claims involving outcomes missing at random

Throughout this section, we make the missing at random assumption. Below, we refer to the modified version of PLEASE given in Section 1 of the Supplementary Material simply as PLEASE. It follows directly from the arguments of [2] that if the working model h is correctly specified, then properties C-E hold for PLEASE.

We next show that if the working model h is correctly specified, PLEASE is guaranteed to be at least as precise as the IPW estimator that uses working models g and h in constructing its weights. Let $\alpha^*, \beta^{(0)*}, \beta^{(1)*}, \eta^*, \mu_0^*, \mu_1^*$ denote the limits (in probability) of $\hat{\alpha}, \hat{\beta}^{(0)}, \hat{\beta}^{(1)}, \hat{\eta}, \hat{\mu}_0, \hat{\mu}_1$, respectively. Let $\alpha_{aug}^*, \gamma_{aug}^*, \eta_{aug}^*, \nu_{aug}^*$ denote the limits of $\tilde{\alpha}, \tilde{\gamma}, \tilde{\eta}, \tilde{\nu}$, respectively. By the regularity assumptions in Section 3.3, these limits exist. Let $u_0^*, u_1^*, u_0'^*, u_1'^*$ denote the functions u_0, u_1, u_0', u_1', respectively, where each of $\hat{\alpha}, \hat{\beta}^{(0)}, \hat{\beta}^{(1)}, \hat{\eta}, \hat{\mu}_0, \hat{\mu}_1, \tilde{\alpha}, \tilde{\gamma}, \tilde{\eta}, \tilde{\nu}$ is replaced by its corresponding limit. Similarly, let g_{aug}^* and h_{aug}^* denote the augmented models g_{aug} and h_{aug}, respectively, where each of u_0, u_1, u_0', u_1' is replaced by the corresponding function $u_0^*, u_1^*, u_0'^*, u_1'^*$.

By the results of [2, pp. 446-448] it suffices to show the following two terms are contained

in the linear span of the scores of the augmented models $g^*_{aug}(\tilde{W}, \alpha, \gamma)$ and $h^*_{aug}(W, A, \eta, \nu)$ at $(\alpha^*_{aug}, \gamma^*_{aug})$ and $(\eta^*_{aug}, \nu^*_{aug})$:

$$\Delta(1-A)\left[h(W,0,\eta^*)\left\{1-g(\tilde{W},\alpha^*)\right\}\right]^{-1}\left\{Q^{(0)}(\tilde{W},\beta^{(0)*})-\mu_0^*\right\}-\left[Q^{(0)}(\tilde{W},\beta^{(0)*})-\mu_0^*\right]; \tag{2}$$

$$\Delta A\left[h(W,1,\eta^*)g(\tilde{W},\alpha^*)\right]^{-1}\left\{Q^{(1)}(\tilde{W},\beta^{(1)*})-\mu_1^*\right\}-\left[Q^{(1)}(\tilde{W},\beta^{(1)*})-\mu_1^*\right]. \tag{3}$$

The above terms represent components of the influence function for certain augmented, inverse probability weighted estimators, as described by [2, Section 2.1].

The score of $g^*_{aug}(\tilde{W}, \alpha, \gamma)$ at $(\alpha^*_{aug}, \gamma^*_{aug})$ is the vector with components

$$\{A - g^*_{aug}(\tilde{W}, \alpha^*_{aug}, \gamma^*_{aug})\}(\tilde{W}^T, u_0^*(W), u_1^*(W)). \tag{4}$$

The score of $h^*_{aug}(W, A, \eta, \nu)$ at $(\eta^*_{aug}, \nu^*_{aug})$ is the vector with components

$$\{\Delta - h^*_{aug}(W, A, \eta^*_{aug}, \nu^*_{aug})\}(v(W), u_0'^*(W, A), u_1'^*(W, A)), \tag{5}$$

where $v(W)$ is the vector of variables in the linear part of the logistic regression model $h(W, A, \eta)$. Since we assumed h is correctly specified (and g is correctly specified by the trial being randomized, as discussed in Section 3.2), it follows by similar arguments as in [2] that

$$g^*_{aug}(\tilde{W}, \alpha^*_{aug}, \gamma^*_{aug}) \equiv g(\tilde{W}, \alpha^*), \text{ and } h^*_{aug}(W, A, \eta^*_{aug}, \nu^*_{aug}) \equiv h(W, A, \eta^*). \tag{6}$$

Since the trial is randomized, we have A and W are independent, i.e., $P(A = 1|W) = P(A = 1) = 1/2$. Therefore, the working model g is correctly specified (since it contains an intercept) and we have $g(\tilde{W}, \alpha^*) = 1/2$ for all values of $\tilde{W}$.

Consider the following linear combination of components of the scores (4) and (5):

$$
\begin{aligned}
& -\{A - g^*_{aug}(\tilde{W}, \alpha^*_{aug}, \gamma^*_{aug})\}2u^*_0(W) + \{\Delta - h^*_{aug}(W, A, \eta^*_{aug}, \nu^*_{aug})\}2u'^*_0(W, A) \\
= \; & -\{A - g(\tilde{W}, \alpha^*)\}2u^*_0(W) + \{\Delta - h(W, A, \eta^*)\}2u'^*_0(W, A) \\
= \; & -\{A - g(\tilde{W}, \alpha^*)\}\{1 - g(\tilde{W}, \alpha^*)\}^{-1}u^*_0(W) \qquad (7) \\
& +\{\Delta - h(W, A, \eta^*)\}\{1 - g(\tilde{W}, \alpha^*)\}^{-1}u'^*_0(W, A) \qquad (8) \\
= \; & -\{A - g(\tilde{W}, \alpha^*)\}\{Q^{(0)}(\tilde{W}, \beta^{(0)*}) - \mu^*_0\}\{1 - g(\tilde{W}, \alpha^*)\}^{-1} \\
& +\{\Delta - h(W, A, \eta^*)\}(1 - A)\{Q^{(0)}(\tilde{W}, \beta^{(0)*}) - \mu^*_0\}\left[h(W, 0, \eta^*)\left\{1 - g(\tilde{W}, \alpha^*)\right\}\right]^{-1} \\
= \; & -\{A - g(\tilde{W}, \alpha^*)\}\{Q^{(0)}(\tilde{W}, \beta^{(0)*}) - \mu^*_0\}\{1 - g(\tilde{W}, \alpha^*)\}^{-1} \\
& +\Delta(1 - A)\{Q^{(0)}(\tilde{W}, \beta^{(0)*}) - \mu^*_0\}\left[h(W, 0, \eta^*)\left\{1 - g(\tilde{W}, \alpha^*)\right\}\right]^{-1} \\
& -(1 - A)\{Q^{(0)}(\tilde{W}, \beta^{(0)*}) - \mu^*_0\}\left\{1 - g(\tilde{W}, \alpha^*)\right\}^{-1} \\
= \; & \Delta(1 - A)\{Q^{(0)}(\tilde{W}, \beta^{(0)*}) - \mu^*_0\}\left[h(W, 0, \eta^*)\left\{1 - g(\tilde{W}, \alpha^*)\right\}\right]^{-1} - \{Q^{(0)}(\tilde{W}, \beta^{(0)*}) - \mu^*_0\},
\end{aligned}
$$

where in (7) and (8) we used that $g(\tilde{W}, \alpha^*) = 1/2$ for all values of $\tilde{W}$, as argued above. The last line in the above display equals (2). This shows (2) is in the linear span of the scores (4) and (5). A similar derivation shows that $\{A - g(\tilde{W}, \alpha^*)\}2u^*_1(W) + \{\Delta - h(W, A, \eta^*)\}2u'^*_1(W, A)$ equals (3), and therefore (3) is also in the linear span of the scores (4) and (5). This completes the verification that (2) and (3) are in the linear span of the scores of $g_{aug}(\tilde{W}, \alpha, \gamma)$ and $h_{aug}(W, A, \eta, \nu)$ at $(\alpha^*_{aug}, \gamma^*_{aug})$ and $(\eta^*_{aug}, \nu^*_{aug})$. By the results of [2], this implies that if the working model h is correctly specified, PLEASE is guaranteed to be at least as precise as the IPW estimator that uses the same working models g and h in constructing its weights.

Also, it follows from the proof in [2] that PLEASE is consistent under the missing at random assumption if either of the following hold: the working model h is correctly specified, or the working models $Q^{(0)}, Q^{(1)}$ are correctly specified.

8 R and SAS code to compute the PLEASE

The R and SAS code is provided in a separate zip file. The R program computes PLEASE with corresponding bootstrap confidence intervals for the case of binary Y where the risk difference is the average treatment effect of interest. The SAS program provides code to compute PLEASE for a variety of outcome distributions; the working outcome regression models may be specified as binomial with logit link, gamma with inverse link, normal with identity link, and poisson with log link. A logistic regression model is used for the working

propensity score model. The output consists of the PLEASE estimators of the mean under treatment 0 and treatment 1; the user can then compute any smooth contrast of interest f (defined in Section 8 of the main paper) by substituting these estimated values into the function f as in Step 2' of Section 8 of the main paper. Standard errors are obtained via bootstrap; the SAS program does not include the bootstrap calculations.

`http://people.csail.mit.edu/mrosenblum/papers/CovAdjRSAS.zip`

References

[1] van der Vaart A. *Asymptotic Statistics*. Cambridge University Press: Cambridge, 1998.

[2] Rotnitzky A, Lei Q, Sued M, Robins J. Improved double-robust estimation in missing data and causal inference models. *Biometrika* 2012; **99**(2):439–456.

[3] Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Double-robustness with improved efficiency in missing and causal inference models. *Technical report, Harvard School of Public Health* 2007; .

Table 1: Implications of missing data on simulated randomized trials with Y dependent on W (scenarios 1 and 2). We compared the relative efficiency of PLEASE to the unadjusted estimator assuming roughly 15% missing data within the context of the MISTIE II trial. The missing data was generated completely at random (MCAR) or at random (MAR). The relative MSE compares the MSE of PLEASE to that of the unadjusted estimator.

MCAR, zero average treatment effect, n = 412				
Estimator	Bias	Variance	MSE	Rel.MSE
Unadjusted	0.00010	0.0025	0.0025	1.000
PLEASE	0.000079	0.0018	0.0018	1.410
MCAR, positive average treatment effect, n = 412				
Estimator	Bias	Variance	MSE	Rel.MSE
Unadjusted	-0.00013	0.0027	0.0027	1.000
PLEASE	-0.00094	0.0021	0.0021	1.270
MAR, zero average treatment effect, n = 412				
Estimator	Bias	Variance	MSE	Rel.MSE
Unadjusted	0.020	0.0012	0.0016	1.000
PLEASE	0.000015	0.00077	0.00077	1.610
MAR, positive average treatment effect, n = 412				
Estimator	Bias	Variance	MSE	Rel.MSE
Unadjusted	0.013	0.0011	0.0012	1.000
PLEASE	-0.00084	0.0010	0.0010	1.060
MAR, zero average treatment effect, n = 1236				
Estimator	Bias	Variance	MSE	Rel.MSE
Unadjusted	0.019	0.0012	0.0016	1.000
PLEASE	0.000015	0.00077	0.00077	2.112
MAR, positive average treatment effect, n = 1236				
Estimator	Bias	Variance	MSE	Rel.MSE
Unadjusted	0.012	0.0011	0.0012	1.000
PLEASE	-0.00084	0.0010	0.0010	1.214

Table 2: Sensitivity of efficiency gains to exclusion of a predictive baseline variable. Data are simulated from the MISTIE II trial assuming Y is dependent on W (scenarios 1 and 2). The propensity score and outcome regression model include an intercept, age and ICH volume. The NIHSS is not included even though this variable is highly correlated with the outcome. The relative efficiency compares the variance of the adjusted estimators to the variance of the unadjusted estimator.

Scenario 1: Y and W dependent, zero average treatment effect				
Estimator	Bias	Variance	MSE	Rel.Efficiency
Unadjusted	-0.00019	0.0021	0.0021	1.000
IPW	-0.00018	0.0019	0.0019	1.130
Model standardization	-0.00017	0.0018	0.0018	1.150
DR-WLS	-0.00017	0.0018	0.0018	1.150
Tan	-0.00014	0.0017	0.0017	1.210
PLEASE	-0.00011	0.0018	0.0018	1.190
Rotnitzky *et al.*$_{K=1}$	-0.00010	0.0017	0.0017	1.270
Gruber and van der Laan	-0.00013	0.0018	0.0018	1.160
Scenario 2: Y and W dependent, positive average treatment effect				
Estimator	Bias	Variance	MSE	Rel.Efficiency
Unadjusted	-0.00026	0.0023	0.0023	1.000
IPW	-0.00028	0.0021	0.0021	1.100
Model standardization	-0.00031	0.0020	0.0020	1.110
DR-WLS	-0.00031	0.0020	0.0020	1.110
Tan	-0.0011	0.0020	0.0020	1.120
PLEASE	-0.00056	0.0020	0.0020	1.130
Rotnitzky *et al.*$_{K=1}$	-0.00043	0.0019	0.0019	1.180
Gruber and van der Laan	-0.000098	0.0020	0.0020	1.120