The Annals of Applied Statistics
2019, Vol. 13, No. 3, 1370–1396
https://doi.org/10.1214/19-AOAS1239

IMPUTATION AND POST-SELECTION INFERENCE IN MODELS WITH MISSING DATA: AN APPLICATION TO COLORECTAL CANCER SURVEILLANCE GUIDELINES[1]

BY LIN LIU, YUQI QIU, LOKI NATARAJAN AND KAREN MESSER

University of California, San Diego

It is common to encounter missing data among the potential predictor variables in the setting of model selection. For example, in a recent study we attempted to improve the US guidelines for risk stratification after screening colonoscopy (*Cancer Causes Control* **27** (2016) 1175–1185), with the aim to help reduce both overuse and underuse of follow-on surveillance colonoscopy. The goal was to incorporate selected additional informative variables into a neoplasia risk-prediction model, going beyond the three currently established risk factors, using a large dataset pooled from seven different prospective studies in North America. Unfortunately, not all candidate variables were collected in all studies, so that one or more important potential predictors were missing on over half of the subjects. Thus, while variable selection was a main focus of the study, it was necessary to address the substantial amount of missing data. Multiple imputation can effectively address missing data, and there are also good approaches to incorporate the variable selection process into model-based confidence intervals. However, there is not consensus on appropriate methods of inference which address both issues simultaneously. Our goal here is to study the properties of model-based confidence intervals in the setting of imputation for missing data followed by variable selection. We use both simulation and theory to compare three approaches to such post-imputation-selection inference: a multiple-imputation approach based on Rubin's Rules for variance estimation (*Comput. Statist. Data Anal.* **71** (2014) 758–770); a single imputation-selection followed by bootstrap percentile confidence intervals; and a new bootstrap model-averaging approach presented here, following Efron (*J. Amer. Statist. Assoc.* **109** (2014) 991–1007). We investigate relative strengths and weaknesses of each method. The "Rubin's Rules" multiple imputation estimator can have severe undercoverage, and is not recommended. The imputation-selection estimator with bootstrap percentile confidence intervals works well. The bootstrap-model-averaged estimator, with the "Efron's Rules" estimated variance, may be preferred if the true effect sizes are moderate. We apply these results to the colorectal neoplasia risk-prediction problem which motivated the present work.

Received April 2018; revised December 2018.

[1] Supported in part by National Institutes of Health Grants R01CA166293 and UL1TR001442, and by a pilot grant from UCSD Department of Family Medicine and Public Health.

Key words and phrases. Post-selection inference, missing data, multiple imputation, model selection, model averaging.

1. Introduction and background.

It is not uncommon that both model selection and missing data are important aspects of the data analysis pipeline. Frequentist approaches which properly incorporate the model selection process into statistical inference include the bootstrap (Efron (2014)) and model averaging (Claeskens (2016)). Multiple imputation is a practical and widely used approach for inference with missing data (Little and Rubin (2002), Rubin (1987), Tanner and Wong (1987), Tsiatis (2006)). However, when both issues are important to the analysis, there is less guidance in the literature on practical and efficient approaches to inference.

In a recent study we attempted to improve the US colorectal cancer prevention guidelines which specify the recommended surveillance interval for colonoscopy following polypectomy (Liu et al. (2016)). As background, annual US incidence of colorectal cancer is about 130,000 individuals, with about 50,000 deaths annually (Siegel, Miller and Jemal (2015)). Colorectal cancer can be prevented by the identification and removal of colorectal polyps during colonoscopy (Lieberman et al. (2012)). During screening colonoscopy, which is recommended every 10 years for adults ages 50–75 years, 20% to 50% of patients are found to have colorectal polyps. US guidelines for the frequency of subsequent surveillance colonoscopy are based on the number, size and histology of these resected polyps. However, these surveillance guidelines have only moderate sensitivity and specificity (Liu et al. (2016)), leading to missed opportunity for cancer prevention on the one hand, and unnecessary colonoscopy on the other. Incorporation of additional prognostic factors might improve risk prediction, with potential large cost reductions and net gain in public health (Martinez et al. (2012)). Our aim was to determine whether a statistical model incorporating known additional clinical risk factors—age, sex, history of prior polyps, and polyp location and grade—could improve estimated sensitivity and specificity above the current US practice (Liu et al. (2016)). However, the cost of adding variables to the current widely used guidelines would be increased complexity in the medical decision process and increased difficulty in assembling the needed information for a given patient. Thus variable selection was a primary goal of the original study—it was of primary interest to determine whether inclusion of a few selected additional factors could meaningfully improve model-based estimates of sensitivity and specificity, using a set of training data. This is the focus of the current paper. In the original study, the predictive performance of the final selected model was then assessed on an independent set of validation data.

Data to estimate the model were aggregated from seven different prospective studies of over 8000 individuals who underwent polypectomy and subsequent surveillance colonoscopy (Liu et al. (2016)). However, not all studies collected all variables. Thus, for three important variables the missing data rate was 24%, 20% and 11%, respectively, with 56% of subjects missing at least one predictor. Hence, both variable selection and missing data were important considerations during the model training phase. The model and associated risk cut-points were developed on a randomly drawn training sample of approximately 5500 subjects. This model-based risk stratification was then applied to the remaining ≈2500 subjects to provide estimates of the potential for improvement in US population-level rates of colonoscopy overuse and underuse.

In the original study, we used an indicator variable method to handle the missing data among predictors at the variable selection step (Jones (1996)), arguing that it would likely be appropriate in this case. However, given the known limitations of this approach, our goal here is to find a practical and general method to incorporate both imputation and selection variation into confidence intervals for model-based estimates, when there is a substantial amount of missing data among the predictor variables. We use both theory and simulation to study the question, and then apply our results to the colorectal cancer risk modeling problem described above.

1.1. *Related literature.* Bootstrap-based methods for consistent and efficient model selection in the presence of imputation for missing data have been reviewed in Long and Johnson (2015). In these papers, B bootstrap resamples are generated from the observed data and the imputation mechanism is applied to each bootstrap resample. Model selection is then applied to each bootstrap-imputed data set. The resulting collection of B selected models is then combined to produce a single final model, using various approaches. A common approach is majority voting, that is, retaining variables which are included in at least πB of the selected models for a given threshold π, and using these to construct the final model (Heymans et al. (2007), Lachenbruch (2011)). Variance estimates are then constructed in the usual way, conditional on the final selected model. Other approaches are based on stability selection (Long and Johnson (2015), Meinshausen and Bühlmann (2010)) and a multiple imputation approach (Wood, White and Royston (2008)) in which a candidate model is estimated in each of M imputed data sets, and Rubin's Rules (Rubin (1987)) are applied to obtain estimated model parameters and their variances, Wald tests are used for backwards model selection, and then Rubin's Rules applied again to the reduced model. Notably, these two latter approaches have a potential efficiency advantage in that they each incorporate the imputation variability into the model selection criteria. However, and importantly from our perspective, in all of these proposals the final estimated parameter variances are conditional on the final model, and thus ignore the model selection process.

Schomaker and Heumann (2014) address inference which incorporates both imputation and model selection variability, in the context of proposing model-averaged estimators. One approach presented (Section 2.1) is multiple imputation, followed by model selection on each of the M imputed data sets. Rubin's Rules are then used to average the estimated coefficients and their variances across the M models; a variable which is not selected into a given model has an assigned coefficient and model-based variance of zero. This model-averaging procedure shrinks the final estimated coefficients towards zero in proportion to their nonselection probability, and produces coefficent and variance estimates which incorporate both imputation and, at least partially, selection variability. Normal theory confidence intervals were computed using Rubin's Rules, but exhibited under coverage. More complex approaches using the bootstrap for multiply imputed model-averaged estimators with exponential weights were also investigated, which did not show under coverage in simulation studies.

1.2. *Aims and organization of the paper.* In this paper, we study three approaches to post imputation-selection inference, and then apply them to the colonoscopy risk modeling problem. The first approach is a single imputation step to fill in missing values, followed by model selection to produce a final model. Confidence intervals and variance estimates are then computed by the bootstrap, using the sequence resample, impute, select. The second approach is multiple imputation followed by model selection and Rubin's Rules to compute variance estimates (Schomaker and Heumann (2014)). The third approach is novel, based on bootstrap model-averaging as in Efron (2014): we again use the sequence resample, impute, select; however, model parameters are estimated by the bootstrap average. We then use a computationally efficient approach to computing the estimated variance using the same bootstrap distribution (Efron (2014)), and thus avoid a second level of bootstrapping. This variance formula is derived from a Hájek projection, similar to the arguments used in the theory of U statistics. In our context, this variance estimate can be construed as analogous to Rubin's Rules for multiple imputation, and hence we call it "Efron's Rules" for bootstrap imputation.

The paper is organized as follows: we first define the estimators (Section 2). Theoretical properties are discussed in Section 3, using the framework of Hjort and Claeskens (2003) and Hjort (2014). In Section 4, we use simulation to study the MSE of the estimators, their distributions and their associated confidence intervals. In Section 5, we apply the estimators to the colorectal cancer surveillance data from Liu et al. (2016). Section 6 is discussion and concluding remarks.

2. Algorithms for imputation, selection then estimation.

2.1. *Notation.* Let the data be n independent observations $(y_i, \mathbf{x}_i)$, where the outcome y is fully observed and the predictors $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})$ may have missing values, with missing indicator $\delta_{i,j} = 1$ if $x_{i,j}$ is missing. The outcome and predictors are related by a statistical model $f(y|\mathbf{x}\boldsymbol{\theta}, \phi)$, where $\boldsymbol{\theta}$ is a $p \times 1$ vector of coefficients, with θ_j set to 0 if predictor x_j is not in the model. Let $\mathbf{y} = (y_1, \dots, y_n)^T$, and denote by $\mathbf{X}$ and Δ the matrices with entries $x_{i,j}$ and $\delta_{i,j}$. Let $\mathbf{X}_I$ denote the imputed dataset under a given imputation model, $x_{Iij} = x_{ij} + \delta_{ij}\gamma_{ij}$, where γ_{ij} represents the imputation error. Let $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$ denote the complete data, $\mathbf{Z}_{\text{obs}}$ denote the observed data and $\mathbf{Z}_I = (\mathbf{y}, \mathbf{X}_I)$ denote the imputed dataset. We assume the data are missing at random (MAR), that is, that $f(\Delta|\mathbf{Z}) = f(\Delta|\mathbf{Z}_{\text{obs}})$, and that this distribution does not depend on $\boldsymbol{\theta}$ or ϕ.

2.2. *Efron's Rules for bootstrap-averaged imputation-selection.* The algorithm is bootstrap imputation followed by model selection; parameter estimates are averaged over the bootstrap distribution, with variance estimates following Efron (2014).

1. Generate B bootstrap datasets $\{\mathbf{Z}_{\text{obs}}^{(b)}, \Delta^{(b)}, b = 1, \ldots, B\}$ from the observed data, including missing indicators, by resampling subjects.

2. On each bootstrap dataset, perform a single imputation, using an imputation method of choice, to obtain the imputed dataset $\{\mathbf{Z}_I^{(b)}, b = 1, \ldots, B\}$.

3. On each imputed dataset, select the best model and compute the associated parameter estimates $\{\hat{\boldsymbol{\theta}}^{(b)}, b = 1, \ldots, B\}$. If predictor x_j is not selected in bootstrap sample b, then $\hat{\theta}_j^{(b)}$ is set to zero.

4. For each component θ_j of $\boldsymbol{\theta}$, compute the smoothed bootstrap estimator $\hat{\theta}_{j\text{ER}}$. This is the average of the $\hat{\theta}_j^{(b)}$ over the bootstrap samples:

$$\hat{\theta}_{j\text{ER}} = \sum_{b=1}^{B} \hat{\theta}_j^{(b)} / B. \tag{2.1}$$

5. Use Efron's Rules to compute $\hat{V}_{j\text{ER}}$, a nonparametric bootstrap estimate of the variance of $\hat{\theta}_{j\text{ER}}$, as follows:

Let $C_i^{(b)}$ be the count of the number of times subject i appears in bootstrap sample b, and let $\bar{C}_i = \sum_{b=1}^{B} C_i^{(b)} / B$ be the average number of times subject i is selected. Let

$$\widehat{\text{cov}}_i(j) = 1/B \sum_{b=1}^{B} (C_i^{(b)} - \bar{C}_i)(\hat{\theta}_j^{(b)} - \hat{\theta}_{j\text{ER}}) \tag{2.2}$$

and let

$$Z_i^{(b)}(j) = (C_i^{(b)} - 1)(\hat{\theta}_j^{(b)} - \hat{\theta}_{j\text{ER}}).$$

Then

$$\hat{V}_{j\text{ER}} = \widehat{\text{var}}(\hat{\theta}_{j\text{ER}}) = \sum_{i=1}^{n} \widehat{\text{cov}}_i(j)^2 - \frac{1}{B^2} \sum_{i=1}^{n} \sum_{b=1}^{B} (Z_i^{(b)} - \widehat{\text{cov}}_i(j))^2. \tag{2.3}$$

Formula (2.3) follows from a Hájeck projection of the bootstrapped average estimator, which is given in equations (3.6), (3.7) and (7.25) of Efron (2014): equations (3.6), (3.7) directly yield (2.2), which is the first right-hand term in (2.3), and equation (7.25) is the second right-hand term in (2.3). The term (7.25) in Efron (2014) is a second-order term in the Hájeck projection (van der Vaart (1998)), and our simulations showed that this term is needed in order to reduce bias in our context.

6. Compute the 95% confidence interval using a normal distribution centered at $\hat{\theta}_{\text{ER}}$ with variance $\hat{V}_{\text{ER}}$.

2.3. *Rubin's Rules for multiple imputation-selection.* The algorithm is multiple imputation followed by model selection; parameter estimates are averaged over the imputation distribution, with variance estimates following Rubin's Rules (Rubin (1987)).

1. From the observed data $\{\mathbf{Z}_{\text{obs}}, \Delta\}$ generate M imputed datasets using a multiple imputation method of choice, to obtain the imputed dataset $\{\mathbf{Z}_I^{(m)}, m = 1, \dots, M\}$.

2. For each imputed dataset m, select the best model and compute the associated parameter estimates $\{\hat{\boldsymbol{\theta}}^{(m)}, m = 1, \dots, M\}$, along with the associated variance estimates (conditional on the selected model) $\widehat{\text{var}}(\hat{\theta}_j^{(m)})$. If predictor x_j is not selected in imputed dataset m, then both $\hat{\theta}_j^{(m)}$ and $\widehat{\text{var}}(\hat{\theta}_j^{(m)})$ are set to zero.

3. For each component θ_j of $\boldsymbol{\theta}$, compute the multiple imputation estimator $\hat{\theta}_{j\text{RR}}$. This is the average of the $\hat{\theta}_j^{(m)}$ over the multiple imputation samples:

$$\hat{\theta}_{j\text{RR}} = \sum_{m=1}^{M} \hat{\theta}_j^{(m)} / M. \tag{2.4}$$

4. Use Rubin's Rules to compute $\hat{V}_{j\text{RR}}$, an estimate of the variance of $\hat{\theta}_{j\text{RR}}$, conditional on $\{\mathbf{Z}_{\text{obs}}, \Delta\}$.

Let

$$B_j = \sum_{m=1}^{M} (\theta_j^{(m)} - \hat{\theta}_{j\text{RR}})^2 / (M - 1).$$

Then

$$\hat{V}_{j\text{RR}} = 1/M \sum_{m=1}^{M} \widehat{\text{var}}(\hat{\theta}_j^{(m)}) + (1 + 1/M) B_j. \tag{2.5}$$

5. Compute the 95% confidence interval using a normal distribution centered at $\hat{\theta}_{\text{RR}}$ with variance $\hat{V}_{\text{RR}}$.

2.4. *A single imputation-selection step, with bootstrap percentile confidence intervals.* Both bootstrap-averaged imputation-selection with Efron's Rules and multiple imputation-selection with Rubin's Rules give rise to model-averaged estimators, averaged across conditional or unconditional imputation-selection distributions. For comparison, we also study the single Impute-Select estimator given by:

1. From the observed data $\{\mathbf{Z}_{\text{obs}}, \Delta\}$ generate a single imputed dataset using an imputation method of choice, to obtain the imputed dataset $\mathbf{Z}_I$.

2. Use $\mathbf{Z}_I$ to select the best model and compute the associated parameter estimates $\hat{\boldsymbol{\theta}}_{\text{IS}}$. If predictor x_j is not selected, then $\hat{\theta}_j$ is set to zero.

3. Generate B bootstrap datasets $\{\mathbf{Z}_{\text{obs}}^{(b)}, \Delta^{(b)}, b = 1, \dots, B\}$ from the observed data, including missing indicators, by resampling subjects.

4. Perform a single imputation on each bootstrap dataset, using an imputation method of choice, to obtain the imputed dataset $\{\mathbf{Z}_I^{(b)}, b = 1, \dots, B\}$.

5. Compute the bootstrap distribution of $\hat{\boldsymbol{\theta}}_{\text{IS}}$, and use it to compute $\hat{V}_{j\text{IS}}$, a nonparametric bootstrap estimate of the variance of $\hat{\theta}_{j\text{IS}}$.

6. Compute the percentile bootstrap confidence interval for $\hat{\theta}_{j\text{IS}}$.

2.5. *Other methods included for comparison.* For comparison, in the simulations below we include a naive estimator based on the complete case analysis, which does not incorporate model selection variability and does not adjust for missing data. This estimator, the "complete case" analysis, drops any case with missing data and computes standard estimates of variance conditional on the selected model. We also include an estimator which does imputation but not model selection (i.e., it includes all variables), the "wide" estmator, to allow comparison to Theorem 3.1 and its Corollary, and to investigate sources of bias in the Rubin's Rules estimate of variance.

3. Theoretical properties of the Efron's Rules and Impute-Select estimators. The asymptotic distribution of post-selection estimators has been elucidated in the work of Claeskens and Hjort (Claeskens and Hjort (2003, 2008a), Hjort (2014), Hjort and Claeskens (2003)). This allows comparison of the MSE of model averaged estimators vs un-averaged estimators. Specializing the Claeskens–Hjort framework to the case of the Efron's Rules estimator, we state Theorem 3.1, which is similar to less formally stated results in their prior work. In the proof, we develop some explicit analytical expressions for the mean squared error of the estimators for the complete data case. This allows us to incorporate the measurement error due to imputation, in Corollary 3.1. These results provide insight into the simulation results of Section 4.

3.1. *Notation and assumptions.* Consider an i.i.d. normal linear regression model of sample size n with p covariates, where $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, with $\mathbf{X}$ an $n \times p$ design matrix of full rank and $\epsilon_i \sim N(0, \sigma^2)$. Without loss of generality, assume the columns of $\mathbf{X}$ are standardized to have mean 0 and variance 1. Let $\hat{\boldsymbol{\theta}}_W$ be the estimate of $\boldsymbol{\theta}$ which includes all p variables (the "wide" model), and let $\hat{\boldsymbol{\theta}}_s$ be the estimate of $\boldsymbol{\theta}$ for the model which contains a subset $s \subset \{1, \dots, p\}$ of selected variables, of size $|s|$, with corresponding design matrix $\mathbf{X}_s$. Because $\mathbf{X}$ is of full rank, $\hat{\boldsymbol{\theta}}_s$ can be written as $\hat{\boldsymbol{\theta}}_s = \mathbf{M}_s\hat{\boldsymbol{\theta}}_W$, where $\hat{\boldsymbol{\theta}}_W$ is $N(\boldsymbol{\theta}, \mathbf{V}_W)$, and where $\mathbf{M}_s = (\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{S}\mathbf{X}'\mathbf{X}$ with $\mathbf{S}$ the selection matrix consisting of the rows of the $p \times p$ identity matrix which correspond to the selected variables. The selection criterion is Akaike's information criterion (AIC), which depends on $\hat{\boldsymbol{\theta}}_s$ and $|s|$. Thus we may write the selection indicator $w(s, \hat{\boldsymbol{\theta}}_W) = I[s = \operatorname{argmin} \text{AIC}(s, \hat{\boldsymbol{\theta}}_W)]$, which is 1 if s is the subset with minimum AIC$(s, \hat{\boldsymbol{\theta}}_W)$ and 0 otherwise. Note that $w(s, \hat{\boldsymbol{\theta}}_W)$ induces a partition of R^p (up to a set of measure 0), with components $\mathcal{A}_s \subset \mathbb{R}^p$ the set where $w(s, \mathbf{u}) = 1$.

3.2. *Distribution of* $\hat{\theta}_{\mathrm{ER}}$ *and* $\hat{\theta}_{\mathrm{IS}}$ *in the complete data case.* For the complete data case, the model selection estimator $\hat{\boldsymbol{\theta}}_{\mathrm{IS}}$ can be written in terms of $\hat{\boldsymbol{\theta}}_W$ as

$$\hat{\boldsymbol{\theta}}_{\mathrm{IS}}(\hat{\boldsymbol{\theta}}_W) = \sum_{s \in S} w(s, \hat{\boldsymbol{\theta}}_W) M_s \hat{\boldsymbol{\theta}}_W, \tag{3.1}$$

where the sum is over all subsets S of $\{1, \dots, p\}$. Hence, the distribution of the model selection estimator is a mixture of conditional normals (conditional on $\boldsymbol{\theta}_W \in \mathcal{A}_s$) and point mass at 0 (for those cases where variable i is not selected).

For the complete data case, the Efron's Rules estimator is the bootstrap-averaged version of the impute select estimator, averaged over the estimated sampling distribution of $\hat{\boldsymbol{\theta}}_W$, and equals within Monte Carlo error to

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{ER}}(\hat{\boldsymbol{\theta}}_W) &= \int \Bigl(\sum_{s \in S} w(s, \mathbf{u}) \mathbf{M}_s \mathbf{u}\Bigr) \boldsymbol{\phi}_W(\mathbf{u} - \hat{\boldsymbol{\theta}}_W)\, d\mathbf{u} \\ &= \sum_{\mathcal{A}_s} \mathbf{M}_s \int_{\mathcal{A}_s} \mathbf{u} \boldsymbol{\phi}_W(\mathbf{u} - \hat{\boldsymbol{\theta}}_W)\, d\mathbf{u},\end{aligned} \tag{3.2}$$

where $\boldsymbol{\phi}_W(\mathbf{u})$ is the density of a $N(\mathbf{0}, \hat{\mathbf{V}}_W)$ random variable.

The relation of $\hat{\boldsymbol{\theta}}_{\mathrm{ER}}(\hat{\boldsymbol{\theta}}_W)$ and $\hat{\boldsymbol{\theta}}_{\mathrm{IS}}(\hat{\boldsymbol{\theta}}_W)$ to $\hat{\boldsymbol{\theta}}_W$ is illustrated in Figure 1. The well-known highly nonnormal nature of the sampling distributions of these estimators is illustrated in the simulations, in Figure 3.

3.3. *MSE of the model selection estimators in the complete data case.* Theorem 3.1 gives the rank ordering of the mean squared error (MSE) for the Impute-Select estimator, the Efron's Rules estimator, and the wide estimator, as a function of the magnitude of the true parameter value $\|\boldsymbol{\theta}^*\|$, in the complete data case and where any dependence between columns of $\mathbf{X}$ is not too strong. The MSE of the wide estimator, which always includes all variables, does not depend on $\boldsymbol{\theta}^*$ and has the smallest minimax risk. For independent predictors the following three statements hold: (1) For very small effect sizes, when it is appropriate to leave the corresponding predictor out of the model, the model selection estimator has the smallest risk. (2) For large effect sizes, where the corresponding variable will be selected with high probability, the choice of estimator is unimportant. (3) For moderate effect sizes where there is considerable uncertainty regarding the variable selection, the model averaged estimator performs best and can have much better minimax risk than the model selection estimator. Figure 2 illustrates these results for the special case $\mathbf{X}'\mathbf{X} = \mathbf{I}$, as has been presented by simulation in Hjort (2014). An appeal to the dominated convergence theorem then shows that these

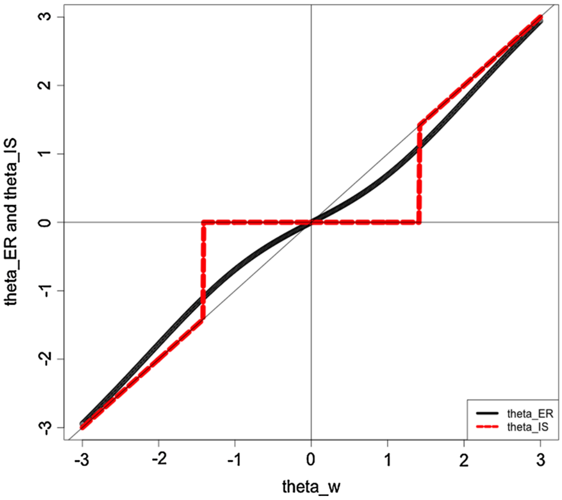

FIG. 1. *The model selection estimator* $\hat{\theta}_{\text{IS}}(\hat{\theta}_W)$ (*dotted red line*) *and the bootstrap averaged estimator* $\hat{\theta}_{\text{ER}}(\hat{\theta}_W)$ (*solid black line*) *as functions of the wide estimator* $\hat{\theta}_W$ *which always includes the predictor variable, with selection by* $\text{AIC}(\hat{\theta}_W)$. *Shown is the univariate case where* $s = \varnothing$ *or* $s = \{x\}$, *scaled so that* $\sigma/\sqrt{n} = 1$. *Here,* $\hat{\theta}_{\text{IS}}$ *is set to zero for* $|\hat{\theta}_W| < \sqrt{2\sigma/n}$, *and otherwise is equal to* $\hat{\theta}_W$. *As shown by equation* (3.2) $\hat{\theta}_{\text{ER}}(\hat{\theta}_W)$ *is a smoothed version of* $\hat{\theta}_{\text{IS}}(\hat{\theta}_W)$, *smoothed by the gaussian kernel which is the sampling distribution of* $\hat{\theta}_W$.

statements still hold for correlated predictors, as long as the correlation is small enough. For more general correlation structures, the situation is more complicated and the reader is referred to Claeskens and Hjort (2008b).

THEOREM 3.1 (MSE of $\hat{\boldsymbol{\theta}}_{\text{IS}}$ and $\hat{\boldsymbol{\theta}}_{\text{ER}}$). *Let* $\boldsymbol{\theta}^*$ *be the true value of* $\boldsymbol{\theta}$, *and consider an i.i.d. normal regression model as above. Let* $\mathbf{X}^j$ *be the* j*th column of* $\mathbf{X}$, *and let* $\max_{j \neq k} \mathbf{X}'^j \mathbf{X}^k < \nu$. *Then, if* ν *is small enough*:

1. *For small values of* $\|\boldsymbol{\theta}^*\|$,
$$\text{MSE}(\hat{\boldsymbol{\theta}}_{\text{IS}}) < \text{MSE}(\hat{\boldsymbol{\theta}}_{\text{ER}}) < \text{MSE}(\hat{\boldsymbol{\theta}}_W).$$
2. *For intermediate values of* $\boldsymbol{\theta}^*$,
$$\text{MSE}(\hat{\boldsymbol{\theta}}_{\text{ER}}) < \text{MSE}(\hat{\boldsymbol{\theta}}_{\text{IS}}).$$
3. *The minimax risk of* $\hat{\boldsymbol{\theta}}_{\text{ER}}$ *is less than that of* $\hat{\boldsymbol{\theta}}_{\text{IS}}$. *In particular,*
$$\max_{\boldsymbol{\theta}^*} \text{MSE}(\hat{\boldsymbol{\theta}}_W) < \max_{\boldsymbol{\theta}^*} \text{MSE}(\hat{\boldsymbol{\theta}}_{\text{ER}}) < \max_{\boldsymbol{\theta}^*} \text{MSE}(\hat{\boldsymbol{\theta}}_{\text{IS}}).$$
4. *As* $\|\boldsymbol{\theta}^*\| \to \infty$, $\text{MSE}(\hat{\boldsymbol{\theta}}_{\text{ER}}) \to \text{MSE}(\hat{\boldsymbol{\theta}}_W)$ *from above. The same holds for* $\text{MSE}(\hat{\boldsymbol{\theta}}_{\text{IS}})$.

The proof is given in the Supplementary Material Section A (Liu et al. (2019)).

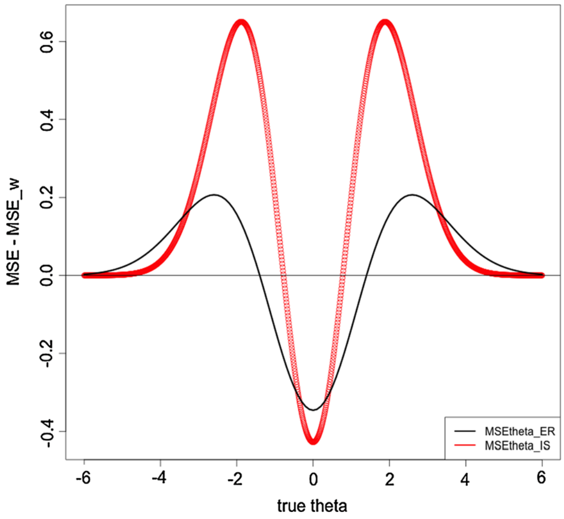

FIG. 2. *The difference* $\mathrm{MSE}(\hat{\theta}_{\mathrm{IS},i}) - \mathrm{MSE}(\hat{\theta}_{W,i})$ (*heavy red line*) *and* $\mathrm{MSE}(\hat{\theta}_{\mathrm{ER},i}) - \mathrm{MSE}(\hat{\theta}_{W,i})$ (*black line*) *as a function of the true value* θ_i^* (x-*axis*). *As in Theorem* 3.1, *the model is an i.i.d. normal linear regression with variable selection by AIC, in the canonical case* $\mathbf{X}'\mathbf{X} = \mathbf{I}$ *with complete data and known* σ, *scaled so that* $\sigma^2/n = 1$. *As seen,* $\hat{\theta}_{\mathrm{IS}}$ *is best for very small effect sizes of* θ^*, $\hat{\theta}_{\mathrm{ER}}$ *is much better than* $\hat{\theta}_{\mathrm{IS}}$ *for intermediate effect sizes, and it doesn't matter which estimator is used for large effect sizes. Plotted are equations* (1.3), *and equation* (1.4) *minus* 1 *from the proof of Theorem* 3.1 (*found in the Supplementary Material*); *the graph is similar to the simulation results in Hjort* (2014).

3.4. *Extension to predictors with missing values and imputation.* As in Section 2.1 suppose some elements of $\mathbf{X}^j$ are unobserved, that is $\delta_{ij} = 1$ with probability π_j, and let $\mathbf{X}_I^j$ denote the imputed dataset under an imputation model. We assume the imputation satisfies a classical measurement error model, $x_{Iij} = x_{ij} + \delta_{ij}\gamma_{ij}$, where the imputation errors γ_{ij} are Gaussian with mean zero and variance $\sigma^2_{\gamma_j,n}$, and where the true value x_{ij}, the missing indicator δ_j, and the imputation error γ_j are independent. Further, suppose $\sigma^2_{\gamma_j,n} \to 0$ as $n \to \infty$, so that x_{Iij} is consistent for x_{ij}. Suppose we form the estimate $\hat{\tilde{\boldsymbol{\theta}}}$ from $\mathbf{Y}$ and $\mathbf{X}_I$, using ordinary least squares as in the complete data case. Then in the simple case in which $\mathbf{X}^j$ is the only variable with missing values, $\hat{\tilde{\theta}}_j$ is a consistent estimator of $\tilde{\theta}_j^* = \theta_j^* \frac{\sigma^2_{X^j|X^{-j}}}{\sigma^2_{X^j|X^{-j}} + \pi\sigma^2_\gamma}$. A similar but more complicated formula for $\tilde{\boldsymbol{\theta}}$ holds in the general case (Carroll et al. (2006)).

COROLLARY 3.1 (Extension to imputation under a classical error model). *Under the above assumptions, as* $n \to \infty$, *and for* ν *small enough, Theorem* 3.1 *holds with* $\tilde{\boldsymbol{\theta}}$ *substituted everywhere for* $\boldsymbol{\theta}$, *with probability approaching* 1.

In summary, we expect the bootstrap-averaged "Efron's Rules" Impute-Select estimator to have better MSE than the single-step Impute-Select estimator in the case of moderate uncertainty regarding the inclusion of variables, and moderate correlation between variables. In the case of very small or large effect sizes, we would expect the Impute-Select estimator to have better performance. In the case of strong correlation between predictors, or a different measurement error model, the MSE relations may be more complex. The proof is given in the Supplementary Material Section A (Liu et al. (2019)).

4. Simulation. In this section we use simulation to compare (a) the bootstrap-averaged imputation-selection estimator, with variance estimates using Efron's Rules ($\hat{\boldsymbol{\theta}}_{\mathrm{ER}}$, Section 2.2); (b) the multiple imputation-selection estimator with variance estimates using Rubin's Rules ($\hat{\boldsymbol{\theta}}_{\mathrm{RR}}$, Section 2.3); and (c) a single imputation-selection step, with bootstrap percentile confidence intervals ($\hat{\boldsymbol{\theta}}_{\mathrm{IS}}$, Section 2.4). Confidence intervals for $\hat{\boldsymbol{\theta}}_{\mathrm{ER}}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{RR}}$ are based on the normal distribution. We study both a normal linear model and a logistic regression model, using simulation settings that have appeared in related literature. The R code for simulation is available in the Supplementary Material Section B.

Imputation uses the R package *mice* with default methods; that is, chained equations with predictive mean matching, using a normal imputation model for continuous variables and a logistic regression imputation model for dichotomous variables (van Buuren and Groothuis-Oudshoorn (2011)). For each simulation scenario below, all variables, including both predictors and response, are used in each imputation step. This imputation algorithm is proper in the sense of Rubin (1987) and Tsiatis (2006), in that the mth imputation is based on a new parameter $\hat{\beta}^m$, drawn from a $N(\hat{\beta}, \hat{\Sigma})$ distribution, where $\hat{\beta}$ and $\hat{\Sigma}$ are the maximum likelihood estimates for the imputation model (although the consistency arguments do not strictly apply in the case of predictive mean matching). As a sensitivity analysis, for the single-step Impute-Select estimator we also investigated improper imputation, using the maximum likelihood estimate $\hat{\beta}$ in each bootstrap imputation model.

Model selection uses the LASSO, and selects the tuning parameter with the minimum Akaike Information Criterion (AIC), as implemented in the R package *glmpath*. Standard maximum likelihood estimates of θ_j are then computed using the corresponding selected variables. Maximum likelihood estimators from the complete data and complete cases are included as benchmarks. If a variable is excluded from a model, both the parameter estimate and the associated variance estimate are set to 0.

The simulation performance metrics include the mean squared error (MSE) of $\hat{\theta}_j$ and the percent bias in the estimated standard deviation, $\hat{V}_j^{1/2}$, of $\mathrm{var}(\hat{\theta}_j)$. Performance of the $(1-\alpha)$ confidence interval is assessed by comparing coverage probability, median length and a metric combining interval coverage and length, namely the interval score of Gneiting and Raftery (2007), given by

$$S(\hat{l}, \hat{u}, \theta) = (\hat{u} - \hat{l}) + \frac{2}{\alpha}\big((\hat{l} - \theta)\mathbb{1}\{\theta < \hat{l}\} + (\theta - \hat{u})\mathbb{1}\{\hat{u} < \theta\}\big), \tag{4.1}$$

where the interval limits are $(\hat{l}, \hat{u})$ and $\mathbb{1}\{\}$ denotes the indicator function. We also compare the distributions of the estimators graphically. The simulation-based estimate of MSE is given by $\mathrm{MSE}(\hat{\theta}_j) = \sum_{r=1}^{R}(\hat{\theta}_{j,r} - \theta_j)^2/R$ and the variance of the estimate by $V_j = \sum_{r=1}^{R}(\hat{\theta}_{j,r} - \hat{\theta}_{j,\cdot})^2/(R-1)$, with $\hat{\theta}_{j,\cdot} = \sum_{r=1}^{R}\hat{\theta}_{j,r}/R$. The mean of the estimated variance $\hat{V}_j$ will be given by $\sum_{r=1}^{R}\hat{V}_{j,r}/R$. Percent bias of the estimated standard deviation is given by $100 \times (E[\hat{V}_j^{1/2}] - V_j^{1/2})/V_j^{1/2}$. The simulation size is $R = 500$.

4.1. *Linear regression model.* We consider an i.i.d. normal linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with $\epsilon_i \sim N(0, \exp(1.25))$, and use simulation scenarios and missing data mechanisms similar to Schomaker and Heumann (2014). There are six potential covariates, each with variance 1, plus an intercept (θ_0). The sample size is $n = 250$. Coefficients are

$$\boldsymbol{\theta} = (2.5, -3, -0.25, 0, -1.5, 0, 0.35).$$

Missing values are generated among the predictors so that about half of observations have at least one missing value. The data are MAR, but with strong dependence of the missing mechanism on the values of observed predictors. Approximately 27% of values for X_1 are missing, with probability of observation depending on y, 15% of values for X_4 are missing, depending on X_2, and 18% of values for X_5 are missing, depending on X_3, see Schomaker and Heumann (2014). The covariates are simulated as random draws from $X_1 \sim N(0.5, 1)$, $X_2 \sim \mathrm{Lognormal}(0.5, 0.5)$, $X_3 \sim \mathrm{Weibull}(1.75, 1.9)$, $X_4 \sim \mathrm{Exp}(1)$, $X_5 \sim \mathrm{Gamma}(0.25, 2)$ and $X_6 \sim N(0.25, 1)$. Note that all but the first and last are heavy-tailed compared to a normal distrubtion. Moderately strong dependence between variables is generated using the R package *copula*; correlations ranged from 0.21 to 0.50. Imputation increased these correlations somewhat, and the imputation error was correlated with the missing values, with correlations for X_1, X_4 and X_5 equal to -0.47, -0.58, -0.61, respectively. The bootstrap size was taken to be $B = 200$, and the imputation size was $M = 5$. As a sensitivity analysis, we also report $M = 200$ and $B = 400$.

4.1.1. *The distribution of the post-selection estimators.* As expected from equation (3.1), $\hat{\theta}_{\mathrm{IS}}$ appears to be a mixture of quasi-normals (each conditional on $\boldsymbol{\theta}_W \in \mathcal{A}_s$), each centered at the best estimate for a given selected best model. As expected from equation (3.2), $\hat{\theta}_{\mathrm{ER}}$ appears to be a model averaged version of $\hat{\theta}_{\mathrm{IS}}$, smoothed against a normal distribution centered at $\hat{\theta}_W$ with variance ~ 0.10. The Rubin's Rule estimator is intermediate between these two, smoothed somewhat by

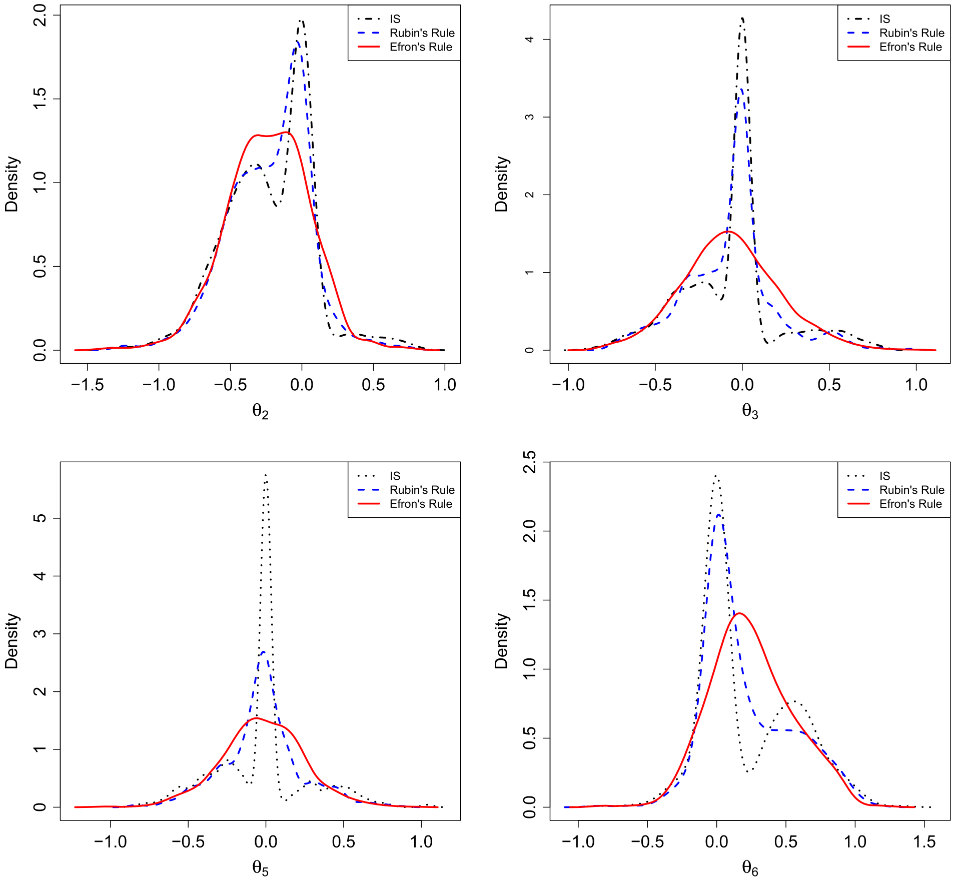

FIG. 3. *Sampling distribution of $\hat{\theta}_j$ for $j = 2, 3, 5, 6$ for the linear regression model of Section 4.1. IS: Impute-Select estimator, $\hat{\theta}_{\mathrm{IS}}$. Rubin's Rule: Rubin's Rules estimator, $\hat{\theta}_{\mathrm{RR}}$. Efron's Rule: Efron's Rules estimator, $\hat{\theta}_{\mathrm{ER}}$. As expected from equation* (3.1), *$\hat{\theta}_{\mathrm{IS}}$ appears to be a mixture of conditional normals. As expected from equation* (3.2), *$\hat{\theta}_{\mathrm{ER}}$ appears to be a model-averaged version of $\hat{\theta}_{\mathrm{IS}}$, smoothed against a normal distribution centered at the wide model estimate. $\hat{\theta}_{\mathrm{RR}}$ is intermediate between these two, smoothed somewhat by averaging over the imputation distribution conditional on the observed sample.*

averaging over the imputation distribution rather than the bootstrap distribution. The distribution of the Rubin's Rules estimator does not appear to change appreciably using $M = 200$ (not shown).

4.1.2. *MSE of estimated coefficients.* The MSE for each method relative to the complete data analysis is given in Table 1. For comparison with Theorem 3.1 we also include the wide model, which uses a single imputation step but no variable selection; all of the other scenarios incorporate variable selection using AIC. The complete case analysis excludes any case with missing data.

TABLE 1

Comparison of post-imputation-selection estimators in a linear regression model: relative MSE of $\hat{\theta}_j$, % bias in $\hat{V}_j^{1/2}$, and mean interval score for a 95% Confidence Interval (CI), from 500 simulation runs. For Rubin's Rules, $M = 5$, and for the bootstrap, $B = 200$

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
Coefficient θ_j:	−3.0	−0.25	0	−1.5	0	0.35
Missing data rate:	27%	0	0	15%	18%	0
Complete Data MSE:	0.09	0.07	0.06	0.07	0.04	0.11
Relative MSE of $\hat{\theta}_j$						
Complete Case Analysis*	7.9	2.5	1.4	3.7	2.0	1.4
Wide Model: Single imputation#	1.7	1.6	1.7	2.0	2.5	1.0
Rubin's Rules: Multiple Impute-Select	1.5	1.2	1.2	1.6	1.4	1.0
Impute-Select	1.7	1.3	1.4	2.1	1.9	1.2
Efron's Rules: Bootstrap-average Impute-Select	1.4	1.2	1.3	1.5	1.6	0.9
Percent bias in $\hat{V}_j^{1/2}$						
Complete Case Analysis*	−10	−51	−57	−15	−62	−61
Wide Model: Multiple Impute##	−4	<1	1	−3	4	3
Rubin's Rules: Multiple Impute-Select	−7	−18	−25	−4	−16	−31
Impute-Select: Bootstrap variance	2	8	10	−3	16	3
Efron's Rules: Bootstrap-average Impute-Select	−7	−5	−4	−9	0	1
Interval Score for 95% confidence interval+						
Complete Case Analysis*	9.3	6.4	0.8	3.3	0.9	9.1
Rubin's Rules: Multiple Impute-Select	1.7	2.7	1.0	1.7	1.0	5.0
Impute-Select: Bootstrap percentile CI	1.7	1.5	1.3	1.7	1.3	1.7
Efron's Rules: Bootstrap-average Impute-Select	1.6	1.4	1.3	1.8	1.3	1.6

*Naive estimate which does not incorporate selection variability. #Missing data is imputed and all variables are included, without selection; compare Impute-Select, and see Theorem 3.1 and Figure 2. ##Variance estimates by Rubin's Rules, $M = 5$, which is seen to work well if there is no variable selection. +Mean interval score, given by equation (4.1). Lower is better.

As expected, using complete cases only gives a much higher MSE than using the complete data, by a factor ranging from 1.4 to 7.9. All of the imputation-selection based methods have smaller MSE than the complete case analysis. The two model-averaged estimators, Efron's Rules and Rubin's Rules, have similar MSE to each other. Single imputation (Impute-Select) has uniformly higher MSE than the two model-averaged estimators, up to 39% higher than Efron's Rules for θ_4. Hence both Efron's Rules and Rubin's Rules appear to benefit from their model averaging, perhaps as might be expected from Theorem 3.1 for coefficients with moderate true effect sizes. Consistent with Theorem 3.1 and Figure 2, the wide model (with a single imputation and without variable selection) has MSE similar to the Impute-Select method for large coefficients, and a larger MSE by up to 25% for the zero coefficients, indicating that model selection is favorable in this scenario.

Looking more closely, Theorem 3.1 and its Corollary also suggest that the Impute-Select MSE should be a little *lower* than the MSE for the model-averaged estimators for θ_3 and θ_5, where the true effect size is zero, and this was not the case. This may be due to the small effect size for the difference at 0 (see Figure 2), or to the moderately strong dependence structures we have incorporated in both the data generating model and in the missing data mechanism. For example, the imputed version of X_3 has correlation of over 0.40 with each of the strong predictors X_1 and X_6, and the correlation between the unobserved values and the imputation error is over 0.40. Thus, either the idealized case of the Theorem and its Corollary does not appear to apply, or the MSE advantage of Impute-Select was too small to be picked up by our simulation. It is also possible that Impute-Select has a relative disadvantage because it does not average over the imputation distribution (i.e., Impute-Select uses $M = 1$, contrasted with the Rubin's Rules use of $M = 5$ or $M = 200$, or the Efron's Rules average over 200 bootstrap draws). However, in a sensitivity analysis, use of improper imputation in which the imputation model uses the maximum likelihood estimates for each imputation draw, the relative MSE rankings between Impute-Select and Efron's rules did not change. Also, these results on relative MSE do not change appreciably using a larger bootstrap size of $B = 400$. Further elucidating model-selection vs model-averaging MSE rankings is left to future work.

4.1.3. *Bias of estimated standard errors.* As expected, the naive complete case estimator, which ignores the model selection variability, has standard error estimates which are biased downwards, by up to 62% in this example (Table 1). These variance estimates are reasonable only for the two coefficients with large effect sizes, θ_1 and θ_4, because these variables are nearly always selected into the model. Both the Impute-Select estimator, with its bootstrap-based variance estimate, and the Efron's Rules estimator, with variance estimated from equation (2.3), give reasonably unbiased estimates of standard error, within 16% of the true value. On the other hand, the Rubin's Rules variance estimator is biased downwards, by up to 31% for the smaller effect sizes. This bias is somewhat reduced by increasing the imputation size from $M = 5$ to $M = 200$, but remains above 25%. Note that the Rubin's Rules downward bias in the variance estimate occurs despite the use of a proper or nearly proper imputation rule, and is entirely due to the failure to fully incorporate the selection variability into the post-selection variance estimate; performance of the Rubin's Rules variance estimate for the wide estimator (with no variable selection) is comparable to the Impute-Select and the Efron's Rules estimators. The greater accuracy of the two bootstrap based imputation-selection methods is at the cost of greater CPU time (about 8 seconds) compared to imputation-selection with Rubin's Rules (0.03 seconds, $M = 5$; 1.25 seconds $M = 200$). Results are not qualitatively changed by increasing the bootstrap size to 400 or the imputation size to 200.

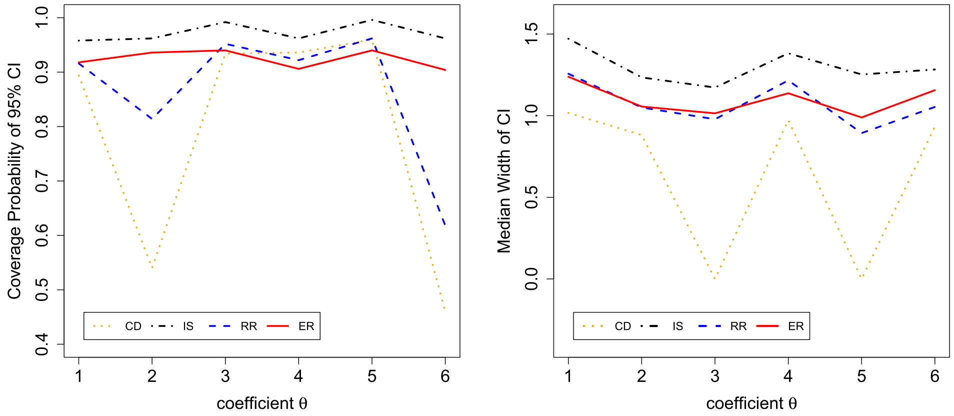

FIG. 4. *Coverage probability* (*left*) *and median width* (*right*) *of* 95% *confidence interval* (*CI*), *for post-imputation-selection estimators in the linear regression model of Section* 4.1. *The* x*-axis indexes the coefficient*. *CC*: *Naive complete case estimator. IS*: *Impute-Select*, *with bootstrap confidence interval. RR*: *Multiple imputation followed by model selection and Rubin's Rules*. *ER*: *Bootstrap imputation followed by model selection and Efron's Rules. Both the IS and ER estimators have coverage above* 90%; *the CI width for the ER estimator is smaller because of model averaging*.

4.1.4. *Confidence interval coverage*, *precision and interval score*. Coverage probabilities of the nominal 95% confidence intervals are given in Figure 4(left). The naive complete case method, which neglects to incorporate the model selection variability, has coverage probabilities well below 50% for some coefficients. The Rubin's Rules estimator has coverage close to 60% for some coefficients, and this is not appreciably improved by using $M = 200$ (not shown). The bootstrap-percentile confidence interval of the Impute-Select method has the best coverage probability, at or above the nominal 95% level, and the normal distribution based Efron's Rules interval has coverage near 90% for all coefficients. Increasing the bootstrap size to $B = 400$ increases coverage of both the Efron's Rules and the impute select estimator, for some but not all coefficients.

The median width of these intervals is given in Figure 4(right). On average, the Efron's Rules width is about 15% (range 10–21%) less than Impute-Select across the coefficients, reflecting the benefit from model averaging.

The mean interval score is given in Table 1. Impute-Select and Efron's Rules score almost identically, for both $B = 200$ and $B = 400$, reflecting the trade off between width of the confidence intervals and coverage probability. The Rubin's Rules post-selection estimator can score poorly and is not recommended.

4.1.5. *Bootstrap*, *imputation and simulation sizes*. The bootstrap size B was checked using the jacknife-after-bootstrap, following Efron (2014) (equation 3.11), with $J = 10$ groups in the cross-validation. In Table 1, the estimated standard error of the $\hat{V}_j$ divided by the estimate mean value for $\hat{V}_j$, is under 10%

for the wide model, but this increased to under 20% for the Efron's Rules estimator, indicating that increasing the bootstrap size from $B = 200$ to $B = 400$ might further improve performance of the Efron's Rules estimator. However, when we tried $B = 400$, results remained largely similar. A similar jackknife calculation with $J = 20$ showed the simulation size of $R = 500$ to be adequate, using the simulation estimate of the true MSE of the Efron's Rules estimator, where the relative standard error was under 10% for the six coefficients in this scenario. Also, note that the standard error of our coverage probabilities is about 1 to 2 percentage points, sufficient to identify the large differences in coverage observed between methods, indicating that R is adequate.

4.1.6. *Summary.* In this linear regression example with imputation followed by model selection, the Rubin's Rules estimator gives confidence intervals with very poor coverage and is not recommended. The two bootstrap-based estimators, Impute-Select and Efron's Rules, both provide confidence intervals with comparable and reasonable performance. The first has confidence intervals that are slightly too wide; the second has coverage that is slightly too small. Impute-Select and Efron's Rules perform almost identically when compared using the interval score, reflecting the trade-off between width of the confidence intervals and coverage probability, and so either one can be recommended.

The poor performance of the Rubin's Rules estimator can be attributed to downwardly biased estimates of variance, by up to 30%. This bias occurs despite our use of a proper or nearly imputation procedure, and it is not significantly improved by increasing the imputation size M. The issue is that the model selection variability is assessed only against the imputation distribution, and not against the full sampling distribution. To see this, consider that in steps 1 through 5 of the Rubin's Rules algorithm (Section 2.3), the entries of $\mathbf{X}_I$ that are observed remain fixed in the calculation of the variance estimate; only the entries that are imputed are varied. Thus, Rubin's Rules provides an estimate of the model selection variability assessed against the imputation distribution, but conditional on the observed data.

The two bootstrap-based estimators both have reasonable coverage, although neither appears to be optimal. The estimated variance of the Efron's Rules estimator is slightly biased downwards, by under 10%, and may also suffer somewhat because the normal-theory confidence intervals used are applied to sampling distributions which are not normal; however it has relatively low MSE. Single imputation with bootstrap confidence intervals adapts to the nonnormal distribution of the post-selection estimators, and appears to give valid inference, however at the cost of higher MSE and wider confidence intervals. While neither estimator is optimal, either estimator can be recommended.

4.2. *Generalized linear regression model.* We next simulate a logistic regression model based on the "Birthweight" data (Hosmer and Lemeshow (1989), Hjort and Claeskens (2003)). We study the association between mother's age (x_1), mother's weight in last menstrual period (x_2), race ($x_3 = 1$ for race "black", $x_4 = 1$ for race "other", with reference category "white") and the baby's birth weight ($y = 1$ if baby's birth weight is less than 2.5 kg and 0 otherwise). The outcome y for our simulation study is generated based on the fitted logistic regression model from the original complete data. There are 4 predictors plus the intercept, with $\boldsymbol{\theta} = (1.307, -0.025, -0.014, 1.004, 0.44)$. The sample size is $n = 189$ and the matrix of predictors $\mathbf{X}$ from the Hosmer and Lemeshow data is used.

As before, we use simulation to study the performance of the coefficient estimates $\hat{\theta}_j$ and $\hat{V}_j$. We also study predicted probabilities of low birthweight for race "white" and race "black", at average age and weight for these subgroups, given by $\hat{p} = \exp(\mathbf{X}\hat{\boldsymbol{\theta}})/(1 + \exp(\mathbf{X}\hat{\boldsymbol{\theta}}))$.

Missing values were randomly generated with probability of missingness $\sim$28% for weight in last menstrual period (LWT), 16% for race "black" and 10% for race "other", given by: $p(x_2) = (y + 0.007 * x_1^2)^{-1}$, $p(x_3) = 1 - (1 + 0.008 * x_1)^{-1}$, $p(x_4) = 1 - (1 + 0.005 * x_1)^{-1}$. The bootstrap size is $B = 200$, and the imputation size is $M = 5$.

4.2.1. *The distribution of the estimators.* Figure 5 shows the sampling distribution of the estimators for the coefficient of race(black) and for the predicted probability of low birthweight outcome for race(black) and average age and weight. Results are similar to the linear model case; the distribution of model outputs (predicted probabilities) generally appear to be closer to normal than the underlying coefficients.

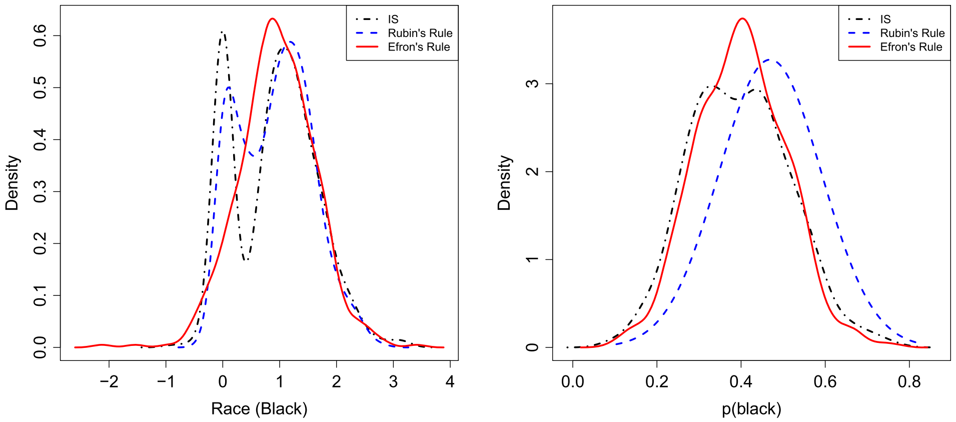

FIG. 5. *Sampling distribution of $\hat{\theta}_{\text{Black}}$ and of a predicted probability for the generalized linear model of Section 4.2. IS: Impute-Select estimator, $\hat{\theta}_{\text{IS}}$. Rubin's Rule: Rubin's Rules estimator, $\hat{\theta}_{\text{RR}}$. Efron's Rule: Efron's Rules estimator, $\hat{\theta}_{\text{ER}}$. As before, $\hat{\theta}_{\text{IS}}$ appears to be a mixture of conditional normals and $\hat{\theta}_{\text{ER}}$ appears to be a model averaged version of $\hat{\theta}_{\text{IS}}$. $\hat{\theta}_{\text{RR}}$ is intermediate between these two.*

TABLE 2

Comparison of post-imputation-selection estimators in a generalized linear model: relative MSE of $\hat{\theta}_j$, % bias in $\hat{V}_j^{1/2}$, and mean interval score for a 95% *confidence interval (CI), from* 500 *simulation runs*

	Weight	Race (black)	Race (other)	$p(white)$	$p(black)$
Coefficient θ:	−0.014	1.004	0.44	0.230	0.414
Missing data rate:	28%	16%	10%	N/A	N/A
Complete data MSE:	0.0001	0.398	0.163	0.002	0.013
Relative MSE of $\hat{\theta}$, $\hat{p}$					
Complete Case Analysis*	2.23	1.99	1.76	2.43	1.89
Rubin's Rules: Multiple-Impute-Select	1.37	1.05	1.07	1.10	0.98
Impute-Select: Bootstrap percentile CI	1.65	1.33	1.19	1.24	1.12
Efron's Rules: Bootstrap-average Impute-Select	1.35	1.12	1.01	0.97	0.88
Percent bias in $\hat{V}_j^{1/2}$					
Complete Case Analysis*	−38	−53	−54	−34	−46
Rubin's Rules: Multiple-Impute-Select	−8	−17	−25	−13	−30
Impute-Select: Bootstrap variance	8	19	4	6	4
Efron's Rules: Bootstrap-average Impute-Select	−6	−5	−2	−6	−7
Interval Score for 95% confidence interval$^+$					
Complete Case Analysis*	0.19	19.6	10.6	0.51	1.27
Rubin's Rules: Multiple-Impute-Select	0.06	7.1	5.1	0.29	0.80
Impute-Select: Bootstrap percentile CI	0.04	3.1	2.0	0.23	0.54
Efron's Rules: Bootstrap-average Impute-Select	0.04	3.2	1.9	0.26	0.63

*Naive estimate which does not incorporate selection variability. $^+$Mean interval score, given by equation (4.1). Lower is better.

4.2.2. *MSE of estimates.* The MSE for each method relative to the complete data analysis is given in Table 2 for selected coefficients. Results are similar to those for linear regression for both estimated coefficients and estimated probabilities. In this setting, results appear to be concordant with expectations from Theorem 3.1, as these effect sizes may be considered moderate.

4.2.3. *Bias of estimated standard deviations.* As before, the naive complete case analysis underestimates standard errors by 20% to 54%. Both Impute-Select and Efron's Rules give reasonably unbiased estimates. Rubin's Rules is again substantially biased downwards for all estimated standard errors, up to 30% below the true value. The CPU time takes about 16 seconds for Efron's Rule compared to 0.06 seconds for Rubin's Rules.

4.2.4. *Confidence interval coverage, precision and interval score.* Coverage probabilities of the nominal 95% confidence intervals are given in Figure 6(left).

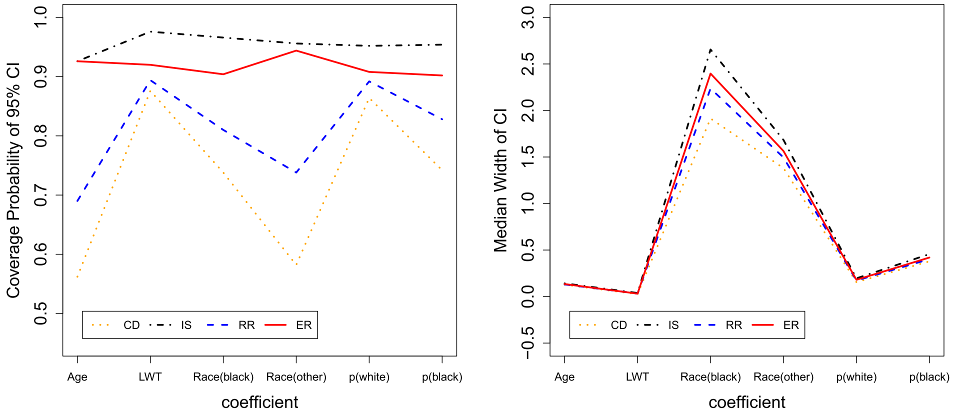

FIG. 6. *Coverage probability* (*left*) *and median width* (*right*) *of* 95% *confidence interval* (*CI*), *for the generalized linear model of Section* 4.2. *CD*: *Naive complete data estimator. IS*: *Impute-Select estimator, with bootstrap percentile confidence interval. RR*: *Multiple imputation followed by model selection and Rubin's Rules. ER*: *Bootstrap imputation followed by model selection and Efron's Rules. Both the IS and ER estimators have coverage near* 90%. *The CI width for the ER estimator is a little smaller because of model averaging.*

The naive complete data estimator and the Rubin's Rules estimator again have unacceptably low coverage. As before, the bootstrap-percentile confidence interval of the Impute-Select method has coverage probability near the nominal 95%, and Efron's Rules has coverage above 90% for all coefficients. As before, the Impute-Select estimator gives the widest CI, up to 23% wider than Efron's Rules (Figure 6(right)). Comparing these estimators using the interval score, Impute-Select and Efron's Rules score almost identically, reflecting the trade-off between width of the confidence intervals and coverage probability. The Rubin's Rules estimator again scores poorly and is not recommended.

5. Application to post-polypectomy risk modeling.

Here we apply the three approaches for post-imputation selection inference to the analysis of colorectal cancer risk which motivated the present paper (Liu et al. (2016)). In the original study, data from 8228 individuals were pooled from seven different prospective studies with baseline polypectomy and repeat surveillance colonoscopy within 3 to 5 years. These subjects were randomly assigned to training and validation datasets. We used the 5483 subjects in training dataset in the following analysis. The outcome was presence of advanced neoplasia on followup. As in our original study, we considered as potential predictors the three variables used by the US guidelines, namely number, size and histopathology of resected polyps, plus the five additional known risk factors of age, sex, history of prior polyps (yes/no), polyp grade (high grade or other) and polyp location (distal only, proximal only, both). This set of eight predictors was selected by univariate screens at $p < 0.15$ in the original study.

Because one or more studies failed to collect several variables, three important predictors had high rates of missing data (history of polyps: 21.4%, histopathologoy: 11.1% and grade: 24.3%). The majority of data are close to MCAR in this example, and so a complete case analysis would be unbiased, but, given the high rate of missing data, underpowered. Hence, in the original study, we used indicator variables for the missing data, despite the known bias of that approach in the case of correlated predictors, even if the data are missing completely at random (Jones (1996)). Because correlation between predictors was low, we expected the induced bias to be modest. For model selection, we used an ad hoc combination of bootstrap-based LASSO regression using AIC, together with Bayesian model averaging using the Bayesian Information Criterion, and used the set of variables that was deemed most compatible with these two approaches. The final selected model was then applied to independent validation data to make population level estimates.

Here we apply an imputation model to fill in the missing values, and then apply the three post-imputation-selection methods for inference. We compare performance of the three methods and consider the implications for our original study. We focus here on the estimated coefficients of the model for the variables history of prior polyps (hx: indicator of positive history), polyp grade (hiGd: indicator of high grade) and polyp location (Prox: indicator for proximal location; ProxDist: indicator for both proximal and distal polyps). These variables were chosen for investigation because they are known to be clinically important and because they have large univariate estimated effect sizes on risk of future advanced neoplasia. In our published study, history and location were selected into the final model, and grade was not. Age was also selected into the final model with strong estimated effects by all methods considered, so we don't report it here. We also investigate predicted probabilities from three different risk scenarios, as specified in Table 3. Here, for each combination of polyp location and grade, other variables were set to their median values, as denoted in the footnote to the table. Imputation used the R package *mice* with $M = 5$ for the multiple imputation approach. We used LASSO for variable selection in a logistic regression model with criterion AIC, as in our original study. The bootstrap sample size was $B = 300$.

TABLE 3
Colonoscopy cohort data: *scenarios for estimated risks*

	Polyp location	Polyp grade
Risk scenario		
1	Proximal only	High
2	Proximal and distal	High
3	Proximal only	Low

Scenarios 1, 2, 3, respectively: Male, age 70, 64, 68; size of largest polyp 10,15,10; number of polyps 1, 3.5, 1.

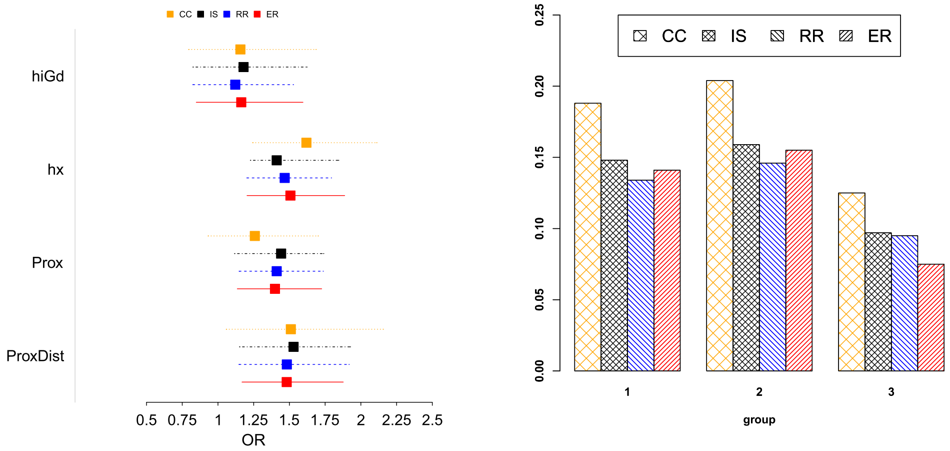

FIG. 7. *Left: forest plot showing 95% confidence intervals for four selected parameters; Right: confidence interval widths for estimated probabilities for three different risk scenarios. Methods are: CC, complete case; IS, Impute-Select; RR, Rubin's Rules; ER, Efron's Rules. Variables are indicators for: hiGd, high grade; hx, history of prior polyps; Prox, proximal location only, ProxDist, both proximal and distal locations.*

5.1. *Results for the colonoscopy data.* Estimated odds ratios were generally close across the imputation methods (each within 3% to 7%; Figure 7, left panel). These estimates were also within 3% to 7% of the published study estimates, consistent with minimal bias from using the indicator method for missing values in our original analysis. There were larger observed differences for the complete case estimates, which also had the widest confidence intervals, confirming that the complete case approach is not recommended.

The Rubin's Rules method was seen to have the narrowest confidence intervals for several coefficients, but is not recommended because of its demonstrated low coverage in our simulation studies above. The Efron's Rules method gave generally, but not always, narrower confidence intervals than Impute-Select with bootstrap confidence intervals. We expect that this would be balanced by somewhat lower coverage, as shown by the interval score from our simulation studies, so that either of these two bootstrap-based methods can be recommended.

Estimated probabilities for the three risk scenarios in Table 3 were again similar across methods (within 5%) and did not differ too much from the complete case analysis (within 8%; data not shown). Comparing the two recommended estimators, Efron's Rules gave the narrowest intervals, up to 23% narrower than Impute-Select for scenario 3. However, as shown by the interval score in our simulation studies, we expect the increased precision of Efron's Rules compared to Impute-Select might be balanced by somewhat lower coverage, although in the smoother setting of estimating model output Efron's Rules might be expected to retain some advantage.

Had we used either recommended imputation-based method in our original study, we would have arrived at the same final model, adding the subject's age, history of polyps and polyp location to the three risk factors currently incorporated into the US guidelines. Thus, fortunately, our ad hoc approach using indicator variables for missing values worked adequately in this case. However, in other cases with highly correlated predictors or data not MCAR, the indicator method would be expected to fail, while Impute-Select or Efron's Rules could still be recommended.

6. Discussion. In this paper we have studied the performance of three practical imputation-based methods to construct confidence intervals after model selection when there is missing data: a bootstrap-based method which uses a single imputation-selection step to produce the final model, followed by bootstrapped imputation-selection for inference; a related bootstrap-based method which uses the average of the imputed and selected models over the bootstrap distribution as the final model, with computationally efficient confidence intervals following results in Efron (2014); and a method based on multiple imputation-selection that was originally studied in Schomaker and Heumann (2014) and which uses Rubin's Rules for inference. The Rubin's Rules method showed severe under-coverage and is not recommended, while the two bootstrap-based methods performed reasonably well with similar computational load, and either one can be recommended.

Both the Rubin's Rules estimator and the Efron's Rules estimator produce model-averaged final estimates, in contrast to the single model produced by the Impute-Select approach. The Rubin's Rules estimator averages the selected models across the imputation-selection distribution. The Efron's Rules estimator averages the selected models across the bootstrap imputation-selection distribution. Following the model averaging literature, we expected that both model-averaged estimators might often have lower MSE than the single imputation estimator, and this was confirmed by our simulations.

Despite its good MSE, the Rubin's Rules post-selection estimator cannot be recommended, as it consistently underestimated the standard deviation by up to 25% to 30%. Variance under-estimation was also seen in a prior publication (Schomaker and Heumann (2014)), which our simulation studies confirmed. It is well known that an "improper" imputation method will cause the Rubin's Rules estimate of variance to be biased downwards. To construct the imputations, we used chained equations as implemented in the popular R package *mice*, with the robust predictive mean matching approach. This is a "proper" or nearly proper imputation method, in that it explicitly incorporates the estimated sampling variability of the parameters for the imputation model into the imputation mechanism. However, despite our explicit use of a proper imputation rule, large downward bias remained. Furthermore, while this downwards bias did not improve with increasing imputation size M, it disappeared in a comparable model with no model selection. Comparison with the two bootstrap-based methods below helps understand why.

Both the Efron's Rules and Impute-Select estimators produced reasonable estimates of variance and adequate confidence interval coverage in all our simulation scenarios. Each of these estimators (1) resamples the data, (2) then estimates the imputation model and uses it to impute missing data in the bootstrap sample, (3) then does model selection on the bootstrapped and imputed dataset. Steps 1 and 2 ensure via the bootstrap that the sampling variability of the imputation model parameters is incorporated into the imputation mechanism, thus the imputation is "proper". Step 3 then assures that this proper imputation variability and *also* the underlying sampling variability are incorporated into the selection mechanism. Consider that, for Rubin's Rules, at each multiple imputation step only the missing observations are replaced: thus, the Rubin's Rules estimator assesses the variability of model selection against the imputation distribution, conditional on the selected sample. By contrast the two bootstrap-based methods use the full sampling distribution of the imputation-selection mechanism to assess variation, and produced acceptable estimates of variance.

In our simulation examples, the Efron's Rules model-averaged estimator always had better MSE than the single Impute-Select estimator. This was expected from our Theorem in the case of moderately large true effect sizes, but it did not agree with our Theorem in case a variable has zero or nearly zero effect (and thus might be omitted from the model). In the no-effect case, single Impute-Selection would be predicted to have slightly smaller MSE than model-averaged imputation-selection. It is possible that our simulation was not sharp enough to pick up the difference. Also, the Theorem was proved in an idealized setting of small dependence and small imputation variability. However, our simulation scenarios were deliberately chosen to be far from ideal, with strong dependence between predictors and within the imputation error structure. While both the Impute-Select and the Efron's Rules estimators worked very well in these messy but realistic settings, the results of our Theorem on MSE did not always apply. Reassuringly, the Theorem did correctly predict that omitting the variable selection step (the "wide" estimator) performed about as well as Impute-Select for larger effect sizes, but worse for the zero effect sizes; indeed, the wide estimator performed the worst among the methods studied, for our selected scenarios.

There are other methods that have been proposed in the literature, including a bootstrapped version of a weighted model-averaged estimator, weighted over all possible sub-models, which also incorporates multiple imputation (Schomaker and Heumann (2014)). This approach, while computationally complex, produces good coverage confidence intervals. Also, a recent paper (Schomaker and Heumann (2018)) discusses various approaches to bootstrapping multiple imputation, without addressing model selection. These approaches might combine the potential efficiency advantages of large M in Rubin's Rules, with the correct coverage of bootstrap confidence intervals, however at what we suspect may be the cost of some computational redundancy; this might be the subject of future investigation. Also, elucidating the kinds of dependence under which single imputation-selection will have better MSE than bootstrap-averaged imputation-selection remains the subject of future work. Finally, we note that our model selection step is conditional on the imputation, and so does not incorporate the imputation variability into the selection criteria. Other methods have been proposed which do properly incorporate the imputation variability into the selection statistics (Claeskens and Consentino (2008), Wood, White and Royston (2008)), and these approaches might be expected to choose a better model in some circumstances. Post-imputation-selection inference remains to be studied in this setting.

Efron (2014) introduced the model averaged bootstrap approach used here as a "more dependable" way of setting confidence intervals for the "jumpy" estimates produced by model selection—the model averaging smooths out the multi-modal sampling distribution of post-selection estimators. Indeed we see this behavior in for example, Figure 3 from our simulations. Interestingly, however, the confidence intervals of the bootstrap-averaged approach have lower coverage than the raw bootstrap percentile confidence intervals for the very jumpy single step Impute-Select estimator. The Efron's Rules estimator uses a symmetric normal confidence interval for $\hat{\theta}_{\mathrm{ER}}$, which however still has a somewhat nonnormal post-selection sampling distribution, while the bootstrap distribution for the impute select estimator adapts appropriately to its highly nonnormal distribution. The advantage of the Efron's Rules estimator comes from its potential for improved MSE, but this does not always translate into improved coverage probability.

It is well known that the bootstrap can be inconsistent for the sampling distribution of post-selection estimators (Chatterjee and Lahiri (2010)); however it is less well appreciated that this can be a failure of convergence in probability, which does not necessarily imply a failure of convergence in distribution. Indeed, in a different simulation setting Chatterjee and Lahiri (2011) noted the good performance bootstrap confidence intervals for post-selection inference. Thus, an important message of the present paper is that the bootstrap remains a good choice for post-selection inference.

In sum, both the bootstrapped Impute-Select estimator and the bootstrap-averaged Efron's Rules estimator appeared to work well in our theoretical investigations and in our simulations. In our simulations, Impute-Select had slightly higher MSE than Efron's Rules, but this was balanced by its better coverage, inherited from using bootstrap percentile confidence intervals. Balancing coverage against confidence interval width, the interval scores for the Impute-Select estimator and the Efron's Rules estimator were similar, as is their computational burden. Thus either the Efron's Rules or the Impute-Select bootstrap-based estimators can be recommended in practice as a good choice for post-selection inference after imputation for missing data.

Acknowledgments. Thanks to Samir Gupta, Elena Martinez and coauthors on the original prognostic modeling study—Drs. Baron, Lieberman, Jacobs, Cross and Murphy—for graciously allowing use of the data. The authors are grateful to the Editor and referees for their suggestions which greatly improved the paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. *Conflict of Interest*: None declared.

SUPPLEMENTARY MATERIAL

Supplement to "Imputation and post-selection inference in models with missing data: An application to colorectal cancer surveillance guidelines" (DOI: 10.1214/19-AOAS1239SUPP; .zip). We provided the proofs of Theorem 3.1 and Corollary 3.1. Also we provided the R code for simulation results in Table 1 and Table 2.

REFERENCES

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417

CHATTERJEE, A. and LAHIRI, S. N. (2010). Asymptotic properties of the residual bootstrap for Lasso estimators. *Proc. Amer. Math. Soc.* **138** 4497–4509. MR2680074

CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. MR2847974

CLAESKENS, G. (2016). Statistical model choice. *Ann. Rev. Stat. Appl.* **3** 233–256.

CLAESKENS, G. and CONSENTINO, F. (2008). Variable selection with incomplete covariate data. *Biometrics* **64** 1062–1069. MR2522253

CLAESKENS, G. and HJORT, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98** 900–945. MR2041482

CLAESKENS, G. and HJORT, N. L. (2008a). Minimizing average risk in regression models. *Econometric Theory* **24** 493–527. MR2422864

CLAESKENS, G. and HJORT, N. L. (2008b). *Model Selection and Model Averaging*. *Cambridge Series in Statistical and Probabilistic Mathematics* **27**. Cambridge Univ. Press, Cambridge. MR2431297

EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

HEYMANS, M. W., VAN BUUREN, S., KNOL, D. L., VAN MECHELEN, W. and DE VET, H. C. W. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Med. Res. Methodol.* **7**.

HJORT, N. L. (2014). Comment [MR3265671]. *J. Amer. Statist. Assoc.* **109** 1017–1020. MR3265676

HJORT, N. L. and CLAESKENS, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98** 879–899. MR2041481

HOSMER, D. W. and LEMESHOW, S. (1989). *Applied Logistic Regression*. Wiley-Interscience, New York.

JONES, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J. Amer. Statist. Assoc.* **91** 222–230. MR1394076

LACHENBRUCH, P. A. (2011). Variable selection when missing values are present: A case study. *Stat. Methods Med. Res.* **20** 429–444. MR2829120

LIEBERMAN, D. A., REX, D. K., WINAWER, S. J., GIARDIELLO, F. M., JOHNSON, D. A. and LEVIN, T. R. (2012). Guidelines for colonoscopy surveillance after screening and polypectomy: A consensus update by the US multi-society task force on colorectal cancer. *Gastroenterology* **143** 844–857.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR1925014

LIU, L., MESSER, K., BARON, J. A., LIEBERMAN, D. A., JACOBS, E. T., CROSS, A. J., MURPHY, G., MARTINEZ, M. E. and GUPTA, S. (2016). A prognostic model for advanced colorectal neoplasia recurrence. *Cancer Causes Control* **27** 1175–1185. DOI:10.1007/s10552-016-0795-5.

LIU, L., QIU, Y., NATARAJAN, L. and MESSER, K. (2019). Supplement to "Imputation and post-selection inference in models with missing data: An application to colorectal cancer surveillance guidelines." DOI:10.1214/19-AOAS1239SUPP.

LONG, Q. and JOHNSON, B. A. (2015). Variable selection in the presence of missing data: Resampling and imputation. *Biostatistics* **16** 596–610. MR3365449

MARTINEZ, M. E., THOMPSON, P., MESSER, K. et al. (2012). One-year risk of advanced colorectal neoplasia: United States vs. United Kingdom risk-stratification guidelines. *Ann. Intern. Med.* **12** 856–864.

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519

SCHOMAKER, M. and HEUMANN, C. (2014). Model selection and model averaging after multiple imputation. *Comput. Statist. Data Anal.* **71** 758–770. MR3132004

SCHOMAKER, M. and HEUMANN, C. (2018). Bootstrap inference when using multiple imputation. *Stat. Med.* **37** 2252–2266. MR3810720

SIEGEL, R. L., MILLER, K. D. and JEMAL, A. (2015). Cancer statistics. *CA Cancer J. Clin.* **65** 5–29.

TANNER, M. A. and WONG, W. H. (1987). An application of impuation to an estimation problem in grouped lifetime analysis. *Technometrics* **29** 23–32.

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

WOOD, A. M., WHITE, I. R. and ROYSTON, P. (2008). How should variable selection be performed with multiply imputed data? *Stat. Med.* **27** 3227–3246. MR2523914

DIVISION OF BIOSTATISTICS AND
BIOINFORMATICS
DEPARTMENT OF FAMILY MEDICINE AND
PUBLIC HEALTH
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093
USA
E-MAIL: l2liu@ucsd.edu
yuq052@ucsd.edu
loki@math.ucsd.edu
kmesser@ucsd.edu