Biometrics , 1–25

Variance Estimation in Inverse Probability Weighted Cox Models

Di Shu*[1], Jessica G. Young[1], Sengwee Toh[1] and Rui Wang[1,2]

[1]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, USA

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

**email:* Di_Shu@harvardpilgrim.org

Summary: Inverse probability weighted Cox models can be used to estimate marginal hazard ratios under different treatments interventions in observational studies. To obtain variance estimates, the robust sandwich variance estimator is often recommended to account for the induced correlation among weighted observations. However, this estimator does not incorporate the uncertainty in estimating the weights and tends to overestimate the variance, leading to inefficient inference. Here we propose a new variance estimator that combines the estimation procedures for the hazard ratio and weights using stacked estimating equations, with additional adjustments for the sum of non-independent and identically distributed terms in a Cox partial likelihood score equation. We prove analytically that the robust sandwich variance estimator is conservative and establish the asymptotic equivalence between the proposed variance estimator and one obtained through linearization by Hajage et al., 2018. In addition, we extend our proposed variance estimator to accommodate clustered data. We compare the finite sample performance of the proposed method with alternative methods through simulation studies. We illustrate these different variance methods in an inverse probability weighted application to estimate the marginal hazard ratio for postoperative hospitalization under sleeve gastrectomy versus Roux-en-Y gastric bypass in a large medical claims and billing database. To facilitate implementation of the proposed method, we have developed an R package **ipwCoxCSV**.

Key words: Clustered data; Cox model; Inverse probability weighting; Marginal hazard ratio; Sandwich variance estimator.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Inverse probability weighting, a tool to address missing data or unequal selection probabilities, has been widely used in various fields such as causal inference (e.g., Rosenbaum, 1987; Lunceford and Davidian, 2004; Hernán and Robins, 2019) and survey sampling (e.g., Horvitz and Thompson, 1952; Pfeffermann, 1993; Höfler et al., 2005; Seaman and White, 2013; Miratrix et al., 2018). With time-to-event outcomes, the inverse probability weighted (IPW) Cox model is frequently used to estimate the marginal hazard ratio comparing hazard functions of counterfactual failure times under different hypothetical treatment interventions in observational studies (Hernán and Robins, 2019). When interest is in the comparison of binary point treatment interventions (e.g., "treat" versus "do not treat"), the weights are a function of the estimated propensity score; i.e., the probability of receiving treatment conditional on the measured baseline confounders (Rosenbaum and Rubin, 1983). The consistency of the resulting estimator depends on several assumptions, including the assumptions of exchangeability between treated and untreated individuals given the baseline measured covariates, correct model specification, and consistent estimation of the propensity score.

The focus of this work is on variance estimation for the treatment effect estimators from IPW Cox models. Previous authors have discussed an efficiency paradox such that estimators constructed using the estimated nuisance parameters are more efficient than those constructed using the true values of these nuisance parameters (e.g., Robins et al., 1992; van der Laan and Robins, 2003; Henmi and Eguchi, 2004). Henmi and Eguchi (2004) gave a sufficient condition for this paradox based on the orthogonality of the components of the projected estimating functions – the projections of the score function on to a given set of estimating functions – corresponding to the parameters of interest and the nuisance parameters. In IPW estimation of average treatment effects for non-survival outcomes, it was found that estimating parameters in a propensity score model leads to a smaller

asymptotic variance for the IPW estimator than using the true values (Lunceford and Davidian, 2004). Similarly, with survival outcomes, it has been noted that a robust sandwich variance estimator tends to be conservative in estimating the variance of an IPW estimator when ignoring the uncertainty in estimating the weights (e.g., Robins, 1997, 1999; Hernán et al., 2000). Given the convenient implementation of a robust sandwich variance estimator using off-the-shelf statistical software, this approach to variance estimation has become routine in practical applications of weighted analysis, including in IPW Cox estimation.

Austin (2016) confirmed in extensive simulations that the robust sandwich variance estimator tends to provide conservative estimates of the variance in the case of IPW estimators of Cox models. He suggested that, given the overestimation of the variance, which leads to wider confidence intervals and inefficient inference, bootstrap resampling (Efron and Tibshirani, 1993) should be used in place of the robust sandwich variance estimator. However, given the computational burden of the bootstrap method, an analytical formula for computing a consistent variance estimator is desirable. Analytical variance formulae for IPW estimators for non-survival outcomes have been proposed in various settings (Lunceford and Davidian, 2004; Williamson et al., 2014; Perez-Heydrich et al., 2014), using the standard M-estimation technique (Stefanski and Boos, 2002) based on stacked estimating equations of the treatment effect and propensity score weights. However, the Cox partial likelihood score equation is not a sum of independently and identically distributed (i.i.d.) terms, making it challenging to apply the standard M-estimation technique to obtain a sandwich type variance estimator for the hazard ratio. Mao et al. (2018) proposed to first poissonize the Cox model and then construct the stacked estimating equations, motivated by numerical findings that the poissonized likelihood gave nearly identical point estimates as the Cox model. This method involves penalized splines that require specification of the number and location of knots. Hajage et al. (2018) derived a closed-form variance formula using linearization (Deville,

1999). Their approach involved linearizing the Cox model and the propensity score weights to arrive at a variable whose dispersion can be used to approximate the variance.

In this paper, we take a different approach to derive a new analytical variance formula, by directly correcting the available robust sandwich variance estimator (Lin and Wei, 1989; Binder, 1992) that ignores the uncertainty in weight estimation. Specifically, we combine the estimating equations for the propensity score weights and the estimating equation used for the robust sandwich variance estimation. In the "meat" part of the sandwich variance estimator, we approximate and replace the original non-i.i.d. terms in the weighted partial likelihood score equation with the i.i.d terms proposed by Lin and Wei (1989) and Binder (1992). We establish two properties of the proposed variance estimator. First, we show that it is asymptotically equivalent to the existing linearization estimator (Hajage et al., 2018). Second, we show that it is more efficient than the existing robust sandwich variance estimator through a direct comparison of the two formulae.

We further propose a new variance estimator for clustered data settings. Clustered data occur frequently in practice. For example, each patient may experience recurrent post-surgery infections where times to multiple infections for the same patient are expected to be correlated. There is no available analytical variance formulae for the IPW Cox model to handle clustered survival data. To fill the gap, we extend the robust sandwich variance estimator proposed by Lee et al. (1992) to the IPW context, with uncertainty in weight estimation taken into account using stacked estimating equations.

The manuscript is organized as follows. In Section 2, we review IPW estimation of Cox models. In Section 3, we review four existing variance estimation methods, and propose a new estimator – the *corrected sandwich variance estimator* – for both independent and clustered data settings. We establish the relation between our corrected sandwich variance estimator and the linearization variance estimator and prove analytically that the robust

sandwich variance estimator is conservative. In Section 4, we conduct simulation studies to evaluate the finite sample performance of the proposed method. For illustration, in Section 5 we perform an IPW Cox analysis of a bariatric surgery dataset arising from the IBM® Health MarketScan® Research Databases. We apply various variance estimation methods for IPW estimator of the log hazard ratio for postoperative hospitalization under sleeve gastrectomy versus Roux-en-Y gastric bypass. We conclude the paper with a discussion in Section 6.

2. Estimation of Marginal Hazard Ratios Using Inverse Probability Weighting

Observed Data Structure: Consider an observational study in which the following are measured on each of $i = 1, \ldots, n$ individuals randomly sampled from a target population of interest (we initially assume individuals are i.i.d. and therefore suppress the i subscript here): Let $\boldsymbol{X}$ be a vector of measured baseline covariates, A a binary treatment indicator ($A = 1$ if treated and $A = 0$ otherwise), and $T = \min(T^*, C)$ where T^* is the event time, C is the censoring time. Further define $\delta = I(T^* \leq C)$, where $I(\cdot)$ is the indicator function. We assume a non-informative censoring mechanism that C is independent of (T^*, $\boldsymbol{X}$) conditional on A.

Parameter of Interest: We aim to estimate the log marginal hazard ratio θ of the model:

$$\lambda_a(t) = \lambda_0(t) \exp(\theta a), \tag{1}$$

where $\lambda_a(t)$ is the hazard function for T_a^*, the time to failure for a given individual in the study population that would have been observed had we set the treatment level $A = a$ for $a = 0$ or 1. We can equivalently interpret θ in (1) as the parameter to which an unweighted partial likelihood estimator of a correctly specified unconditional Cox model would converge when applied to data from a randomized controlled trial.

IPW estimator of θ: Inverse probability weighting effectively eliminates or reduces confounding bias such that the weighted data emulate data that would have been collected from

a randomized controlled trial. We consider the IPW estimator $\widehat{\theta}$ which solves the weighted partial likelihood score equation (Cox, 1975; Lin and Wei, 1989; Binder, 1992) for θ

$$\sum_{i=1}^{n} \widehat{w}_i \delta_i \left\{ A_i - \frac{\sum_{l:l\in\Re_i} \widehat{w}_l \exp(A_l\theta) A_l}{\sum_{l:l\in\Re_i} \widehat{w}_l \exp(A_l\theta)} \right\} = 0, \tag{2}$$

where $\Re_i = \{l : l = 1, \ldots, n, T_i \leq T_l, \delta_i = 1\}$ is the risk set for individual i who experiences an event at T_i and $\widehat{w}_i$ is an estimate of a weight w_i. Two types of weight are commonly used: the conventional inverse probability weight

$$w_i = w_{i,ipw} = \frac{A_i}{e_i} + \frac{1 - A_i}{1 - e_i} \tag{3}$$

and the stabilized weight

$$w_i = w_{i,stab} = P(A = 1)\frac{A_i}{e_i} + P(A = 0)\frac{1 - A_i}{1 - e_i}, \tag{4}$$

where $e_i = P(A_i = 1|\boldsymbol{X}_i)$ is the propensity score (Rosenbaum, 1987; Cole and Hernán, 2004, 2008). In an observational study, the propensity score e_i and treatment prevalence $P(A = 1)$ are unknown but may be estimated from the data. We consider the estimator $\widehat{\theta}$ under a logistic regression model for e_i and, when stabilized weights are used, nonparametric estimation of the marginal treatment prevalence by the proportion treated in the sample.

The consistency of $\widehat{\theta}$ for the true value of θ relies on correct specification of the propensity score model. It also requires several identifying assumptions including conditional exchangeability of treated and untreated individuals (A independent of T_a^* given $\boldsymbol{X}$), positivity (individuals with $A = 1$ or $A = 0$ are possibly observed within all levels of $\boldsymbol{X}$) and sufficiently well-defined counterfactual outcomes (Hernán and Robins, 2019).

3. Variance Estimation Methods for Marginal Hazard Ratios

In this section we describe five variance estimation methods for the IPW estimator $\widehat{\theta}$. In Section 3.1, we review four existing variance estimation methods. In Section 3.2, we propose the new *corrected sandwich variance estimator* and establish its relation with the

linearization estimator and the standard robust sandwich variance estimator. We also give an extension of our estimator that handles clustered data.

3.1 *Review of Four Existing Variance Estimation Methods*

3.1.1 *Naive Likelihood-Based Variance Estimator.* With the estimated weights in (2) treated as known constants, an application of the partial likelihood-based variance estimation (Andersen and Gill, 1982) leads to the naive likelihood-based variance estimator for $\widehat{\theta}$:

$$\widehat{var}_{\text{NL}}(\widehat{\theta}) = \left\{ -\sum_{i=1}^{n} \psi_i^{*\prime}(\widehat{\theta}) \right\}^{-1}, \tag{5}$$

where

$$\psi_i^*(\theta) = \widehat{w}_i \delta_i \left\{ A_i - \frac{\sum_{l:l\in\Re_i} \widehat{w}_l \exp(A_l\theta) A_l}{\sum_{l:l\in\Re_i} \widehat{w}_l \exp(A_l\theta)} \right\}.$$

In addition to ignoring the uncertainty in weight estimation, the naive likelihood-based variance estimator (5) incorrectly assumes independence among the weighted observations and thus is biased in general.

3.1.2 *Robust Sandwich Variance Estimator.* To help protect against model misspecification, Lin and Wei (1989) developed the robust sandwich variance estimator for partial likelihood estimates of Cox model parameters, and Binder (1992) extended their results to incorporate known constant weights.

The weighted robust sandwich variance estimator that replaces the true weights w_i with their estimates $\widehat{w}_i$, $i = 1, \ldots, n$ is given by

$$\widehat{var}_{\text{RS}}(\widehat{\theta}) = \left\{ -\sum_{i=1}^{n} \psi_i^{*\prime}(\widehat{\theta}) \right\}^{-1} \sum_{i=1}^{n} \eta_i^*(\widehat{\theta}) \eta_i^*(\widehat{\theta})^{\text{T}} \left[\left\{ -\sum_{i=1}^{n} \psi_i^{*\prime}(\widehat{\theta}) \right\}^{-1} \right]^{\text{T}}, \tag{6}$$

where

$$\eta_i^*(\widehat{\theta}) = \widehat{w}_i \delta_i \left\{ A_i - \frac{S_1(i)}{S_0(i)} \right\} - \widehat{w}_i A_i \exp(A_i\widehat{\theta}) \sum_{j=1}^{n} \frac{\delta_j \widehat{w}_j I(T_j \le T_i)}{S_0(j)} + \widehat{w}_i \exp(A_i\widehat{\theta}) \sum_{j=1}^{n} \frac{\delta_j \widehat{w}_j I(T_j \le T_i) S_1(j)}{S_0^2(j)},$$

$S_0(i) = \sum_{l:l\in\Re_i} \widehat{w}_l \exp(A_l\widehat{\theta})$, and $S_1(i) = \sum_{l:l\in\Re_i} \widehat{w}_l \exp(A_l\widehat{\theta}) A_l$.

Since the log marginal hazard ratio is a constant, both $\psi_i^{*\prime}(\widehat{\theta})$ and $\eta_i^*(\widehat{\theta})$ are scalars, and the robust sandwich variance estimator can be re-written as

$$\widehat{var}_{\text{RS}}(\widehat{\theta}) = \left\{ -\sum_{i=1}^{n} \psi_i^{*\prime}(\widehat{\theta}) \right\}^{-2} \sum_{i=1}^{n} \{\eta_i^*(\widehat{\theta})\}^2. \tag{7}$$

Because the robust sandwich variance estimator (6) or (7) treats the estimated weights as known constants, it does not take into account the uncertainty in weight estimation and is generally a biased estimator of the true variance of $\widehat{\theta}$.

3.1.3 *Bootstrap Variance Estimator.* The bootstrap method (Efron and Tibshirani, 1993) has been frequently used to obtain variance of estimators. In the current context, one resamples data at the individual level with replacement M times, for user-specified M (e.g., $M = 500$) to construct M bootstrap samples each containing the same dimensions as the original data. In each bootstrap sample $m = 1, \ldots, M$, the entire estimation algorithm is repeated, including estimation of the propensity score and corresponding weights, to obtain an estimate of the log hazard ratio (the true θ under the model 1), in that sample. Denote the estimate for sample m by $\widehat{\theta}_m$. The bootstrap variance estimator is then given by

$$\widehat{var}_{\text{BOOT}}(\widehat{\theta}) = \frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\theta}_m - \frac{1}{M} \sum_{m=1}^{M} \widehat{\theta}_m \right)^2. \tag{8}$$

Note that, because the propensity score and the weights are re-estimated in each bootstrap sample, the bootstrap variance estimator (8) incorporates the uncertainty in weight estimation. Austin (2016) found that the performance of the bootstrap variance estimator was superior to the commonly used robust sandwich variance estimator in his simulations.

3.1.4 *Linearization Variance Estimator.* Hajage et al. (2018) derived an analytical variance formula for the IPW estimator $\widehat{\theta}$ that is the solution to (2) using an influence function technique (Deville, 1999). Specifically, they showed that the variance can be approximated by the dispersion of a linearized variable divided by sample size. In their derivation, linearization

was conducted for both the Cox model and the propensity score weights to take into account the uncertainty in weight estimation.

Their proposed linearization variance estimator is

$$\widehat{var}_{\text{LIN}}(\widehat{\theta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\widehat{L}_i - \frac{1}{n} \sum_{i=1}^{n} \widehat{L}_i \right)^2, \tag{9}$$

where $\{\widehat{L}_i : i = 1, \ldots, n\}$ are the linearized terms. Specifically, define

$$\widehat{L}_{0i} = \delta_i \left\{ A_i - \frac{S_1(i)}{S_0(i)} \right\} - \exp(\widehat{\theta} A_i) \left[\sum_{j=1}^{n} \frac{\widehat{w}_j \delta_j I(T_j \le T_i)}{S_0(j)} \left\{ A_i - \frac{S_1(j)}{S_0(j)} \right\} \right],$$

$$\widehat{\boldsymbol{U}} = \frac{1}{n} \sum_{j=1}^{n} \widehat{e}_j (1 - \widehat{e}_j) \boldsymbol{X}_j \boldsymbol{X}_j^{\mathrm{T}}, \quad \text{and} \quad \widehat{V} = \frac{1}{n} \sum_{j=1}^{n} \widehat{w}_j \delta_j \frac{S_1(j)}{S_0(j)} \left\{ 1 - \frac{S_1(j)}{S_0(j)} \right\}.$$

For the conventional inverse probability weights (3), the linearized term $\widehat{L}_i$ in (9) is

$$\widehat{L}_{1i} = \widehat{V}^{-1} \{ \widehat{w}_i \widehat{L}_{0i} + \widehat{\boldsymbol{d}}_1^{\mathrm{T}} (A_i - \widehat{e}_i) \boldsymbol{X}_i \},$$

where

$$\widehat{\boldsymbol{d}}_1 = \widehat{\boldsymbol{U}}^{-1} \left[\frac{1}{n} \sum_{j=1}^{n} \left\{ -A_j \frac{1 - \widehat{e}_j}{\widehat{e}_j} + (1 - A_j) \frac{\widehat{e}_j}{1 - \widehat{e}_j} \right\} \widehat{L}_{0j} \boldsymbol{X}_j \right],$$

and $\widehat{e}_i$ is the estimated propensity score $i = 1, \ldots, n$. For the stabilized weights (4), the linearized term $\widehat{L}_i$ in (9) is given by $\widehat{L}_{2i} = \widehat{V}^{-1} \{ \widehat{w}_i \widehat{L}_{0i} + \widehat{d}_2 (A_i - \widehat{\rho}) + \widehat{\boldsymbol{d}}_3^{\mathrm{T}} (A_i - \widehat{e}_i) \boldsymbol{X}_i \}$, where

$$\widehat{d}_2 = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{A_j}{\widehat{e}_j} - \frac{1 - A_j}{1 - \widehat{e}_j} \right) \widehat{L}_{0j}$$

and

$$\widehat{\boldsymbol{d}}_3 = \widehat{\boldsymbol{U}}^{-1} \left[\frac{1}{n} \sum_{j=1}^{n} \left\{ -A_j \frac{\widehat{\rho}(1 - \widehat{e}_j)}{\widehat{e}_j} + (1 - A_j) \frac{(1 - \widehat{\rho}) \widehat{e}_j}{1 - \widehat{e}_j} \right\} \widehat{L}_{0j} \boldsymbol{X}_j \right].$$

3.2 *New Method: The Corrected Sandwich Variance Estimator*

3.2.1 *Theoretical Development.* We derive a new analytical variance estimator under either the conventional inverse probability weights (3) or stabilized weights (4). Our method extends the robust sandwich variance estimation to account for the uncertainty in estimating weights. We refer to the proposed estimator as the *corrected sandwich* variance estimator.

First, we develop the variance estimator with the conventional inverse probability weights (3). Let $\boldsymbol{\gamma}$ denote the vector of parameters in the propensity score model, which is specified as a logistic regression model. The corresponding system of estimating equations for $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ is given by

$$\sum_{i=1}^{n} \Phi_i(\theta, \boldsymbol{\gamma}) = \begin{cases} \sum_{i=1}^{n} \psi_i(\theta, \boldsymbol{\gamma}) = \sum_{i=1}^{n} w_i \delta_i \left\{ A_i - \dfrac{\sum_{l:l \in \Re_i} w_l \exp(A_l \theta) A_l}{\sum_{l:l \in \Re_i} w_l \exp(A_l \theta)} \right\} = 0 \\ \sum_{i=1}^{n} \pi_i(\boldsymbol{\gamma}) = \sum_{i=1}^{n} [A_i - 1/\{1 + \exp(-\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X}_i)\}] \boldsymbol{X}_i = \mathbf{0} \end{cases} \tag{10}$$

where w_i is individual i's conventional weight defined by (3) and estimated using the score function $\pi_i(\boldsymbol{\gamma})$ for logistic propensity score model (with 1 included in the vector of covariates). Solving (10) for $(\theta, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ gives $(\widehat{\theta}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\theta}$ is the estimated log hazard ratio and $\widehat{\boldsymbol{\gamma}}$ is the estimated propensity score model parameters.

A standard application of M-estimation (e.g., Stefanski and Boos, 2002) to (10) is complicated by the fact that the partial likelihood score equation is not a sum of i.i.d. terms. We propose to estimate the variance of $\widehat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$ by adapting the strategy of Lin and Wei (1989) and Binder (1992) to get around the non-i.i.d. problems.

In Web Appendix A, we prove that the variance of $\widehat{\boldsymbol{\beta}}$ can be consistently estimated by

$$\widehat{var}_{\mathrm{CS}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1} \boldsymbol{B}(\widehat{\boldsymbol{\beta}}) \left\{ \boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1} \right\}^{\mathrm{T}}, \tag{11}$$

where $\boldsymbol{A}(\widehat{\boldsymbol{\beta}}) = -\sum_{i=1}^{n} \Phi_i'(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{B}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \Omega_i(\widehat{\boldsymbol{\beta}}) \Omega_i(\widehat{\boldsymbol{\beta}})^{\mathrm{T}}$, with $\Omega_i(\widehat{\boldsymbol{\beta}}) = \left(\eta_i(\widehat{\theta}, \widehat{\boldsymbol{\gamma}}), \pi_i(\widehat{\boldsymbol{\gamma}})^{\mathrm{T}} \right)^{\mathrm{T}}$ and $\eta_i(\widehat{\theta}, \widehat{\boldsymbol{\gamma}})$ given by

$$\widehat{w}_i \delta_i \left\{ A_i - \frac{S_1(i)}{S_0(i)} \right\} - \widehat{w}_i A_i \exp(A_i \widehat{\theta}) \sum_{j=1}^{n} \frac{\delta_j \widehat{w}_j I(T_j \le T_i)}{S_0(j)} + \widehat{w}_i \exp(A_i \widehat{\theta}) \sum_{j=1}^{n} \frac{\delta_j \widehat{w}_j I(T_j \le T_i) S_1(j)}{S_0^2(j)}.$$

Then the element at the first row and the first column of matrix $\widehat{var}_{\mathrm{CS}}(\widehat{\boldsymbol{\beta}})$, denoted by $\widehat{var}_{\mathrm{CS}}(\widehat{\theta})$, is the proposed variance estimator for $\widehat{\theta}$.

Let ρ denote the treatment prevalence. Under the stabilized weights (4), we define a system

of estimating equations for $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^{\mathrm{T}}, \rho)^{\mathrm{T}}$:

$$\sum_{i=1}^{n} \Phi_i(\theta, \boldsymbol{\gamma}, \rho) = \begin{cases} \sum_{i=1}^{n} \psi_i(\theta, \boldsymbol{\gamma}, \rho) = \sum_{i=1}^{n} w_i \delta_i \left\{ A_i - \dfrac{\sum_{l:l \in \Re_i} w_l \exp(A_l \theta) A_l}{\sum_{l:l \in \Re_i} w_l \exp(A_l \theta)} \right\} = 0 \\ \sum_{i=1}^{n} \pi_i(\boldsymbol{\gamma}) = \sum_{i=1}^{n} [A_i - 1/\{1 + \exp(-\boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{X_i})\}] \boldsymbol{X}_i = \mathbf{0} \\ \sum_{i=1}^{n} \sigma_i(\rho) = \sum_{i=1}^{n} (A_i - \rho) = 0 \end{cases} \tag{12}$$

where w_i is given by (4), and $\psi_i(\theta, \boldsymbol{\gamma}, \rho)$, $\pi_i(\boldsymbol{\gamma})$ and $\sigma_i(\rho)$ are the partial likelihood score function for the weighted Cox model, the score function for the logistic propensity score model (with 1 included in the vector of covariates), and the estimating function for the treatment prevalence, respectively.

Solving (12) for $(\theta, \boldsymbol{\gamma}^{\mathrm{T}}, \rho)^{\mathrm{T}}$ gives $(\widehat{\theta}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}}, \widehat{\rho})^{\mathrm{T}}$, where $\widehat{\theta}$ is the estimated log hazard ratio, $\widehat{\boldsymbol{\gamma}}$ is the estimated propensity score model parameters, and $\widehat{\rho}$ is the estimated treatment prevalence. The variance estimator for $\widehat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}}, \widehat{\rho})^{\mathrm{T}}$ under estimating equations (12) can be derived in a similar way to the variance estimator (11) under estimating equations (10).

3.2.2 *Comparison with the Linearization and Robust Sandwich Estimators.* Both the proposed variance estimator $\widehat{var}_{\mathrm{CS}}(\widehat{\theta})$ and the linearization variance estimator $\widehat{var}_{\mathrm{LIN}}(\widehat{\theta})$ developed by Hajage et al. (2018) incorporate the uncertainty in the estimation of the propensity score weights. It is of interest to establish the intrinsic connections between these two analytical estimators, whose formulae look quite different at first glance.

Although derived from different approaches, the two estimators are asymptotically equivalent. This is justified by showing that $\widehat{var}_{\mathrm{CS}}(\widehat{\theta})$ can be re-written as the empirical second moment of the linearized variable divided by n. Below is a sketch of proof with the conventional inverse probability weights. The detailed proof with both types of weights is available in Web Appendix B.

We re-write $\boldsymbol{A}(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{B}(\widehat{\boldsymbol{\beta}})$ in block matrix form as

$$\boldsymbol{A}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} A_{11} & \boldsymbol{A}_{12} \\ \mathbf{0} & \boldsymbol{A}_{22} \end{bmatrix} \quad \text{and} \quad \boldsymbol{B}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} B_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^{\mathrm{T}} & \boldsymbol{B}_{22} \end{bmatrix}.$$

It can be shown that $\widehat{var}_{\text{CS}}(\widehat{\theta})$, the element at the first row and the first column of matrix $\widehat{var}_{\text{CS}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\boldsymbol{B}(\widehat{\boldsymbol{\beta}})\left\{\boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\right\}^{\text{T}}$, is given by

$$\widehat{var}_{\text{CS}}(\widehat{\theta}) = \frac{1}{A_{11}^2}B_{11} - \frac{2}{A_{11}^2}\boldsymbol{B}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\text{T}} + \frac{1}{A_{11}^2}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{B}_{22}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\text{T}}.$$

On the other hand, it can be shown that

$$\sum_{i=1}^{n}\widehat{L}_{1i}^2/n^2 = \frac{1}{A_{11}^2}\left(B_{11} + 2\widehat{\boldsymbol{d}}_1^{\text{T}}\boldsymbol{B}_{12}^{\text{T}} + \widehat{\boldsymbol{d}}_1^{\text{T}}\boldsymbol{B}_{22}\widehat{\boldsymbol{d}}_1\right).$$

By further showing $\widehat{\boldsymbol{d}}_1 = -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\text{T}}$, we obtain

$$\widehat{var}_{\text{CS}}(\widehat{\theta}) = \sum_{i=1}^{n}\widehat{L}_{1i}^2/n^2, \tag{13}$$

where $\widehat{L}_{1i}$ is the linearized term for $i = 1, \ldots, n$. By (9) and (13), $\widehat{var}_{\text{CS}}(\sqrt{n}\widehat{\theta})$ is the empirical second moment of the linearized variable, and $\widehat{var}_{\text{LIN}}(\sqrt{n}\widehat{\theta})$ is the sample variance of the linearized variable. Because variance is the same as the second moment for a mean-zero variable, $\widehat{var}_{\text{CS}}(\sqrt{n}\widehat{\theta})$ and $\widehat{var}_{\text{LIN}}(\sqrt{n}\widehat{\theta})$ are asymptotically equivalent.

While it is well-known that the standard robust sandwich variance estimator is conservative, the development of a correct variance formula allows explicit comparisons of the two formulae. In Web Appendix C, we derive the large sample difference matrix between the proposed and robust sandwich variance estimators, which is negative definite. In addition to providing an explicit proof that the robust sandwich variance estimator is conservative, examining the components of this difference matrix may provide insights into which settings result in large or negligible differences between the standard robust sandwich variance estimator and the proposed corrected estimator.

3.2.3 *Extension to Handle Clustered Data.* We again use the stacked estimating equations approach to develop a variance estimator for IPW Cox model with clustered data (e.g., recurrent events for each patient). In unweighted situations, Lee et al. (1992) proposed a robust sandwich variance estimator for Cox regression when the data consists of a large

number of independent small-size clusters of correlated failure time observations. We consider its extension to the IPW context and correct the corresponding robust sandwich variance estimator by further accounting for the estimating equations for the propensity score weights.

Suppose cluster i has K_i failure times for $i = 1, \ldots, n$ and $k = 1, \ldots, K_i$, where K_i is relatively small compared to n. For the kth failure time of cluster i, let $\boldsymbol{X}_{ik}$ be the baseline covariates, A_{ik} the treatment indicator, $T_{ik} = \min(T^*_{ik}, C_{ik})$ where T^*_{ik} is the event time and C_{ik} is the censoring time, and δ_{ik} the event indicator. Note that $(\boldsymbol{X}^{\mathrm{T}}_{ik}, A_{ik})^{\mathrm{T}}$ may contain cluster-level factors, which are k-invariant.

We now develop the variance estimator under the conventional inverse probability weights (3). Let $\boldsymbol{\gamma}$ denote the logistic propensity score model parameters. The corresponding system of the estimating equations for $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ is given by

$$\sum_{i=1}^{n}\sum_{k=1}^{K_i}\Phi_{i,k}(\theta,\boldsymbol{\gamma}) = \begin{cases} \sum_{i=1}^{n}\sum_{k=1}^{K_i}\psi_{i,k}(\theta,\boldsymbol{\gamma}) = 0 \\ \sum_{i=1}^{n}\sum_{k=1}^{K_i}\pi_{i,k}(\boldsymbol{\gamma}) = \boldsymbol{0} \end{cases} \tag{14}$$

where

$$\psi_{i,k}(\theta,\boldsymbol{\gamma}) = w_{ik}\delta_{ik}\left\{A_{ik} - \frac{\sum_{j=1}^{n}\sum_{l=1}^{K_j} I(T_{jl} \geq T_{ik}) w_{jl}\exp(A_{jl}\theta)A_{jl}}{\sum_{j=1}^{n}\sum_{l=1}^{K_j} I(T_{jl} \geq T_{ik}) w_{jl}\exp(A_{jl}\theta)}\right\}$$

is the partial likelihood score function for the weighted Cox model,

$$\pi_{i,k}(\boldsymbol{\gamma}) = \left[A_{ik} - 1/\{1+\exp(-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_{ik})\}\right]\boldsymbol{X}_{ik}$$

is the score function for logistic propensity score model (with 1 included in the vector of covariates), and w_{ik} is given by (3). Solving (14) for $(\theta, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ gives $(\widehat{\theta}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, denoted by $\widehat{\boldsymbol{\beta}}$.

Similarly to the development in Section 3.2.1, we derive the corrected sandwich variance estimator of $\widehat{\boldsymbol{\beta}}$, given by $\widehat{var}_{\mathrm{CS}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\boldsymbol{B}(\widehat{\boldsymbol{\beta}})\left\{\boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\right\}^{\mathrm{T}}$, where $\boldsymbol{A}(\widehat{\boldsymbol{\beta}}) = -\sum_{i=1}^{n}\sum_{k=1}^{K_i}\Phi'_{i,k}(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{B}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\Omega_i(\widehat{\boldsymbol{\beta}})\Omega_i(\widehat{\boldsymbol{\beta}})^{\mathrm{T}}$, with $\Omega_i(\widehat{\boldsymbol{\beta}}) = \left(\sum_{k=1}^{K_i}\eta_{i,k}(\widehat{\theta},\widehat{\boldsymbol{\gamma}}), \sum_{k=1}^{K_i}\pi_{i,k}(\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\right)^{\mathrm{T}}$ and

$$\eta_{i,k}(\widehat{\theta},\widehat{\boldsymbol{\gamma}}) = \widehat{w}_{ik}\delta_{ik}\left\{A_{ik} - \frac{S_1(i,k)}{S_0(i,k)}\right\} - \widehat{w}_{ik}A_{ik}\exp(A_{ik}\widehat{\theta})\sum_{j=1}^{n}\sum_{l=1}^{K_j}\frac{\delta_{jl}\widehat{w}_{jl}I(T_{jl}\leq T_{ik})}{S_0(j,l)}$$

$$+\widehat{w}_{ik}\exp(A_{ik}\widehat{\theta})\sum_{j=1}^{n}\sum_{l=1}^{K_j}\frac{\delta_{jl}\widehat{w}_{jl}I(T_{jl}\leq T_{ik})S_1(j,l)}{S_0^2(j,l)},$$

where $S_1(i,k)=\sum_{j=1}^{n}\sum_{l=1}^{K_j}I(T_{jl}\geq T_{ik})\widehat{w}_{jl}\exp(A_{jl}\widehat{\theta})A_{jl}$ and $S_0(i,k)=\sum_{j=1}^{n}\sum_{l=1}^{K_j}I(T_{jl}\geq T_{ik})\widehat{w}_{jl}\exp(A_{jl}\widehat{\theta})$. Then the element at the first row and the first column of matrix $\widehat{var}_{\text{CS}}(\widehat{\boldsymbol{\beta}})$, denoted by $\widehat{var}_{\text{CS}}(\widehat{\theta})$, is the proposed variance estimator for $\widehat{\theta}$. Similarly, the variance estimator under stabilized weights (4) can be obtained by further including the estimating equation for treatment prevalence, i.e., $\sum_{i=1}^{n}\sum_{k=1}^{K_i}(A_{ik}-\rho)=0$. Note $\eta_{i,k}(\widehat{\theta},\widehat{\boldsymbol{\gamma}})$ is the key to address non-i.i.d. issues. In unweighted case, $\eta_{i,k}(\widehat{\theta},\widehat{\boldsymbol{\gamma}})$ reduces to the result of Lee et al. (1992). In non-clustered case, $\eta_{i,k}(\widehat{\theta},\widehat{\boldsymbol{\gamma}})$ reduces to the result of Binder (1992).

4. Simulation Studies

We conducted simulation studies to compare the finite sample performance of the proposed corrected sandwich variance estimation method with alternative methods in two settings: without clustering and with clustering. In Setting 1, we compared the proposed estimator with the naive likelihood-based variance estimator, the robust sandwich variance estimator, the bootstrap variance estimator, and the linearization variance estimator. In Setting 2, we compared the proposed estimator with the (cluster version) robust sandwich variance estimator (Lee et al., 1992) and the cluster bootstrap variance estimator (Davison and Hinkley, 1997; Field and Welsh, 2007).

4.1 *Setting 1: without clustering*

4.1.1 *Data Generation and Simulation Scenarios.* To simulate data that exactly followed model (1), we adapted the simulation method of Young et al. (2008), which was initially designed for time-varying treatment settings, to our point-treatment setting. Specifically, for $i=1,\ldots,n$ individuals, we simulated the following (we assumed individuals were i.i.d. and therefore suppressed the i subscript):

Step 1: counterfactual event time under an intervention that sets $A = 0$, T_0^*, according to an exponential distribution with constant hazard rate $\lambda_0 = 0.01$.

Step 2: vector of covariates $\boldsymbol{X} = (X^{(1)}, X^{(2)}, X^{(3)})^{\intercal}$, where $X^{(1)} = 0.5(T_0^* + 0.2)/(T_0^* + 1) + 0.3Z$ and $X^{(2)} = 1/\log(1.3T_0^* + 3) - 0.3Z$ with Z following the standard normal distribution, and $X^{(3)}$ a binary variable with $P(X^{(3)} = 1|T_0^*) = 0.3 + 0.5/(T_0^* + 1)$.

Step 3: treatment indicator A generated by setting the probability of being treated to be $1/\{1 + \exp(\gamma_0 + X^{(1)} - X^{(2)} - X^{(3)})\}$, where the parameter γ_0 was chosen such that the treatment prevalence was about 10%, 20%, 30%, 40%, or 50%.

Step 4: event time using formula $T^* = T_0^* \exp(-\theta A)$, where θ was specified as $log(0.8)$ such that the true marginal hazard ratio was 0.8.

Step 5: censoring time C generated from an exponential distribution whose rate was chosen to yield a censoring rate about 20%, 40%, 60%, or 80%, to feature different degrees of censoring, and calculated $T = \min(T^*, C)$ and $\delta = I(T^* \leq C)$.

We considered sample sizes of 250 and 5000 and ran 1000 simulations for each parameter configuration. Five hundred bootstrap samples were used when implementing the bootstrap variance estimator.

4.1.2 *Results.* Similar to Austin (2016), to evaluate the accuracy of the proposed variance estimator in comparison to the other four methods, we examined the ratio of the average standard error (ASE) to the empirical standard error (ESE), where ASE was calculated as the average of the estimated standard errors for $\widehat{\theta}$ (obtained using each variance estimation method) across 1000 simulation runs, and ESE was calculated as the empirical estimate of the standard error (i.e., square root of sample variance of $\widehat{\theta}$ across 1000 simulation runs). ESE directly measures the uncertainty in estimation of the log marginal hazard ratio and reflects the true variability. With adequate sample size, a consistent variance estimator is expected to have the ratio of ASE to ESE close to 1.

Figures 1 and 2 depict the ratios of ASE to ESE for the five variance estimation methods under various combinations of censoring rates and treatment prevalence. As expected, the proposed variance estimator generally produced ASE to ESE ratios fairly close to 1 with $n = 5000$ (censoring rates ranging from 20% to 80%, treatment prevalence ranging from 10% to 50%), indicating that it estimated the variance with high accuracy when the sample size was adequate. With a small sample size $n = 250$, the ratios of ASE to ESE can be noticeably smaller than 1 especially when the treatment prevalence was far from 50% and the censoring rate was high, implying that the proposed variance estimator may underestimate the variance when the number of events is small within one or both treatment groups. The robust sandwich variance estimator tended to produce ratios greater than 1, suggesting a tendency to overestimate the truth. In some scenarios, the robust sandwich variance estimator produced ratios as high as 1.4, implying an 40% overestimation of variance. The naive likelihood-based variance method severely underestimated the variance under the conventional inverse probability weights. Under stabilized weights, it can underestimate or overestimate the variance. The linearization method showed almost the same performance as the proposed method, as seen from the overlapping lines in figures. With a large sample size of $n = 5000$, the bootstrap method performed well. With a small sample size of $n = 250$, the bootstrap method severely overestimated the variance under high censoring rate and low treatment prevalence, which is likely due to extreme estimates in some bootstrap samples.

[Figure 1 about here.]

[Figure 2 about here.]

We further examined the empirical coverage rates of the corresponding 95% confidence intervals obtained using the five variance estimation methods, where an empirical coverage rate was calculated as the percentage of 95% confidence intervals in 1000 simulation runs that covered the true log marginal hazard ratio. Results are summarized in Figures 3 and 4.

As in Austin (2016), we drew three horizontal lines (at 93.65%, 95% and 96.35%) to indicate a plausible range of coverage rates. Based on the normality approximation, a consistent variance estimator is expected to have empirical coverage rates that fluctuate around 95% and roughly within the interval of (93.65%, 96.35%) given that we used 1000 simulation runs.

In most scenarios, the proposed method produced empirical coverage rates close to 95% and within range, although it tended to result in undercoverage with high censoring rates and low treatment prevalence. As anticipated, the robust sandwich variance estimator tended to produce conservative confidence intervals with empirical coverage higher than 95% (and 96.35%), due to its overestimation of variance. In some scenarios, its empirical coverage rates could be nearly 100%. The naive likelihood-based variance method produced severe undercoverage under the conventional inverse probability weights due to its underestimation of variance. Using stabilized weights, its coverage rates behaved much better than under the conventional weights, but still could be outside of the range (93.65%, 96.35%). Results from the linearization method were almost the same as those from the proposed method, shown from the overlapping lines in figures. With a large sample size $n = 5000$, the bootstrap method produced reasonable empirical coverage rates that within the interval (93.65%, 96.35%). With a small sample size $n = 250$, the bootstrap method produced slight undercoverage under 20% censoring and overcoverage under 80% censoring.

[Figure 3 about here.]

[Figure 4 about here.]

Finally, we examined the average widths of the 95% confidence intervals (Figures S1 and S2, Web Appendix D). Under a large sample size of $n = 5000$, the three methods that account for uncertainty in weights: the bootstrap method, the linearization method, and the proposed method behaved similarly. Under a small sample size of $n = 250$, the bootstrap method produced the widest confidence intervals. This difference could be substantial, likely

resulting from extreme estimates obtained in some bootstrap samples when the number of events was small. Results for additional settings with sample sizes of $n = 100, 500, 1000$, and 2000 are included in Figures S3-S14 (Web Appendix D). Similar results were observed.

4.2 *Setting 2: with Clustering*

We compared the finite sample performance of the proposed corrected sandwich variance estimator with the (cluster version) robust sandwich variance estimator (Lee et al., 1992) and the cluster bootstrap variance estimator (Davison and Hinkley, 1997; Field and Welsh, 2007). When implementing the cluster bootstrap, we resampled with replacement from the n clusters and used all observations from each selected cluster to form the bootstrap samples.

We simulated n clusters of size K as follows. For the ith cluster where $i = 1, \ldots, n$, we first generated K counterfactual failure events $\{T_0^*(i,1), \ldots, T_0^*(i,K)\}$ from the Frank's family with unit exponential margins and Kendall's tau equals 0.7. Then for $k = 1, \ldots, K$, we specified covariates $\boldsymbol{X}_{ik} = (X_{ik}^{(1)}, X_{ik}^{(2)}, X_{ik}^{(3)})^{\mathrm{T}}$, where $X_{ik}^{(1)} = \frac{1}{K}\sum_{k=1}^{K}[0.5\{T_0^*(i,k) + 0.2\}/\{T_0^*(i,k)+1\}]$, $X_{ik}^{(2)} = \frac{1}{K}\sum_{k=1}^{K}[1/\log\{1.3T_0^*(i,k)+3\}]$, and $X_{ik}^{(3)} = 0.3+0.5/\{T_0^*(i,k)+1\}$. Here $X_{ik}^{(1)}$ and $X_{ik}^{(2)}$ were both made k-invariant to be cluster-level factors. The treatment A_{ik} was generated by setting the propensity score for cluster i and time k to $e_{i,k} = 1/\{1 + \exp(\gamma_0 + 2X_{ik}^{(1)} + X_{ik}^{(2)} + X_{ik}^{(3)})\}$, where γ_0 was chosen to achieve the desired treatment prevalence of approximately 10%, 20%, 30%, 40%, or 50%. Calculate $T_{ik}^* = T_0^*(i,k)\exp(-\theta A_{ik})$. Censoring times for each cluster were drawn independently from an exponential distribution whose rate was chosen to yield a censoring rate of about 20% or 60%. The true marginal hazard ratio was specified as 1.5. We considered 80 clusters with size $K = 3$ or $K = 6$, and ran 1000 simulations for each parameter configuration. Five hundred bootstrap samples were used when implementing the cluster bootstrap method.

Figure 5 reports the ratios of ASE to ESE for the three variance estimation methods under various combinations of censoring rate and treatment prevalence. The proposed variance

estimator generally produced ASE: ESE ratios close to 1 and outperformed the robust sandwich variance estimator and the cluster bootstrap estimator. The robust sandwich variance estimator tended to overestimate the truth, just like in settings without clustering; in some scenarios, the resulting estimates doubled the true variance.

We further reported the empirical coverage rates and the average 95% confidence interval widths (Figures S15-S16, Web Appendix E) and repeated simulations under $n = 200$ and $n = 800$ clusters (Figures S17-S22, Web Appendix E). The simulation results showed that when the number of clusters became larger ($n = 200$ and $n = 800$), both the proposed method and the cluster bootstrap performed well in terms of ratios of ASE to ESE and coverage rates. In general, the proposed method tended to produce narrower confidence intervals than the cluster bootstrap.

[Figure 5 about here.]

5. Application to A Bariatric Surgery Dataset

We applied the naive likelihood-based variance estimator, the robust sandwich variance estimator, the bootstrap variance estimator, the linearization variance estimator, and the proposed corrected sandwich variance estimator to analyze a real-world bariatric surgery dataset arising from the IBM® Health MarketScan® Research Databases, which contained de-identified patient-level healthcare claims information from employers, health plans, hospitals, and Medicare and Medicaid programs fully compliant with the Health Insurance Portability and Accountability Act (HIPAA). The dataset included 6690 patients aged 18 to 79 years who received sleeve gastrectomy (SG) or Roux-en-Y gastric bypass (RYGB) surgery between 1/1/2015 and 9/30/2015. The treatment variable was set to 1 if the patient received SG and 0 if the patient received RYGB. Among the 6690 patients, 4719 (70.5%) received SG and 1971 (29.5%) received RYGB. The outcome was time to the first all-cause

hospitalization during the first 30-day follow-up after patients were discharged from the index hospitalization. As a common feature of administrative databases in safety studies of rare outcomes, the censoring rate was high (97%). There were 210 total failure events.

We conducted IPW Cox regression to estimate the marginal hazard ratio. The propensity score model was specified as a logistic regression model linking the treatment variable to measured baseline covariates $\boldsymbol{X}$, including gender; age; the Charlson/Elixhauser combined comorbidity score; diagnosis of cancer, depression, diabetes, eating disorder, gastroesophageal reflux disease, hypertension, kidney disease, and non-alcoholic fatty liver disease; number of emergency department visits; number of dispensing of unique drug classes; and number of unique generic medications. The resultant propensity score weights achieved reasonable covariate balance across treatment groups, shown from absolute standardized differences in percent (Austin and Stuart, 2015) (Figure S23, Web Appendix F). The estimated marginal hazard ratios under the conventional and stabilized weights were both 0.659.

To obtain the standard error and 95% confidence interval, we used the five variance estimation methods. Table 1 summarizes the results. The proposed corrected sandwich variance estimator and the linearization estimator produced almost the same standard errors and 95% confidence intervals. The robust sandwich variance estimate was only slightly larger than the proposed and linearization variance estimates in this example. The likelihood-based variance method produced a remarkably smaller standard error than the other methods under the conventional inverse probability weights, consistent with the findings in simulation studies. All the variance methods examined produced 95% confidence intervals for the marginal hazard ratio that excluded 1, suggesting a statistically significant lower risk of post-surgery hospitalization at the nominal level of 5% comparing SG to RYGB.

[Table 1 about here.]

6. Discussion

We considered variance estimation for IPW Cox model and proposed the corrected sandwich variance estimator for both independent and clustered data settings. Our simulation studies demonstrated satisfactory performance of the proposed variance estimator and confirmed that the standard robust sandwich variance estimator which incorporates estimated weights is conservative. The performance of the linearization estimator and the proposed estimator was quite similar and tended to provide narrower confidence intervals than the bootstrap estimator. Although the robust sandwich variance estimator ignores the uncertainty in weight estimation, the impact of ignoring such uncertainty on the magnitude of the variance is expected to be negligible when sample size is large. Based on findings from our study and prior studies, the proposed variance estimator, the linearization variance estimator, and the bootstrap variance estimator are generally recommended for practical use. To facilitate the implementation of the proposed method, we developed an R package **ipwCoxCSV**, which is available from the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/package=ipwCoxCSV`.

The idea of correcting available robust sandwich variance estimators is generally applicable to weighted Cox models in causal inference and survey sampling. For example, results of Binder (1992) may be extended to multivariable-adjusted Cox models (targeting conditional hazard ratios) with sampling weights estimated from the data. As another example, in multisite studies, it will be useful to develop a variance estimator for IPW Cox model stratified on data-contributing sites (Shu et al., 2019). It is also possible to handle other weighting strategies such as the one targeting the average treatment effect among treated (ATT).

As the performance of asymptotic methods relies on adequate sample size, in small samples with rare treatment, the proposed estimator may underestimate the variance while the bootstrap estimator may overestimate or underestimate the variance. In this case, analysts

may calculate $\widehat{var}_{\text{CS}}(\widehat{\theta})$, $\widehat{var}_{\text{LIN}}(\widehat{\theta})$, $\widehat{var}_{\text{BOOT}}(\widehat{\theta})$, and $\widehat{var}_{\text{RS}}(\widehat{\theta})$ to see if they are similar. The underestimation of the robust sandwich variance for estimators from generalized linear models with small number of clusters has been extensively studied (see for examples Kauermann and Carroll, 2001; Mancl and DeRouen, 2001). Small sample correction formulae have been proposed. It is not yet apparent and would be useful to investigate, whether or how these corrections may be extended to the survival data settings.

The robust sandwich variance method and the bootstrap method can be applied under any type of propensity score model. The proposed and linearization methods assume a logistic propensity score model, which is widely used in practice. To be flexible within the logistic model form, analysts may include additional terms such as interaction terms between covariates or higher-order polynomial terms of certain covariates if the relationship might be non-linear, to help achieve covariate balance. In principle, the proposed method allows for any type of propensity score model, as long as it has a well-defined estimating equation to be included in the stacked estimating equations. When the estimating equation is intractable (e.g., a black-box machine learning algorithm), it is unclear how to develop an analytical variance estimator. It would be useful to investigate the statistical properties of resulting estimators.

Acknowledgements

The authors thank the Editor, Associate Editor, and two referees for providing thoughtful comments, which led to an improved version of the paper. The authors also thank Qoua Her at Harvard Pilgrim Health Care Institute for the help with the bariatric surgery dataset. Drs. Shu and Toh are partially supported by the National Institutes of Health (U01EB023683) and the Agency for Healthcare Research and Quality (R01HS026214). Drs. Shu and Wang are partially supported by R01 AI136947 from the National Institute of Allergy and Infectious

Diseases. Drs. Toh and Wang are also supported by Harvard Pilgrim Health Care Institute Robert H. Ebert Career Development Awards.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10,** 1100–1120.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* **34,** 3661–3679.

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine* **35,** 5642–5655.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79,** 139–147.

Cole, S. R. and Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine* **75,** 45–49.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* **168,** 656–664.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269–276.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25,** 193–204.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall/CRC.

Field, C. A and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B* **69,** 369–390.

Hajage, D., Chauvet, G., Belin, L., Lafourcade, A., Tubach, F., and De Rycke, Y. (2018). Closed-form variance estimator for weighted propensity score estimators with survival outcome. *Biometrical Journal* **60,** 1151–1163.

Henmi, M. and Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* **91,** 929–941.

Hernán, M. A., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology* **11,** 561–570.

Hernán, M. A. and Robins, J. M. (2019). *Causal Inference.* Boca Raton: Chapman & Hall/CRC, forthcoming.

Höfler, M., Pfister, H., Lieb, R., and Wittchen, H.-U. (2005). The use of weights to account for non-response and drop-out. *Social Psychiatry and Psychiatric Epidemiology* **40,** 291–299.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663–685.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96,** 1387–1396.

Lee, E. W., Wei, L. J., and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Klein, J. P. and Goel, P. K. eds. Survival Analysis: State of the Art. Dordrecht: Kluwer Academic Publishers, 237–247.

Lin, D. Y. and Wei, L.-J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84,** 1074–1078.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*

23, 2937–2960.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57,** 126–134.

Mao, H., Li, L., Yang, W., and Shen, Y. (2018). On the propensity score weighting analysis with survival outcome: estimands, estimation, and inference. *Statistics in Medicine* **37,** 3745–3763.

Miratrix, L. W., Sekhon, J. S., Theodoridis, A. G., and Campos, L. F. (2018). Worth weighting? How to think about and use weights in survey experiments. *Political Analysis* **26,** 275–291.

Perez-Heydrich, C., Hudgens, M. G., Halloran, M. E., Clemens, J. D., Ali, M., and Emch, M. E. (2014). Assessing effects of cholera vaccination in the presence of interference. *Biometrics* **70,** 734–744.

Pfeffermann, D. (2014). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique* **61,** 317–337.

Robins, J. M. (1997). Marginal structural models. In: 1997 Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association, 1–10.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In: Halloran E., Berry D., eds. Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer, 95–134.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48,** 479–495.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82,** 387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in

observational studies for causal effects. *Biometrika* **70,** 41–55.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* **22,** 278–295.

Shu, D., Yoshida, K., Fireman, B. H and Toh, S. (2019). Inverse probability weighted Cox model in multi-site studies without sharing individual-level data. *Statistical Methods in Medical Research* https://doi.org/10.1177/0962280219869742

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56,** 29–38.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer.

Williamson, E. J., Forbes, A., and White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* **33,** 721–737.

Young, J. G., Hernán, M. A., Picciotto, S., and Robins, J. M. (2008). Simulation from structural survival models under complex time-varying data structures. JSM Proceedings, Section on Statistics in Epidemiology, Denver, CO: American Statistical Association.

Supporting Information

Additional supporting information (Web Appendices referenced in Sections 3-5, code and example data) may be found online in the Supporting Information section at the end of the article. An R package **ipwCoxCSV** which implements the proposed method, is available from the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/package=ipwCoxCSV`.

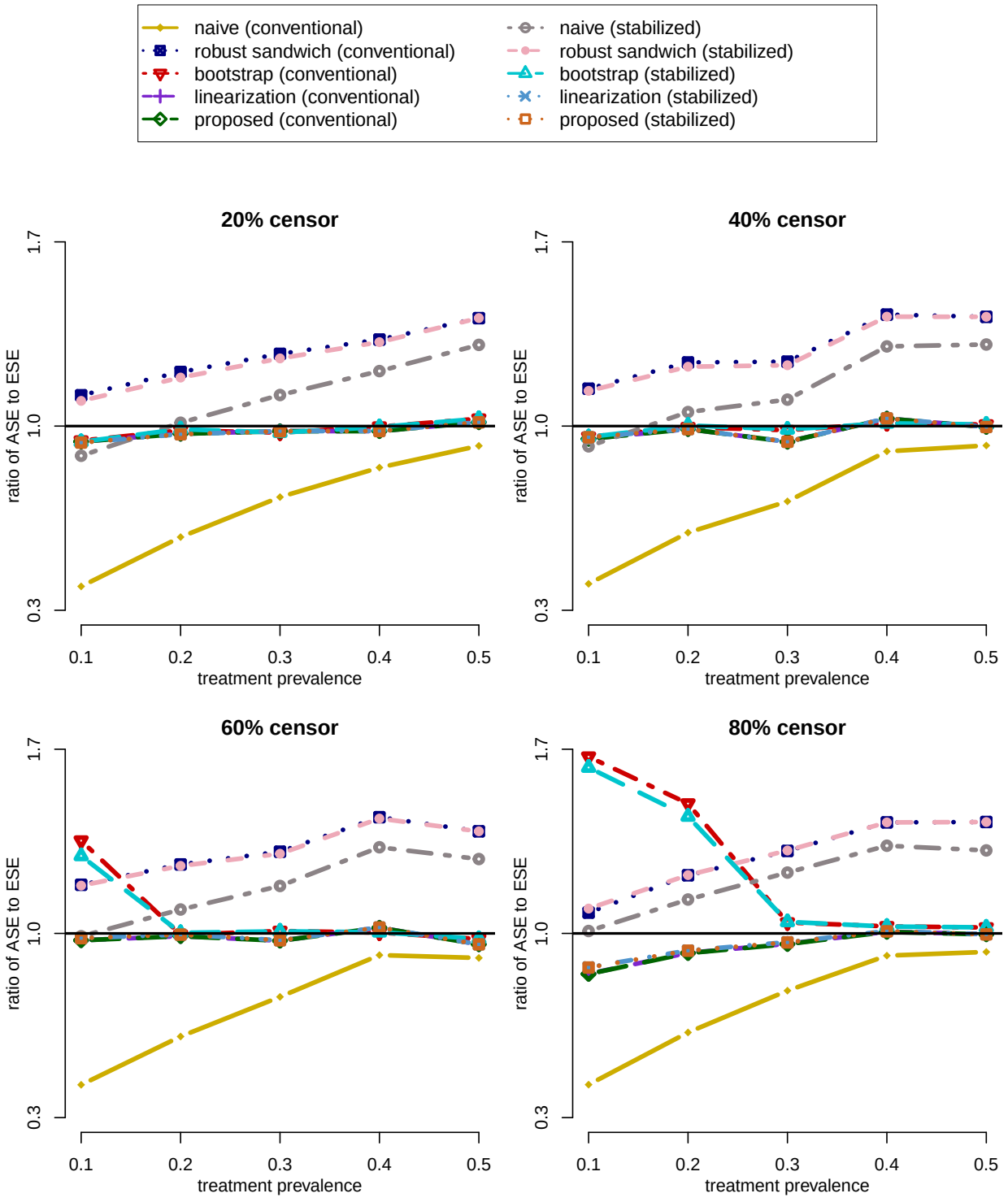

Figure 1: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 250$. Total number of failure events is about 200, 150, 100, or 50.

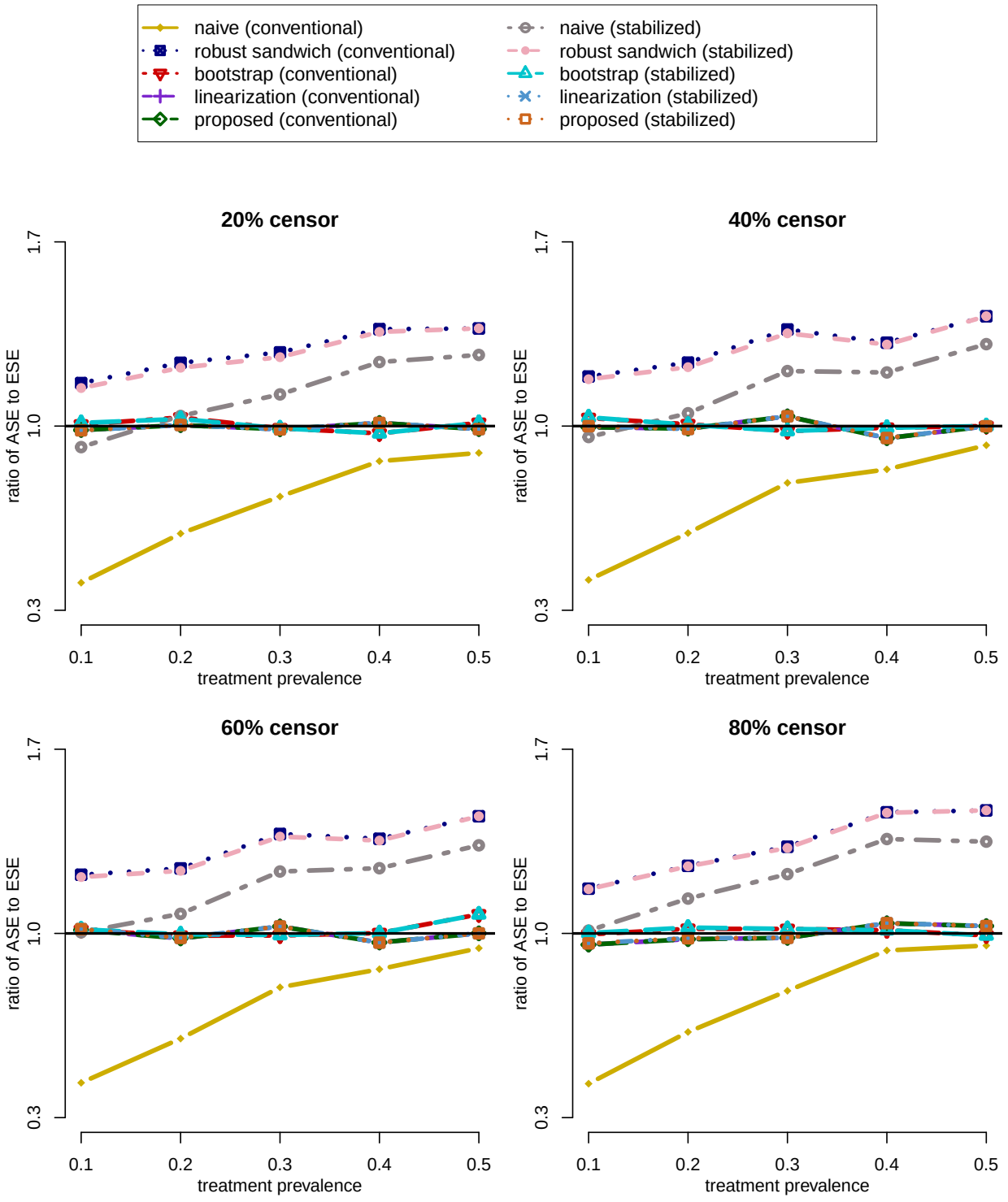

Figure 2: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 5000$. Total number of failure events is about 4000, 3000, 2000, or 1000.

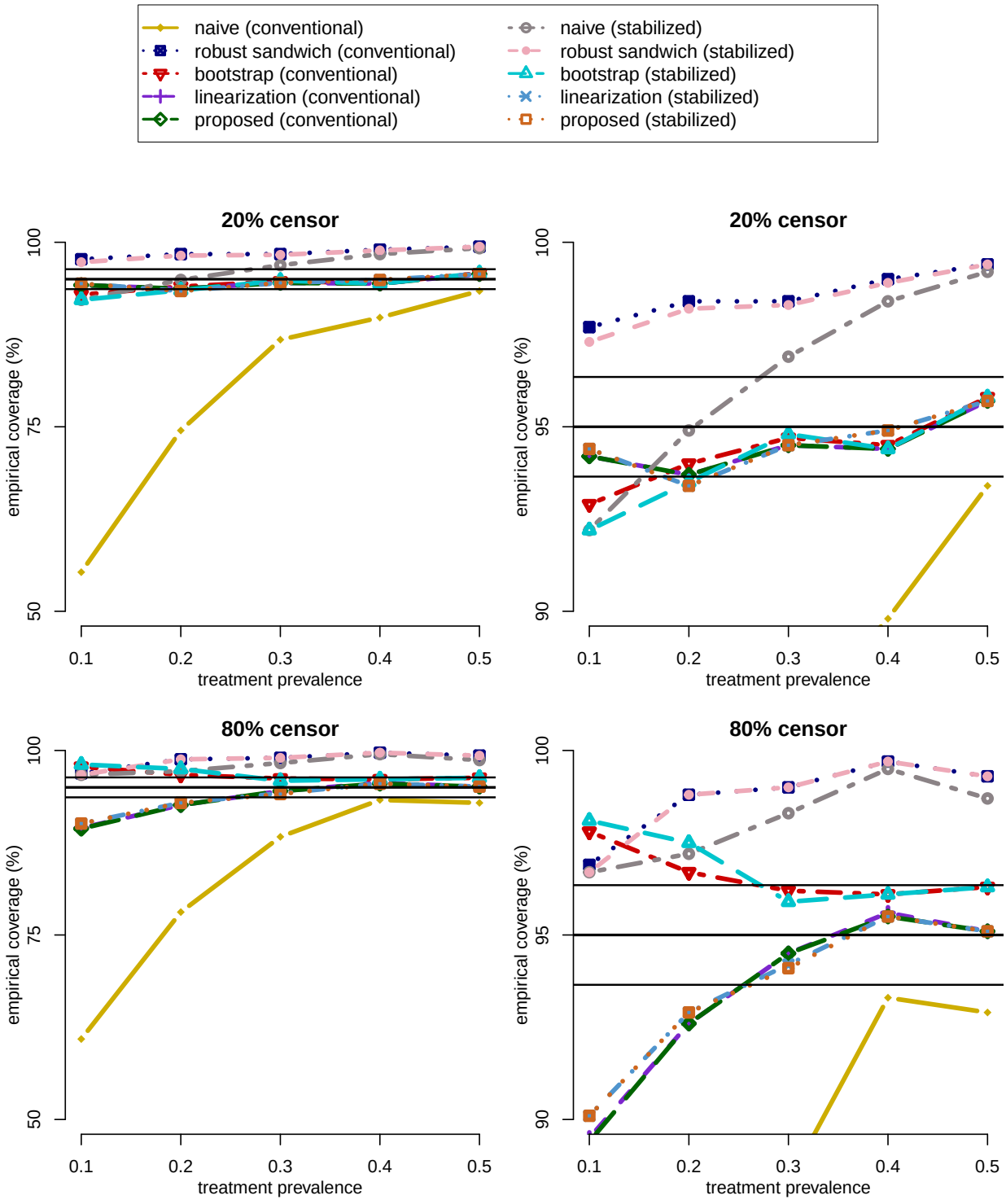

Figure 3: Empirical coverage rates in percent with $n = 250$. Total number of failure events is about 200 or 50. The right panel shows a zoom-in version of the left panel.

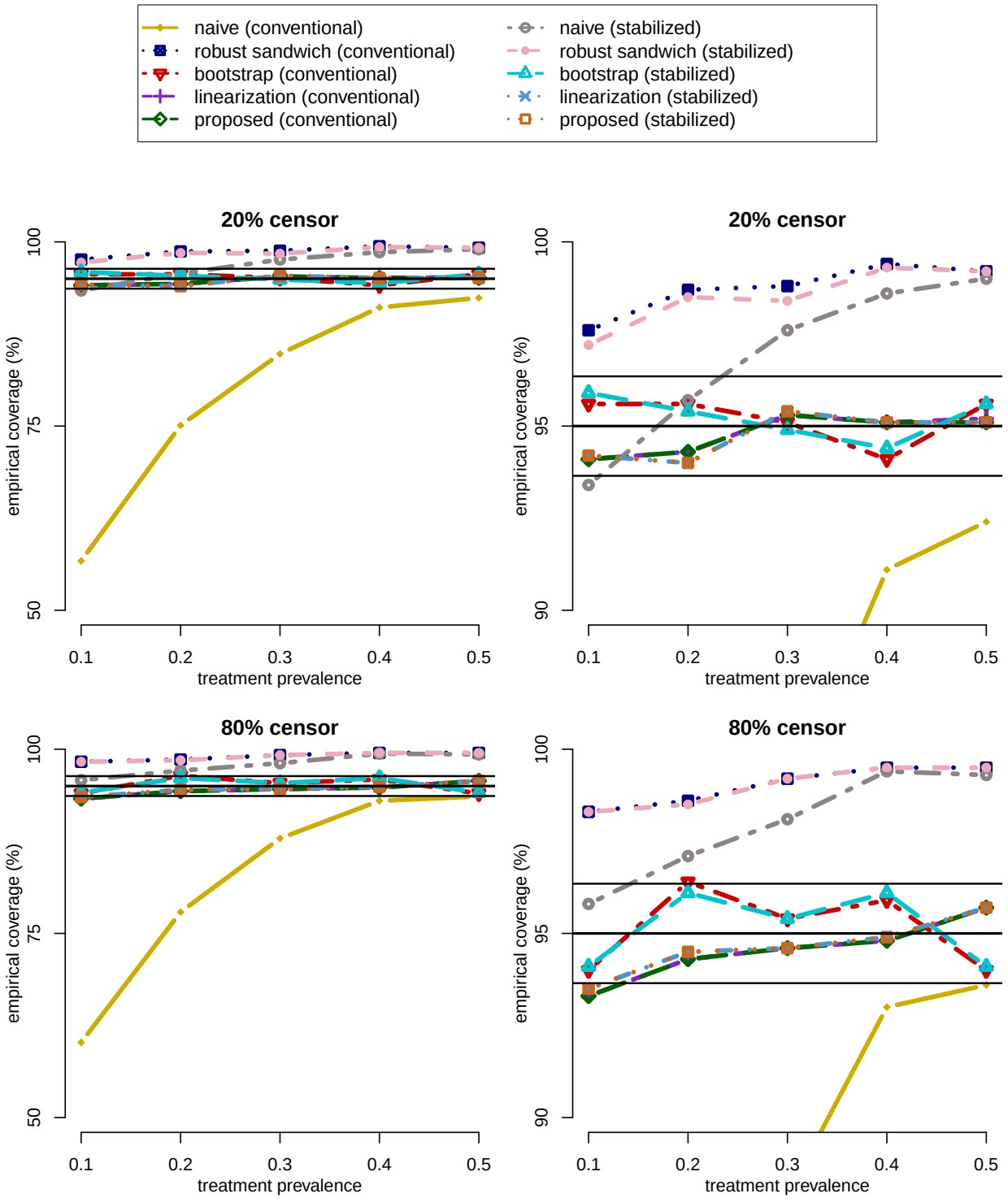

Figure 4: Empirical coverage rates in percent with $n = 5000$. Total number of failure events is about 4000 or 1000. The right panel shows a zoom-in version of the left panel.

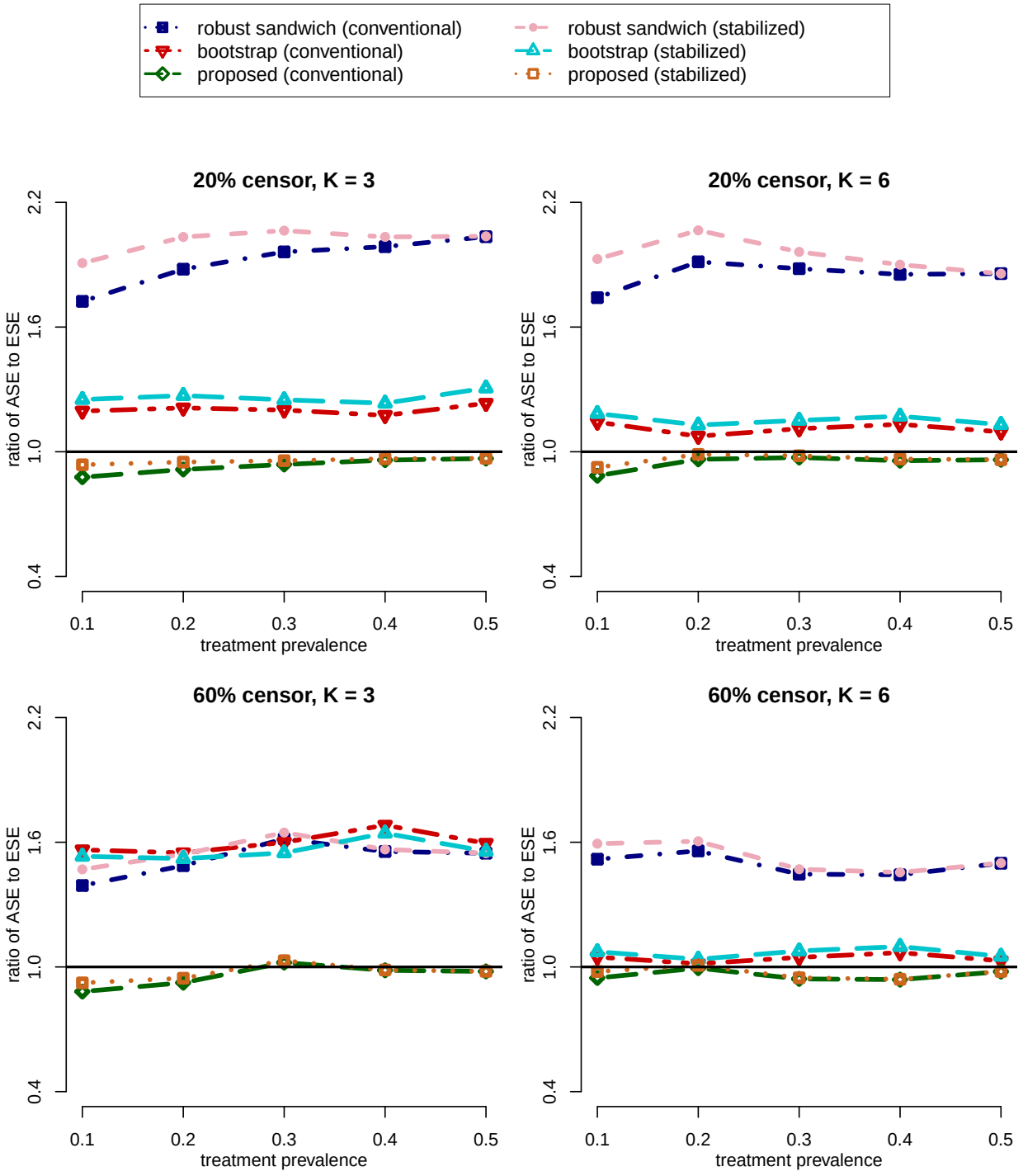

Figure 5: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 80$ clusters each of size K.

Table 1: Analysis results of bariatric surgery data using various variance estimation methods: estimated log marginal hazard ratio (log HR), estimated marginal hazard ratio (HR), standard error, and 95% confidence interval for marginal hazard ratio (95% CI of HR)

Weight	log HR	HR	Variance Method	Standard Error	95% CI of HR
Conventional	-0.417	0.659	Naive likelihood	0.0960	(0.5458, 0.7953)
			Robust sandwich	0.1440	(0.4968, 0.8738)
			Bootstrap (500 times)	0.1450	(0.4958, 0.8755)
			Linearization	0.1436	(0.4973, 0.8729)
			Corrected sandwich	0.1435	(0.4973, 0.8729)
Stabilized	-0.417	0.659	Naive likelihood	0.1426	(0.4982, 0.8714)
			Robust sandwich	0.1440	(0.4969, 0.8738)
			Bootstrap (500 times)	0.1450	(0.4959, 0.8755)
			Linearization	0.1435	(0.4973, 0.8729)
			Corrected sandwich	0.1435	(0.4973, 0.8729)

Supporting Information for

Variance Estimation in Inverse Probability Weighted Cox Models

by

Di Shu*[1], Jessica G. Young[1], Sengwee Toh[1] and Rui Wang[1,2]

[1]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, USA

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

*email: Di_Shu@harvardpilgrim.org

Web Appendix A: Development of the proposed variance estimator

Since $\sum_{i=1}^{n} \Phi_i(\widehat{\boldsymbol{\beta}}) = 0$, Taylor expansion gives

$$\sum_{i=1}^{n} \Phi_i'(\tilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sum_{i=1}^{n} \Phi_i(\boldsymbol{\beta}^*) = 0$$

where $\boldsymbol{\beta}^* = (\theta^*, \boldsymbol{\gamma}^{*\mathrm{T}})^{\mathrm{T}}$ is the limiting value for $\widehat{\boldsymbol{\beta}}$, and $\tilde{\boldsymbol{\beta}}$ is between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$. We obtain

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \left\{ -\frac{1}{n} \sum_{i=1}^{n} \Phi_i'(\tilde{\boldsymbol{\beta}}) \right\}^{-1} n^{-1/2} \sum_{i=1}^{n} \Phi_i(\boldsymbol{\beta}^*).$$

To estimate the asymptotic variance of $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, we need to estimate $\left\{ -\frac{1}{n} \sum_{i=1}^{n} \Phi_i'(\tilde{\boldsymbol{\beta}}) \right\}^{-1}$, and the asymptotic variance of $n^{-1/2} \sum_{i=1}^{n} \Phi_i(\boldsymbol{\beta}^*) = n^{-1/2} \left(\sum_{i=1}^{n} \psi_i(\theta^*, \boldsymbol{\gamma}^*), \sum_{i=1}^{n} \pi_i(\boldsymbol{\gamma}^*)^{\mathrm{T}} \right)^{\mathrm{T}}$. Note that $\left\{ -\frac{1}{n} \sum_{i=1}^{n} \Phi_i'(\tilde{\boldsymbol{\beta}}) \right\}^{-1}$ can be estimated by substituting $\widehat{\boldsymbol{\beta}}$ into the formula. Let

$$\boldsymbol{A}(\widehat{\boldsymbol{\beta}}) = -\sum_{i=1}^{n} \Phi_i'(\widehat{\boldsymbol{\beta}}).$$

Asymptotic variance of $n^{-1/2} \sum_{i=1}^{n} \Phi_i(\boldsymbol{\beta}^*) = n^{-1/2} \left(\sum_{i=1}^{n} \psi_i(\theta^*, \boldsymbol{\gamma}^*), \sum_{i=1}^{n} \pi_i(\boldsymbol{\gamma}^*)^{\mathrm{T}} \right)^{\mathrm{T}}$ is non-trivial since the partial likelihood score equation is not a sum of i.i.d. terms. By adapting the strategy of Lin and

Wei (1989) and Binder (1992) to get around the non-i.i.d. problems, $\psi_i(\theta^*, \boldsymbol{\gamma}^*)$ can be replaced by i.i.d. terms $\eta_i(\theta^*, \boldsymbol{\gamma}^*)$ where $n^{-1/2}\sum_{i=1}^n \eta_i(\theta^*, \boldsymbol{\gamma}^*)$ is asymptotically equivalent to $n^{-1/2}\sum_{i=1}^n \psi_i(\theta^*, \boldsymbol{\gamma}^*)$. Define $\Omega_i(\boldsymbol{\beta}^*) = (\eta_i(\theta^*, \boldsymbol{\gamma}^*), \pi_i(\boldsymbol{\gamma}^*)^{\mathrm{T}})^{\mathrm{T}}$. The asymptotic variance of $\sum_{i=1}^n \Omega_i(\boldsymbol{\beta}^*)$ can be estimated by

$$\boldsymbol{B}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \Omega_i(\widehat{\boldsymbol{\beta}})\Omega_i(\widehat{\boldsymbol{\beta}})^{\mathrm{T}},$$

where $\Omega_i(\widehat{\boldsymbol{\beta}}) = \left(\eta_i(\widehat{\theta}, \widehat{\boldsymbol{\gamma}}), \pi_i(\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\right)^{\mathrm{T}}$ and

$$\eta_i(\widehat{\theta}, \widehat{\boldsymbol{\gamma}}) = \widehat{w}_i\delta_i\left\{A_i - \frac{S_1(i)}{S_0(i)}\right\} - \widehat{w}_i A_i \exp(A_l\widehat{\theta})\sum_{j=1}^{n}\frac{\delta_j\widehat{w}_j I(T_j \le T_i)}{S_0(j)} + \widehat{w}_i \exp(A_l\widehat{\theta})\sum_{j=1}^{n}\frac{\delta_j\widehat{w}_j I(T_j \le T_i)S_1(j)}{S_0^2(j)}.$$

Therefore, we propose to estimate the variance of $\widehat{\boldsymbol{\beta}}$ as $\widehat{var}_{\mathrm{CS}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\boldsymbol{B}(\widehat{\boldsymbol{\beta}})\left\{\boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\right\}^{\mathrm{T}}$.

Web Appendix B: Relation between the proposed variance and the linearization variance

Recall the linearization variance $\widehat{var}_{\mathrm{LIN}}(\widehat{\theta})$ is the sample variance of the linearized terms divided by n. In this proof we show that for both the conventional inverse probability weights and stabilized weights, the proposed variance estimator $\widehat{var}_{\mathrm{CS}}(\widehat{\theta})$ can be re-written as the sample second moment of the linearization variable divided by n. By the mean-zero property of the linearized terms, the linearization and the proposed variance estimators are asymptotically equivalent. Below we give the detailed proof.

Conventional inverse probability weights

We re-write $\boldsymbol{A}(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{B}(\widehat{\boldsymbol{\beta}})$ in block matrix form as

$$\boldsymbol{A}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} A_{11} & \boldsymbol{A}_{12} \\ \mathbf{0} & \boldsymbol{A}_{22} \end{bmatrix} \quad \text{and} \quad \boldsymbol{B}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} B_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{12}^{\mathrm{T}} & \boldsymbol{B}_{22} \end{bmatrix}.$$

The inverse of $\boldsymbol{A}(\widehat{\boldsymbol{\beta}})$ is

$$\begin{bmatrix} A_{11} & \boldsymbol{A}_{12} \\ \mathbf{0} & \boldsymbol{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} C_{11} & \boldsymbol{C}_{12} \\ \mathbf{0} & \boldsymbol{C}_{22} \end{bmatrix}$$

where $C_{11} = \dfrac{1}{A_{11}}$, $\boldsymbol{C}_{22} = \boldsymbol{A}_{22}^{-1}$ and $\boldsymbol{C}_{12} = -\dfrac{1}{A_{11}}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}$.

Some matrix calculations show that $\widehat{var}_{\rm CS}(\widehat{\theta})$, the element at the first row and the first column of matrix $\widehat{var}_{\rm CS}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\boldsymbol{B}(\widehat{\boldsymbol{\beta}})\left\{\boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\right\}^{\rm T}$, is given by

$$\begin{aligned}
& \widehat{var}_{\rm CS}(\widehat{\theta}) \\
= \;& C_{11}B_{11}C_{11} + 2C_{11}\boldsymbol{B}_{12}\boldsymbol{C}_{12}^{\rm T} + \boldsymbol{C}_{12}\boldsymbol{B}_{22}\boldsymbol{C}_{12}^{\rm T} \\
= \;& \frac{1}{A_{11}^2}B_{11} - \frac{2}{A_{11}^2}\boldsymbol{B}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\rm T} + \frac{1}{A_{11}^2}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{B}_{22}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\rm T}. \qquad (1)
\end{aligned}$$

Our task is to show that the proposed variance estimator $\widehat{var}_{\rm CS}(\widehat{\theta})$ can be re-written as $\sum_{i=1}^n \widehat{L}_{1i}^2/n^2$, where $\widehat{L}_{1i}^2$ is the linearized term for $i = 1, \ldots, n$. Observing that $A_{11} = n\widehat{V}$, $B_{11} = \sum_{i=1}^n \widehat{w}_i^2\widehat{L}_{0i}^2$, $\boldsymbol{B}_{12}^{\rm T} = \sum_{i=1}^n \widehat{w}_i\widehat{L}_{0i}(A_i - \widehat{e}_i)\boldsymbol{X}_i$, and $\boldsymbol{B}_{22} = \sum_{i=1}^n (A_i - \widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\rm T}$, we obtain

$$\begin{aligned}
\sum_{i=1}^n \widehat{L}_{1i}^2/n^2 &= \sum_{i=1}^n (n\widehat{V})^{-2}\left\{\widehat{w}_i^2\widehat{L}_{0i}^2 + 2\widehat{w}_i\widehat{L}_{0i}\widehat{\boldsymbol{d}}_1^{\rm T}(A_i - \widehat{e}_i)\boldsymbol{X}_i + \widehat{\boldsymbol{d}}_1^{\rm T}(A_i - \widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\rm T}\widehat{\boldsymbol{d}}_1\right\} \\
&= \frac{1}{A_{11}^2}\left\{B_{11} + 2\sum_{i=1}^n \widehat{w}_i\widehat{L}_{0i}\widehat{\boldsymbol{d}}_1^{\rm T}(A_i - \widehat{e}_i)\boldsymbol{X}_i + \widehat{\boldsymbol{d}}_1^{\rm T}\boldsymbol{B}_{22}\widehat{\boldsymbol{d}}_1\right\} \\
&= \frac{1}{A_{11}^2}\left\{B_{11} + 2\widehat{\boldsymbol{d}}_1^{\rm T}\boldsymbol{B}_{12}^{\rm T} + \widehat{\boldsymbol{d}}_1^{\rm T}\boldsymbol{B}_{22}\widehat{\boldsymbol{d}}_1\right\}. \qquad (2)
\end{aligned}$$

To show the equivalence of (1) and (2), it suffices to show

$$\widehat{\boldsymbol{d}}_1 = -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\rm T}.$$

Observing $\boldsymbol{A}_{22} = \widehat{\boldsymbol{U}}$, we only need to show

$$\frac{1}{n}\sum_{j=1}^n \left[\left\{-A_j\frac{1-\widehat{e}_j}{\widehat{e}_j} + (1-A_j)\frac{\widehat{e}_j}{1-\widehat{e}_j}\right\}\widehat{L}_{0i}\boldsymbol{X}_j\right] = -\boldsymbol{A}_{12}^{\rm T} \qquad (3)$$

by the definition of $\widehat{\boldsymbol{d}}_1$.

For ease of exposition, we denote $g_j = -A_j\frac{1-\widehat{e}_j}{\widehat{e}_j} + (1-A_j)\frac{\widehat{e}_j}{1-\widehat{e}_j}$. The left-hand-side of (3) equals

$$\begin{aligned}
& \frac{1}{n}\sum_{j=1}^n (g_j\boldsymbol{X}_j\widehat{I}_{0j}) \\
= \;& \frac{1}{n}\sum_{j=1}^n g_j\boldsymbol{X}_j\left[\delta_j\left\{A_j - \frac{S_1(j)}{S_0(j)}\right\}\right. \\
& \left. -A_j\exp(\widehat{\theta}A_j)\sum_{l=1}^n \frac{\widehat{w}_l\delta_l I(T_l \le T_j)}{S_0(l)} + \exp(\widehat{\theta}A_j)\sum_{l=1}^n \frac{\widehat{w}_l\delta_l I(T_l \le T_j)S_1(l)}{S_0^2(l)}\right]. \qquad (4)
\end{aligned}$$

Note that

$$-\boldsymbol{A}_{12}^{\mathrm{T}} = \sum_{j=1}^{n} \left(\delta_j \left\{ A_j - \frac{S_1(j)}{S_0(j)} \right\} \cdot \frac{\partial \widehat{w}_i}{\partial \widehat{\gamma}} \right.$$
$$\left. -\widehat{w}_j \delta_j \left[\frac{\sum_{l=1}^{n} I(T_l \geq T_j) \exp(A_l \widehat{\theta}) A_l \frac{\partial \widehat{w}_l}{\partial \widehat{\gamma}}}{\sum_{l=1}^{n} \widehat{w}_l I(T_l \geq T_j) \exp(A_l \widehat{\theta})} - \frac{Re}{\left\{ \sum_{l=1}^{n} \widehat{w}_l I(T_l \geq T_j) \exp(A_l \widehat{\theta}) \right\}^2} \right] \right) \tag{5}$$

where

$$Re = \left\{ \sum_{l=1}^{n} I(T_l \geq T_j) \exp(A_l \widehat{\theta}) \frac{\partial \widehat{w}_l}{\partial \widehat{\gamma}} \right\} \left\{ \sum_{l=1}^{n} \widehat{w}_l I(T_l \geq T_j) \exp(A_l \widehat{\theta}) A_l \right\}.$$

We observe that $\dfrac{\partial \widehat{w}_l}{\partial \widehat{\gamma}} = g_l \boldsymbol{X}_l$, and obtain

$$\begin{aligned}
& \sum_{j=1}^{n} g_j \boldsymbol{X}_j A_j \exp(\widehat{\theta} A_j) \sum_{l=1}^{n} \frac{\widehat{w}_l \delta_l I(T_l \leq T_j)}{S_0(l)} \\
= & \sum_{k=1}^{n} \frac{\widehat{w}_k \delta_k}{S_0(k)} \sum_{l=1}^{n} g_l \boldsymbol{X}_l A_l \exp(\widehat{\theta} A_l) I(T_l \geq T_k) \\
= & \sum_{j=1}^{n} \frac{\widehat{w}_j \delta_j}{S_0(j)} \sum_{l=1}^{n} g_l \boldsymbol{X}_l A_l \exp(\widehat{\theta} A_l) I(T_l \geq T_j) \\
= & \sum_{j=1}^{n} \widehat{w}_j \delta_j \frac{\sum_{l=1}^{n} I(T_l \geq T_j) \exp(A_l \widehat{\theta}) A_l g_l \boldsymbol{X}_l}{\sum_{l=1}^{n} \widehat{w}_l I(T_l \geq T_j) \exp(A_l \widehat{\theta})}
\end{aligned} \tag{6}$$

and

$$\begin{aligned}
& \sum_{j=1}^{n} g_j \boldsymbol{X}_j \exp(\widehat{\theta} A_j) \sum_{l=1}^{n} \frac{\widehat{w}_l \delta_l I(T_l \leq T_j) S_1(l)}{S_0^2(l)} \\
= & \sum_{k=1}^{n} \frac{\widehat{w}_k \delta_k S_1(k)}{S_0^2(k)} \sum_{l=1}^{n} g_l \boldsymbol{X}_l \exp(\widehat{\theta} A_l) I(T_l \geq T_k) \\
= & \sum_{j=1}^{n} \frac{\widehat{w}_j \delta_j S_1(j)}{S_0^2(j)} \sum_{l=1}^{n} g_l \boldsymbol{X}_l \exp(\widehat{\theta} A_l) I(T_l \geq T_j) \\
= & \sum_{j=1}^{n} \widehat{w}_j \delta_j \frac{\left\{ \sum_{l=1}^{n} I(T_l \geq T_j) \exp(A_l \widehat{\theta}) g_l \boldsymbol{X}_l \right\} \left\{ \sum_{l=1}^{n} \widehat{w}_l I(T_l \geq T_j) \exp(A_l \widehat{\theta}) A_l \right\}}{\left\{ \sum_{l=1}^{n} \widehat{w}_l I(T_l \geq T_j) \exp(A_l \widehat{\theta}) \right\}^2}.
\end{aligned} \tag{7}$$

Combining (6) and (7) shows that (4) equals (5), and hence (1) equals (2).

Stabilized weights

We re-write $\boldsymbol{A}(\widehat{\boldsymbol{\beta}})$ and $\boldsymbol{B}(\widehat{\boldsymbol{\beta}})$ in block matrix form as

$$\boldsymbol{A}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} A_{11} & \boldsymbol{A}_{12} & A_{13} \\ \mathbf{0} & \boldsymbol{A}_{22} & \mathbf{0} \\ 0 & \mathbf{0} & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} B_{11} & \boldsymbol{B}_{12} & B_{13} \\ \boldsymbol{B}_{12}^{\mathrm{T}} & \boldsymbol{B}_{22} & \boldsymbol{B}_{23} \\ B_{13}^{\mathrm{T}} & \boldsymbol{B}_{23}^{\mathrm{T}} & B_{33} \end{bmatrix}.$$

The inverse of $\boldsymbol{A}(\widehat{\boldsymbol{\beta}})$ is

$$\begin{bmatrix} A_{11} & \boldsymbol{A}_{12} & A_{13} \\ \mathbf{0} & \boldsymbol{A}_{22} & \mathbf{0} \\ 0 & \mathbf{0} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} C_{11} & \boldsymbol{C}_{12} & C_{13} \\ \mathbf{0} & \boldsymbol{C}_{22} & \mathbf{0} \\ 0 & \mathbf{0} & 1 \end{bmatrix}$$

where $C_{11} = \dfrac{1}{A_{11}}$, $\boldsymbol{C}_{22} = \boldsymbol{A}_{22}^{-1}$, $\boldsymbol{C}_{12} = -\dfrac{1}{A_{11}}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}$, and $C_{13} = -\dfrac{1}{A_{11}}A_{13}$.

Some matrix calculations show that $\widehat{var}_{\mathrm{CS}}(\widehat{\theta})$, the element at the first row and the first column of matrix $\widehat{var}_{\mathrm{CS}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\boldsymbol{B}(\widehat{\boldsymbol{\beta}})\left\{\boldsymbol{A}(\widehat{\boldsymbol{\beta}})^{-1}\right\}^{\mathrm{T}}$, is given by

$$\begin{aligned} & \widehat{var}_{\mathrm{CS}}(\widehat{\theta}) \\ = \;& C_{11}B_{11}C_{11} + 2C_{11}\boldsymbol{B}_{12}\boldsymbol{C}_{12}^{\mathrm{T}} + 2C_{11}B_{13}C_{13}^{\mathrm{T}} + 2\boldsymbol{C}_{12}\boldsymbol{B}_{23}C_{13}^{\mathrm{T}} + \boldsymbol{C}_{12}\boldsymbol{B}_{22}\boldsymbol{C}_{12}^{\mathrm{T}} + C_{13}B_{33}C_{13}^{\mathrm{T}} \\ = \;& \frac{1}{A_{11}^2}B_{11} - \frac{2}{A_{11}^2}\boldsymbol{B}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\mathrm{T}} - \frac{2}{A_{11}^2}B_{13}A_{13}^{\mathrm{T}} \\ & + \frac{2}{A_{11}^2}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{B}_{23}A_{13}^{\mathrm{T}} + \frac{1}{A_{11}^2}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{B}_{22}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\mathrm{T}} + \frac{1}{A_{11}^2}A_{13}B_{33}A_{13}^{\mathrm{T}}. \end{aligned} \tag{8}$$

Our task is to show that the proposed variance estimator $\widehat{var}_{\mathrm{CS}}(\widehat{\theta})$ can be re-written as $\sum_{i=1}^n \widehat{L}_{2i}^2/n^2$, where $\widehat{L}_{2i}^2$ is the linearized term for $i = 1, \ldots, n$. Observing that $A_{11} = n\widehat{V}$, $B_{11} = \sum_{i=1}^n \widehat{w}_i^2\widehat{L}_{0i}^2$, $\boldsymbol{B}_{12}^{\mathrm{T}} = \sum_{i=1}^n \widehat{w}_i\widehat{L}_{0i}(A_i - \widehat{e}_i)\boldsymbol{X}_i$, $\boldsymbol{B}_{22} = \sum_{i=1}^n (A_i - \widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}$, $B_{13} = \sum_{i=1}^n \widehat{w}_i\widehat{L}_{0i}(A_i - \widehat{\rho})$, $\boldsymbol{B}_{23}^{\mathrm{T}} = \sum_{i=1}^n (A_i - \widehat{e}_i)(A_i - \widehat{\rho})\boldsymbol{X}_i$, and $B_{33} = \sum_{i=1}^n (A_i - \widehat{\rho})^2$, we obtain

$$\begin{aligned} \sum_{i=1}^n \widehat{L}_{2i}^2/n^2 &= \sum_{i=1}^n (n\widehat{V})^{-2}\left\{\widehat{w}_i^2\widehat{L}_{0i}^2 + \widehat{d}_2^2(A_i - \widehat{\rho})^2 + \widehat{\boldsymbol{d}}_3^{\mathrm{T}}(A_i - \widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{d}}_3 + 2\widehat{w}_i\widehat{L}_{0i}\widehat{d}_2(A_i - \widehat{\rho})\right. \\ &\quad \left. + 2\widehat{w}_i\widehat{L}_{0i}\widehat{\boldsymbol{d}}_3^{\mathrm{T}}(A_i - \widehat{e}_i)\boldsymbol{X}_i + 2\widehat{d}_2(A_i - \widehat{e}_i)(A_i - \widehat{\rho})\widehat{\boldsymbol{d}}_3^{\mathrm{T}}\boldsymbol{X}_i\right\} \\ &= \frac{1}{A_{11}^2}\left\{B_{11} + \widehat{d}_2^2B_{33} + \widehat{\boldsymbol{d}}_3^{\mathrm{T}}\boldsymbol{B}_{22}\widehat{\boldsymbol{d}}_3 + 2\widehat{d}_2B_{13} + 2\widehat{\boldsymbol{d}}_3^{\mathrm{T}}\boldsymbol{B}_{12}^{\mathrm{T}} + 2\widehat{d}_2\widehat{\boldsymbol{d}}_3^{\mathrm{T}}\boldsymbol{B}_{23}^{\mathrm{T}}\right\}. \end{aligned} \tag{9}$$

To show the equivalence of (8) and (9), it suffices to show

$$\widehat{d}_2 = -A_{13} \quad \text{and} \quad \widehat{\boldsymbol{d}}_3 = -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{12}^{\mathrm{T}},$$

which can be done in a similar way to the proof for the conventional inverse probability weights.

Web Appendix C: Robust sandwich variance is conservative

In this proof we show that the robust sandwich variance estimator $\widehat{var}_{\text{RS}}(\widehat{\theta})$ which ignores the uncertainty in propensity score estimation is conservative, by comparing it with the proposed precise variance estimator $\widehat{var}_{\text{CS}}(\widehat{\theta})$.

Conventional inverse probability weights

The proposed variance $\widehat{var}_{\text{CS}}(\widehat{\theta})$ is

$$\begin{aligned}\sum_{i=1}^{n}\widehat{L}_{1i}^2/n^2 &= \sum_{i=1}^{n}(n\widehat{V})^{-2}\left\{\widehat{w}_i^2\widehat{L}_{0i}^2+2\widehat{w}_i\widehat{L}_{0i}\widehat{\boldsymbol{d}}_1^{\text{T}}(A_i-\widehat{e}_i)\boldsymbol{X}_i+\widehat{\boldsymbol{d}}_1^{\text{T}}(A_i-\widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\text{T}}\widehat{\boldsymbol{d}}_1\right\}\\ &= \frac{1}{A_{11}^2}\sum_{i=1}^{n}\widehat{w}_i^2\widehat{L}_{0i}^2+\frac{1}{A_{11}^2}\left[\widehat{\boldsymbol{d}}_1^{\text{T}}\sum_{i=1}^{n}\{(A_i-\widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\text{T}}\}\widehat{\boldsymbol{d}}_1+2\widehat{\boldsymbol{d}}_1^{\text{T}}\sum_{i=1}^{n}\widehat{w}_i\widehat{L}_{0i}(A_i-\widehat{e}_i)\boldsymbol{X}_i\right],\end{aligned}$$

where $\frac{1}{A_{11}^2}\sum_{i=1}^{n}\widehat{w}_i^2\widehat{L}_{0i}^2$ is the robust sandwich variance estimator $\widehat{var}_{\text{RS}}(\widehat{\theta})$. Thus we obtain

$$\widehat{var}_{\text{CS}}(\sqrt{n}\widehat{\theta})=\widehat{var}_{\text{RS}}(\sqrt{n}\widehat{\theta})+\frac{1}{(A_{11}/n)^2}\left[\widehat{\boldsymbol{d}}_1^{\text{T}}\frac{1}{n}\sum_{i=1}^{n}\{(A_i-\widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\text{T}}\}\widehat{\boldsymbol{d}}_1+2\widehat{\boldsymbol{d}}_1^{\text{T}}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_i\widehat{L}_{0i}(A_i-\widehat{e}_i)\boldsymbol{X}_i\right].$$

To show that the robust sandwich variance estimator is conservative, we need to show

$$\widehat{\boldsymbol{d}}_1^{\text{T}}\frac{1}{n}\sum_{i=1}^{n}\{(A_i-\widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\text{T}}\}\widehat{\boldsymbol{d}}_1+2\widehat{\boldsymbol{d}}_1^{\text{T}}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_i\widehat{L}_{0i}(A_i-\widehat{e}_i)\boldsymbol{X}_i\le 0\quad\text{as}\quad n\to\infty. \tag{10}$$

Let $g_j=-A_j\frac{1-\widehat{e}_j}{\widehat{e}_j}+(1-A_j)\frac{\widehat{e}_j}{1-\widehat{e}_j}$. Then $\widehat{\boldsymbol{d}}_1=\widehat{\boldsymbol{U}}^{-1}\left\{\frac{1}{n}\sum_{j=1}^{n}g_j\widehat{L}_{0j}\boldsymbol{X}_j\right\}$. Note

$$\widehat{w}_i(A_i-\widehat{e}_i)=\left(\frac{A_i}{\widehat{e}_i}+\frac{1-A_i}{1-\widehat{e}_i}\right)(A_i-\widehat{e}_i)=A_i\frac{1-\widehat{e}_i}{\widehat{e}_i}-(1-A_i)\frac{\widehat{e}_i}{1-\widehat{e}_i}=-g_i.$$

The left-hand-side of (10) equals

$$
\begin{aligned}
& \left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\}^{\mathrm{T}} \widehat{\boldsymbol{U}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\{(A_i-\widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\}\widehat{\boldsymbol{U}}^{-1}\left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\} \\
& +2\left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\}^{\mathrm{T}} \widehat{\boldsymbol{U}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_i\widehat{L}_{0i}(A_i-\widehat{e}_i)\boldsymbol{X}_i \\
\approx\ & \left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\}^{\mathrm{T}} \widehat{\boldsymbol{U}}^{-1}\left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\} - 2\left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\}^{\mathrm{T}} \widehat{\boldsymbol{U}}^{-1}\left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\} \\
=\ & -\left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\}^{\mathrm{T}} \widehat{\boldsymbol{U}}^{-1}\left\{\frac{1}{n}\sum_{j=1}^{n} g_j \widehat{L}_{0j}\boldsymbol{X}_j\right\} \\
\leq\ & 0,
\end{aligned}
$$

where $\widehat{\boldsymbol{U}} \approx \frac{1}{n}\sum_{i=1}^{n}\{(A_i-\widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\}$ is used in the first step, because they converge to the same limiting value as $n \to \infty$:

$$\widehat{\boldsymbol{U}} = \frac{1}{n}\sum_{j=1}^{n}\widehat{e}_j(1-\widehat{e}_j)\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}} \to E\{e(1-e)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\{(A_i-\widehat{e}_i)^2\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}\} \to E\{(A-e)^2\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\} = E\{(A-2Ae+e^2)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\} = E\{e(1-e)\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}.$$

Therefore, as $n \to \infty$, $\widehat{var}_{\mathrm{CS}}(\sqrt{n}\widehat{\theta}) \leq \widehat{var}_{\mathrm{RS}}(\sqrt{n}\widehat{\theta})$, indicating that the robust sandwich variance estimator is conservative.

Stabilized weights

In an approximate sense with the treatment prevalence assumed known, it can be shown that the robust sandwich variance estimator with stabilized weights is also conservative in a similar way to the proof for the conventional inverse probability weights.

Web Appendix D: Additional simulation results without clustering

Web Appendix E: Additional simulation results with clustering

Web Appendix F: Balance diagnostics in case study

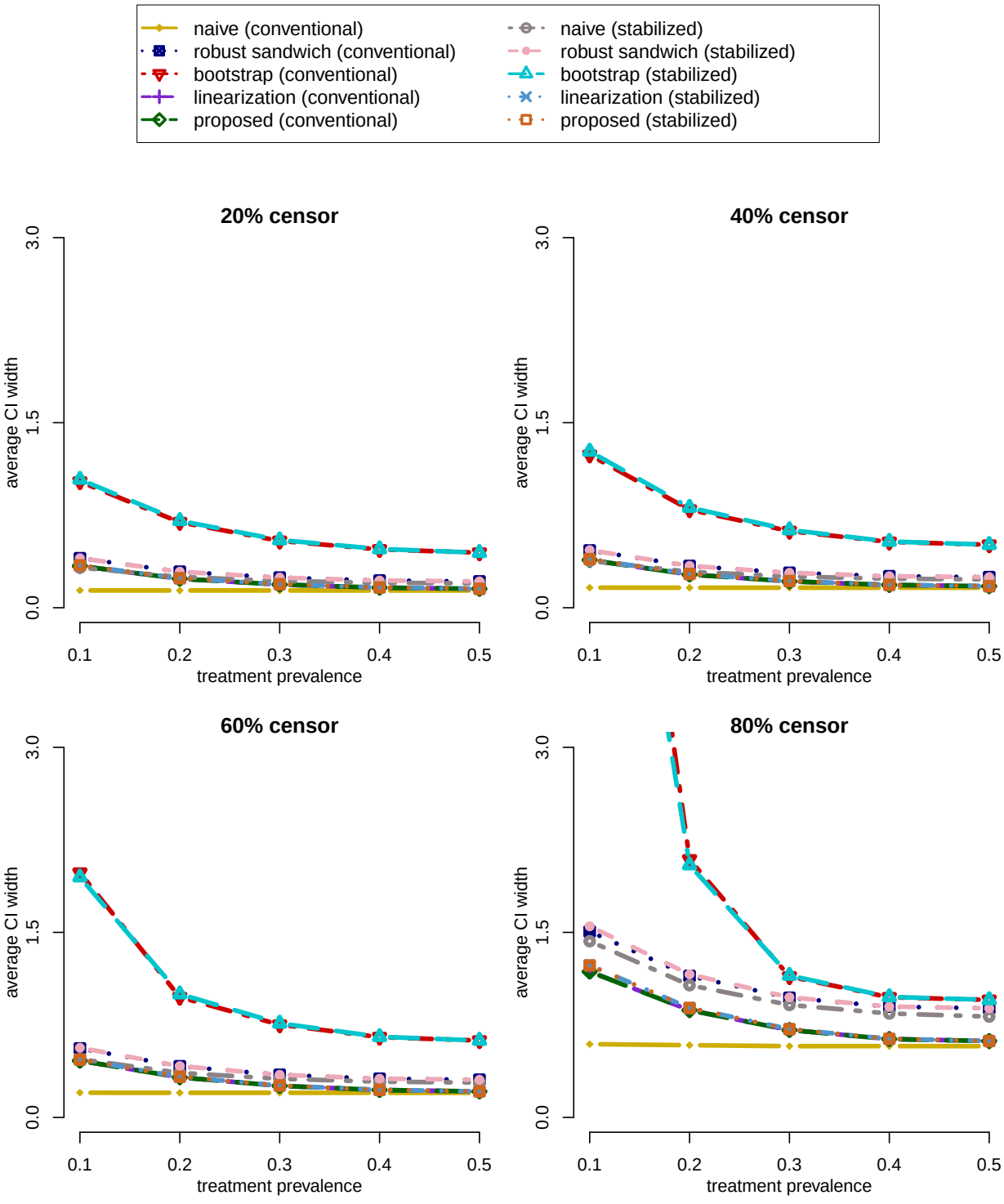

Figure S1: Average widths of 95% confidence intervals with $n = 250$. Total number of failure events is about 200, 150, 100, or 50.

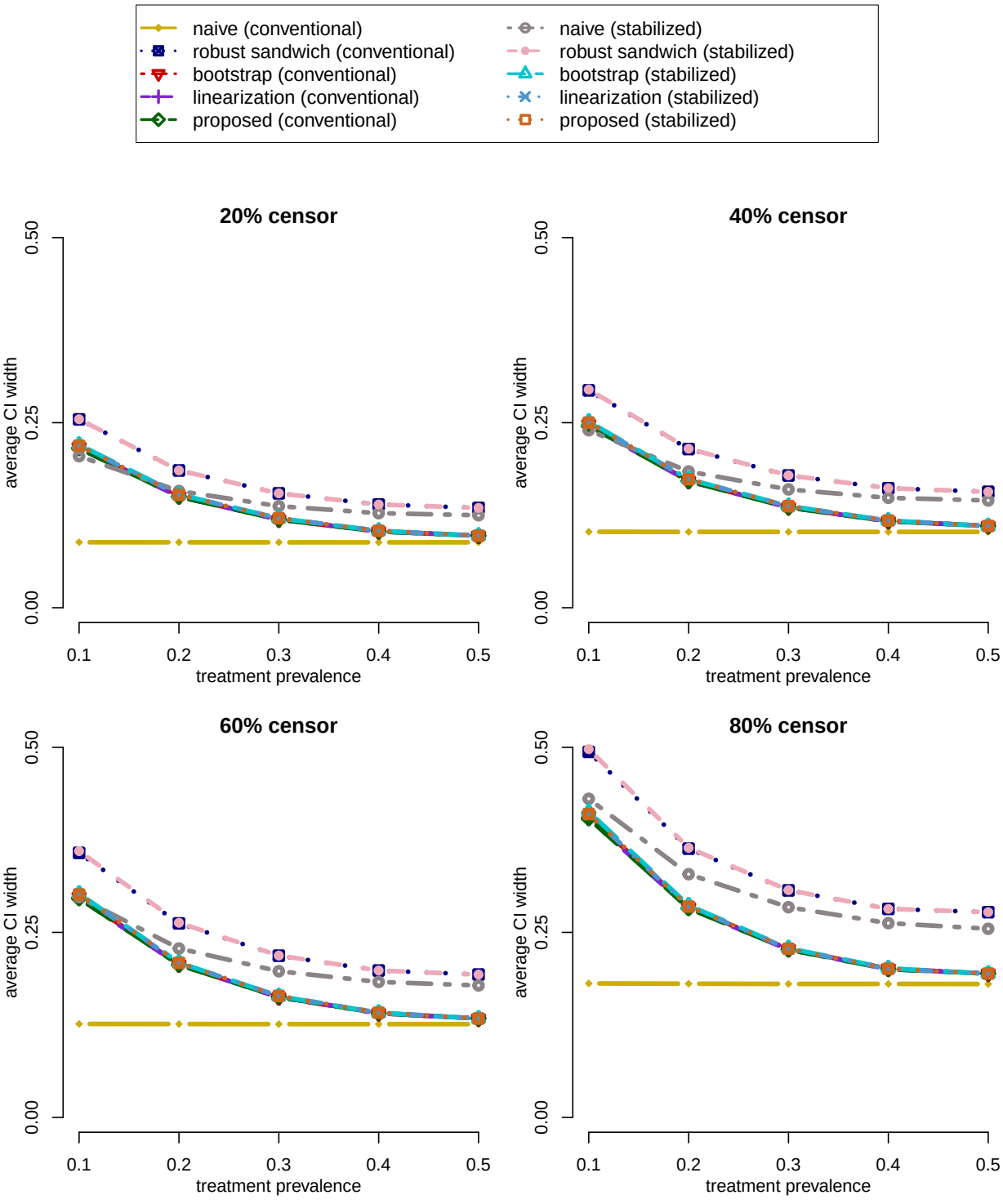

Figure S2: Average widths of 95% confidence intervals with $n = 5000$. Total number of failure events is about 4000, 3000, 2000, or 1000.

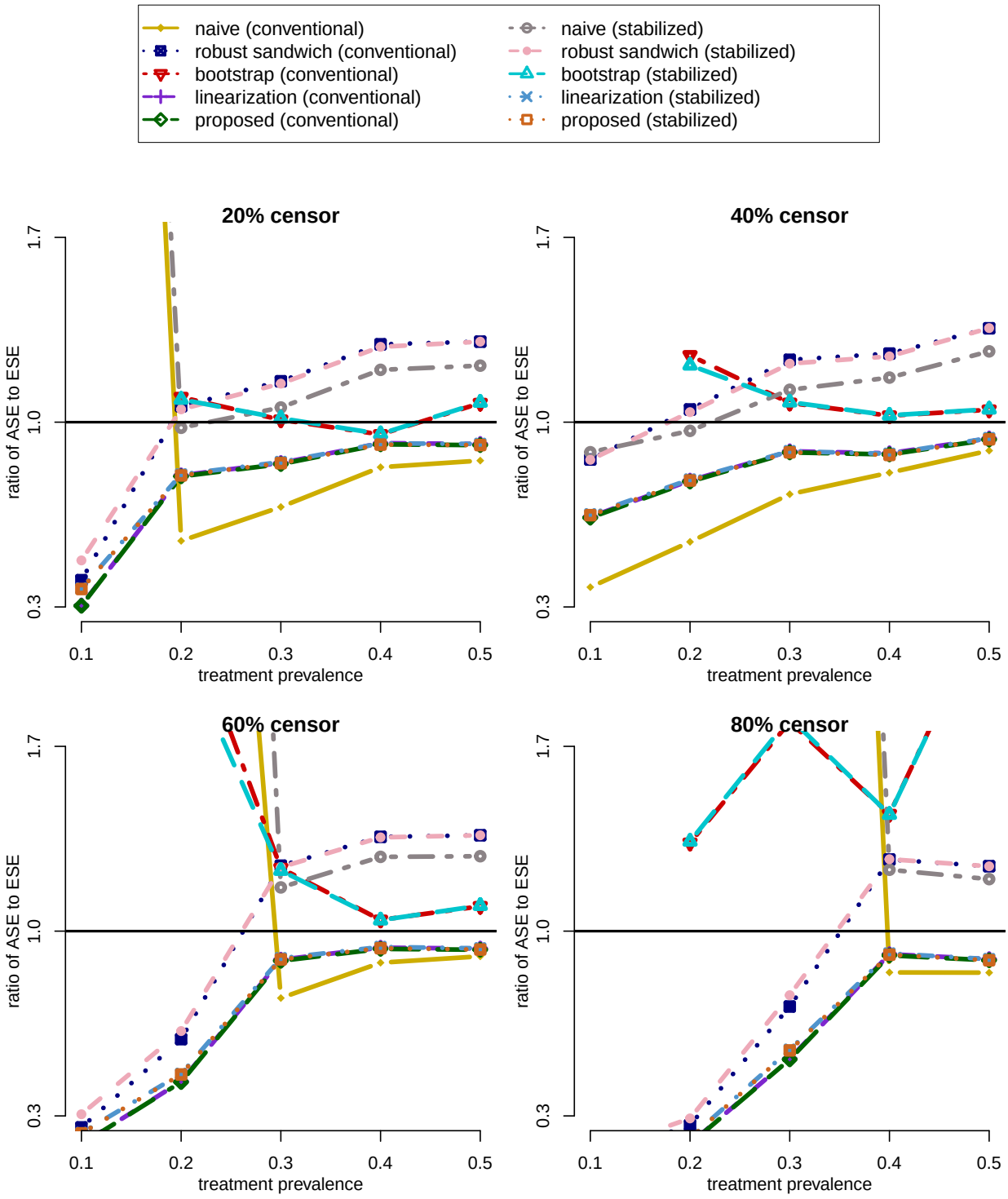

Figure S3: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 100$. Total number of failure events is about 80, 60, 40, or 20. Note: in some scenarios with low prevalence or high censoring rate, unestimable (i.e., error message from R) or extreme results occur and hence not shown in the figure.

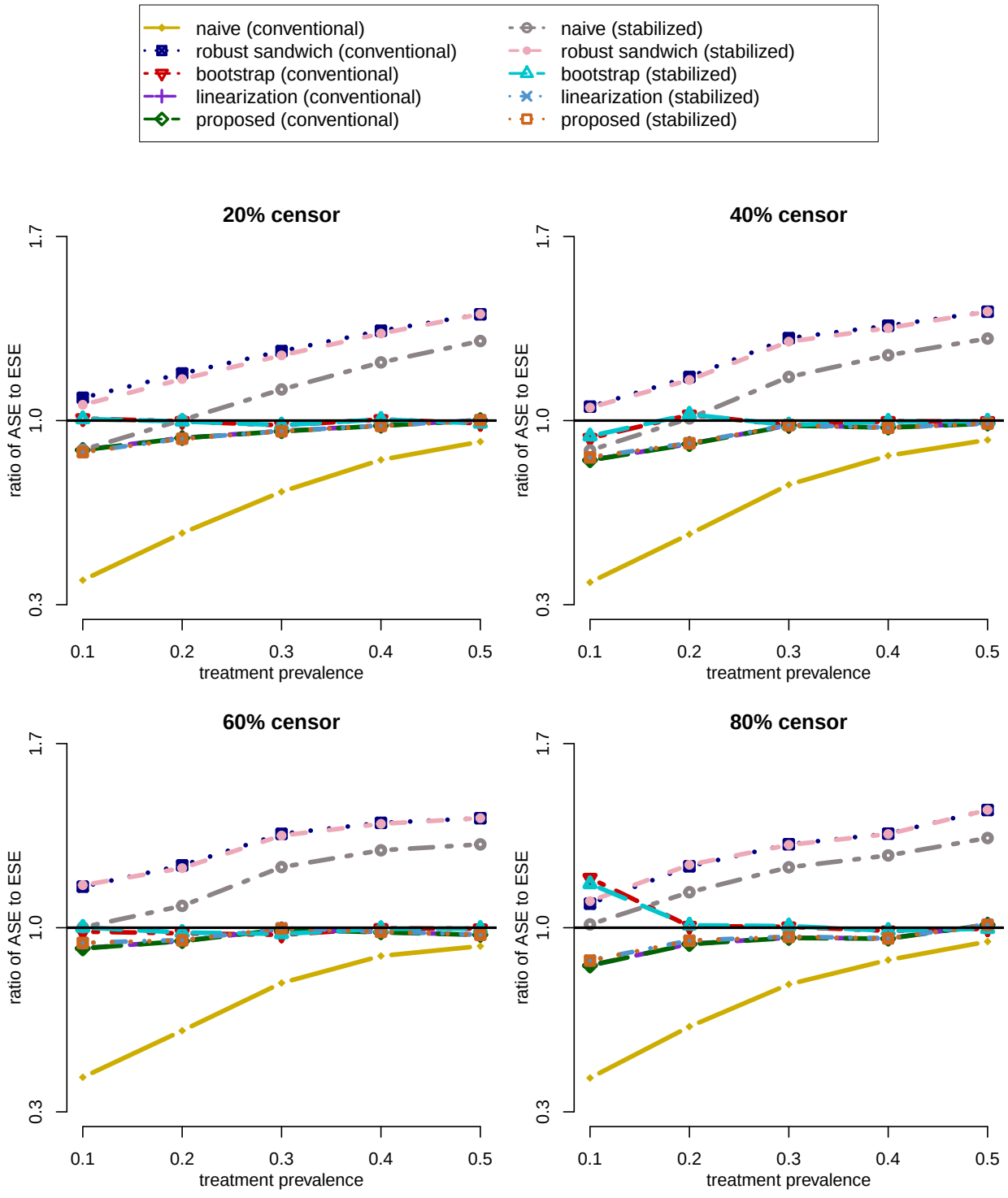

Figure S4: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 500$. Total number of failure events is about 400, 300, 200, or 100.

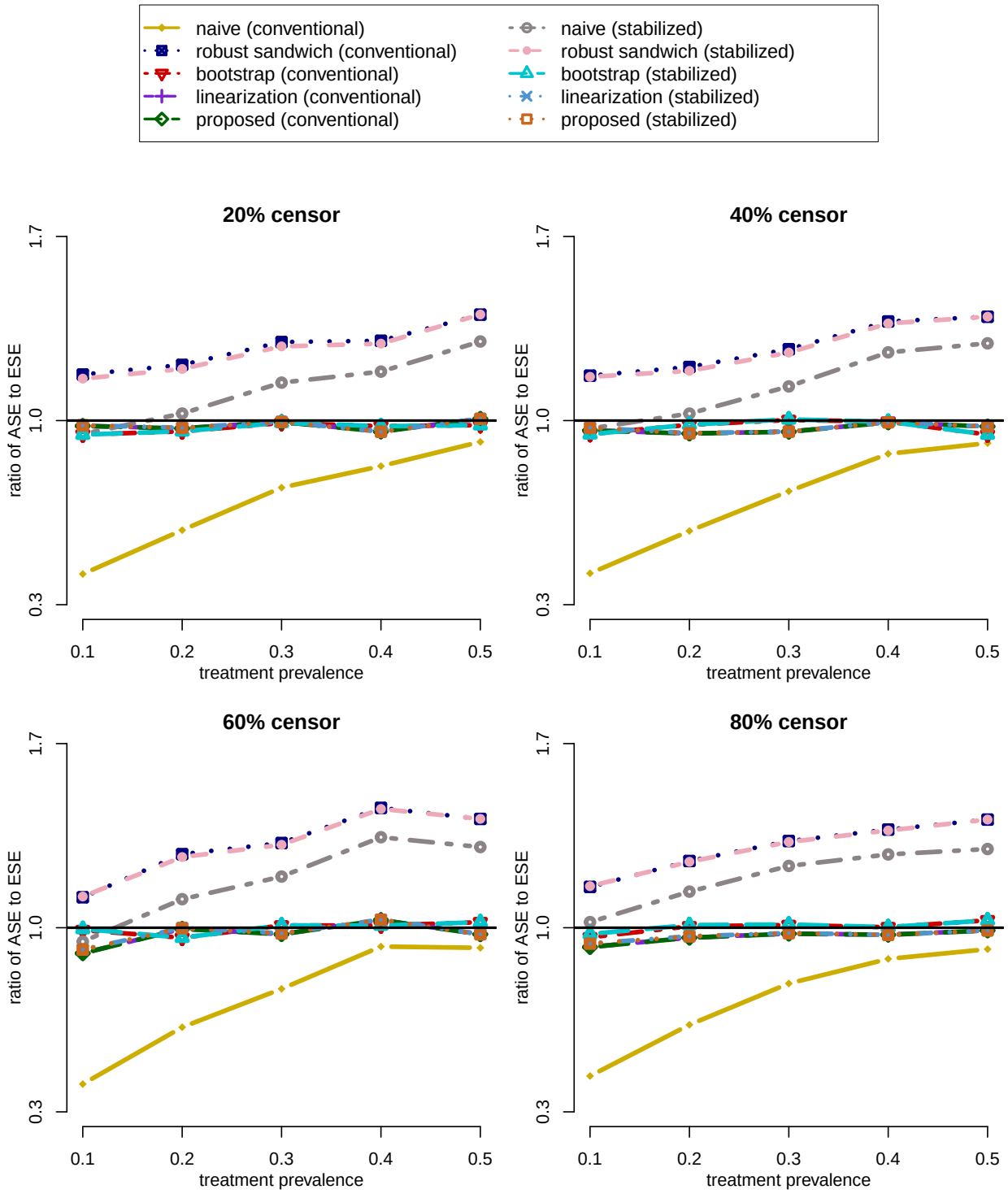

Figure S5: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 1000$. Total number of failure events is about 800, 600, 400, or 200.

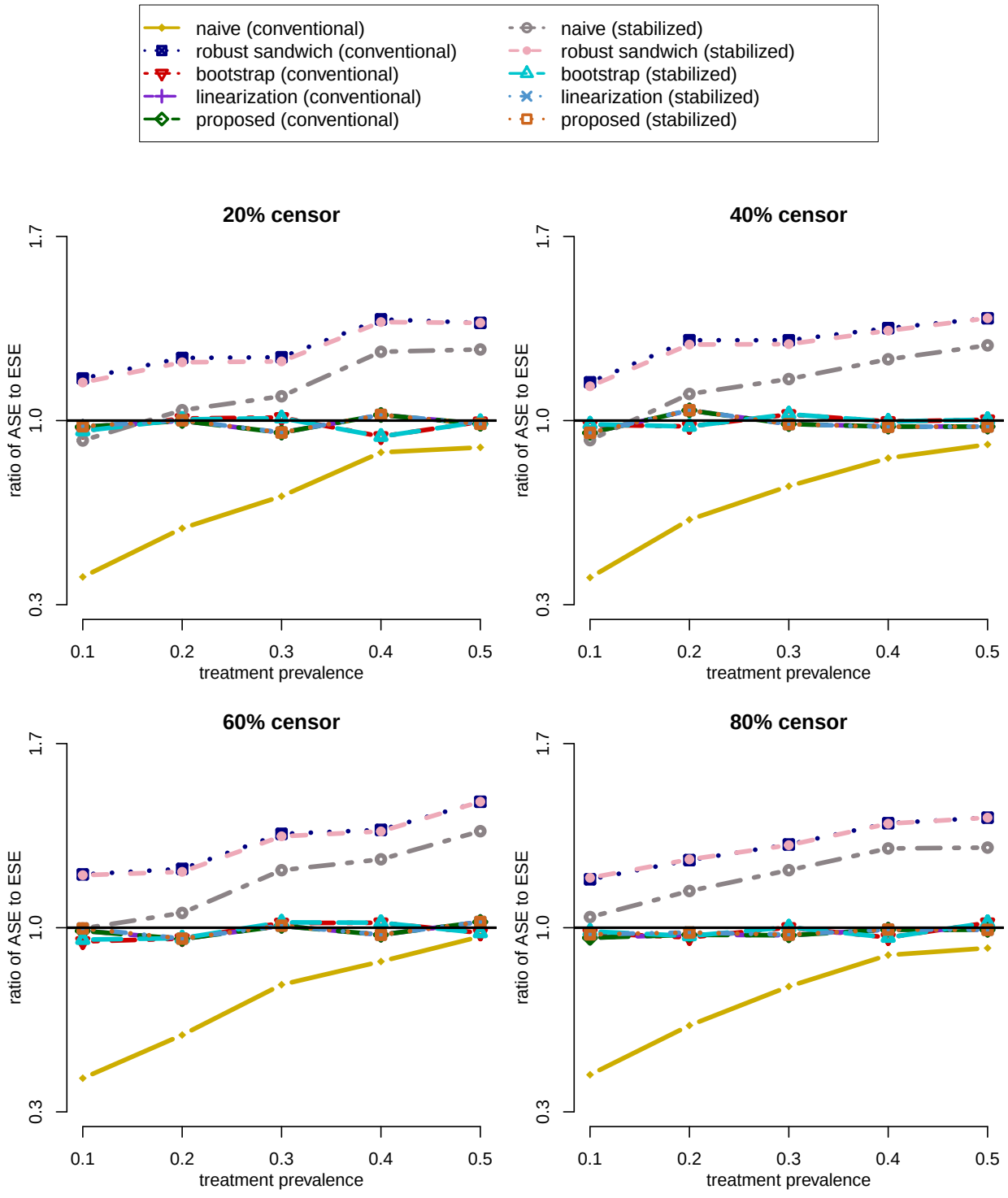

Figure S6: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 2000$. Total number of failure events is about 1600, 1200, 800, or 400.

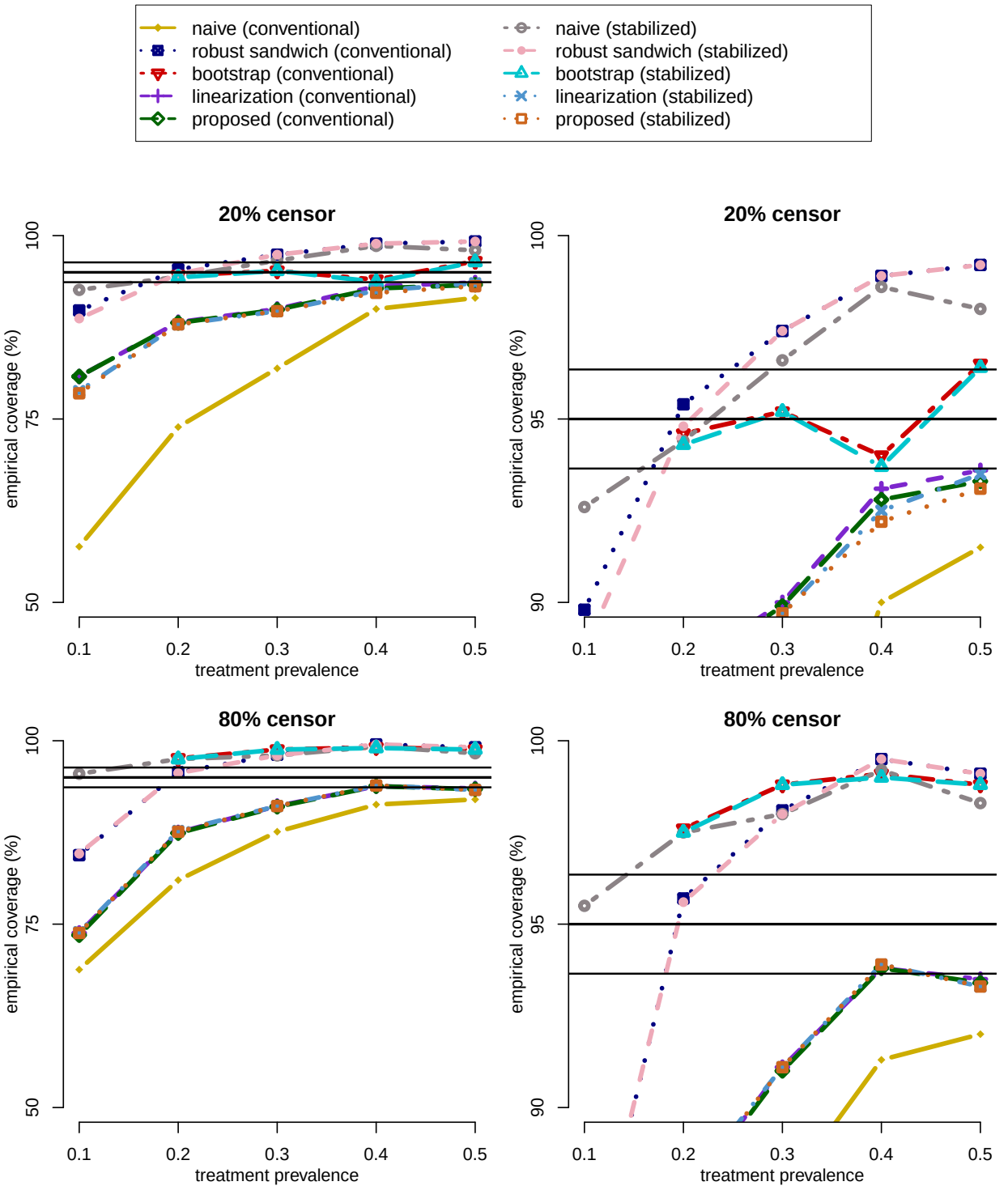

Figure S7: Empirical coverage rates in percent with $n = 100$. Total number of failure events is about 80 or 20. The right panel shows a zoom-in version of the left panel. Note: in some scenarios with low prevalence or high censoring rate, unestimable (i.e., error message from R) results occur and hence not shown in the figure.

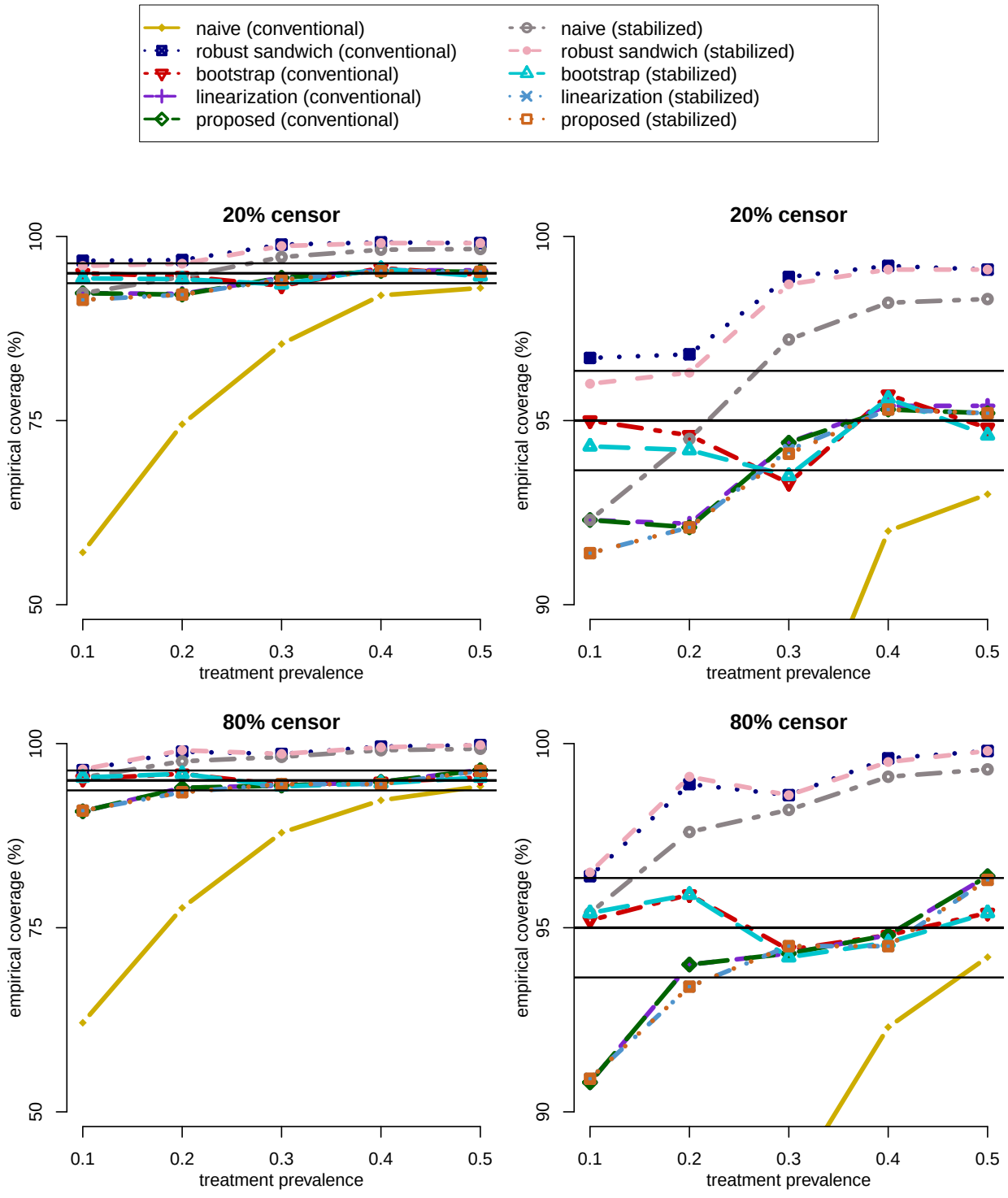

Figure S8: Empirical coverage rates in percent with $n = 500$. Total number of failure events is about 400 or 100. The right panel shows a zoom-in version of the left panel.

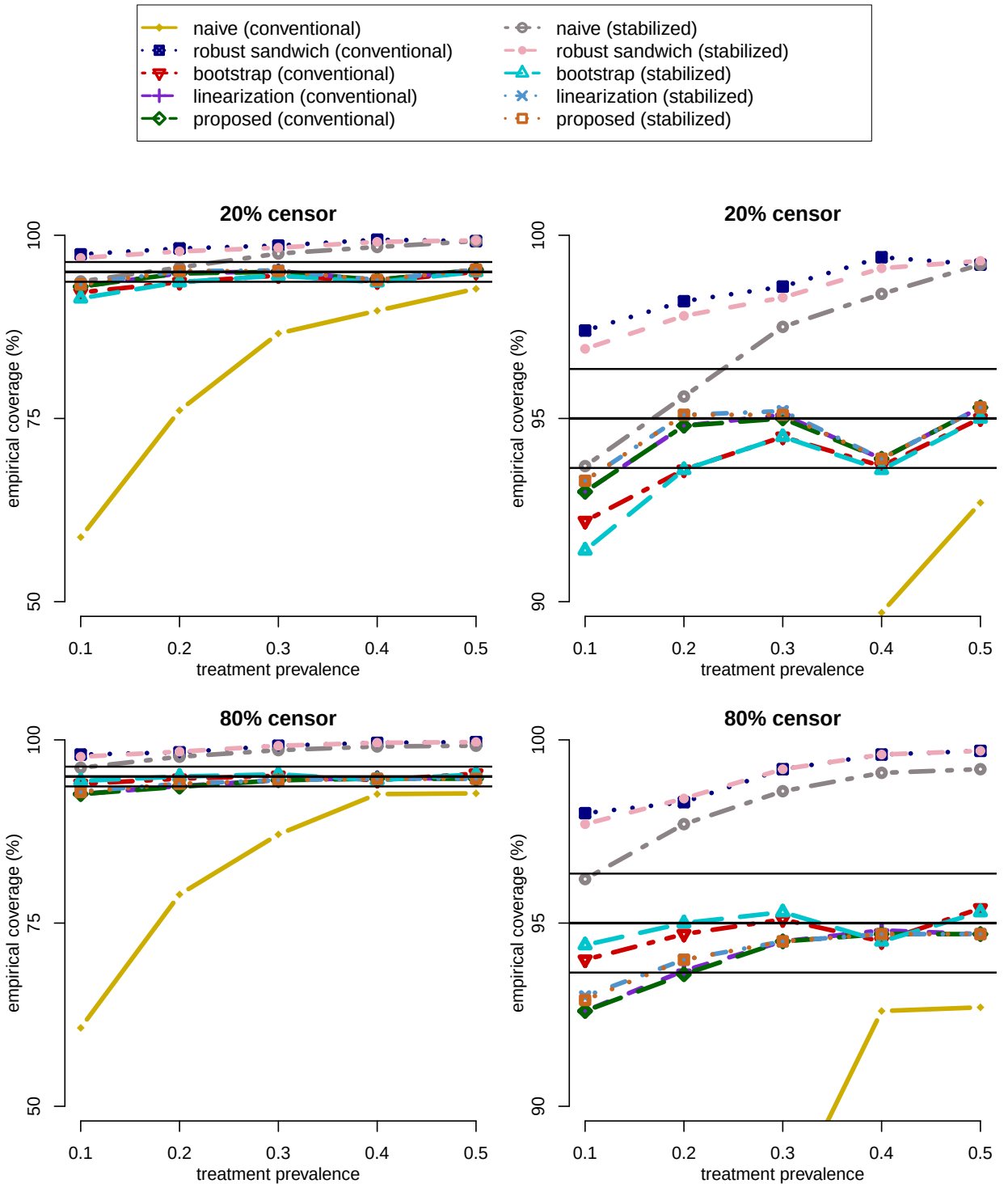

Figure S9: Empirical coverage rates in percent with $n = 1000$. Total number of failure events is about 800 or 200. The right panel shows a zoom-in version of the left panel.

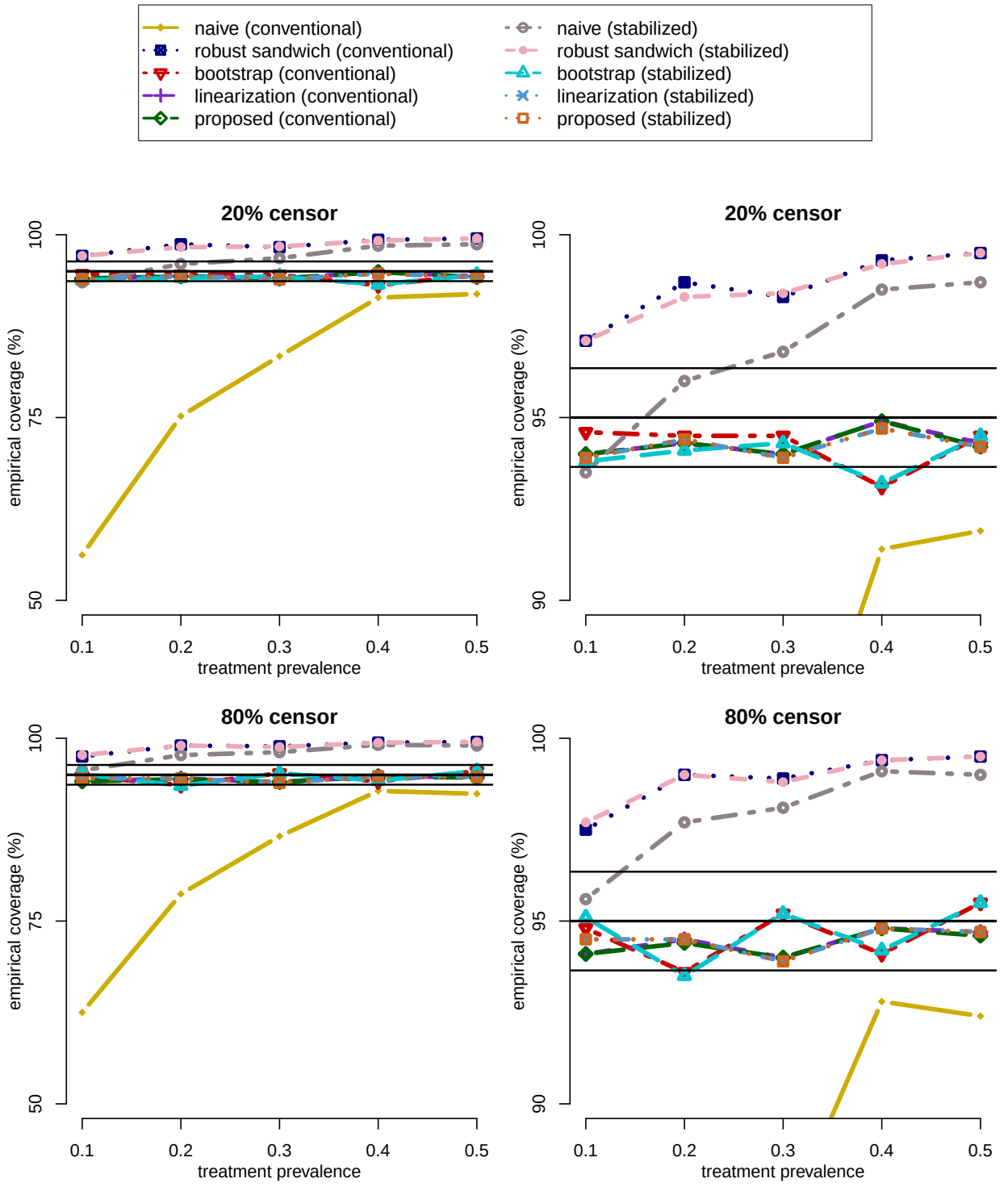

Figure S10: Empirical coverage rates in percent with $n = 2000$. Total number of failure events is about 1600 or 400. The right panel shows a zoom-in version of the left panel.

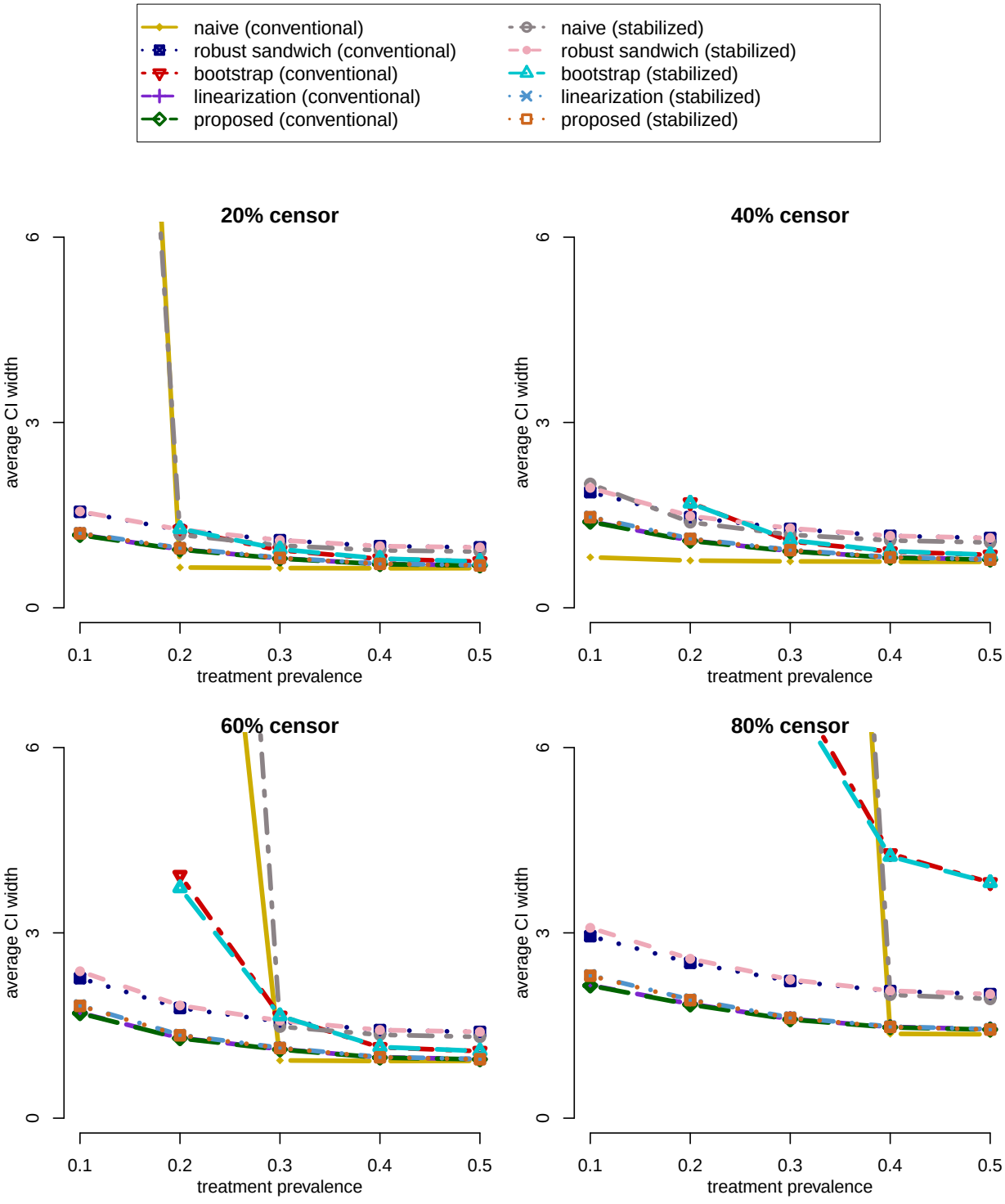

Figure S11: Average widths of 95% confidence intervals with $n = 100$. Total number of failure events is about 80, 60, 40, or 20. Note: in some scenarios with low prevalence or high censoring rate, unestimable (i.e., error message from R) or extreme results occur and hence not shown in the figure.

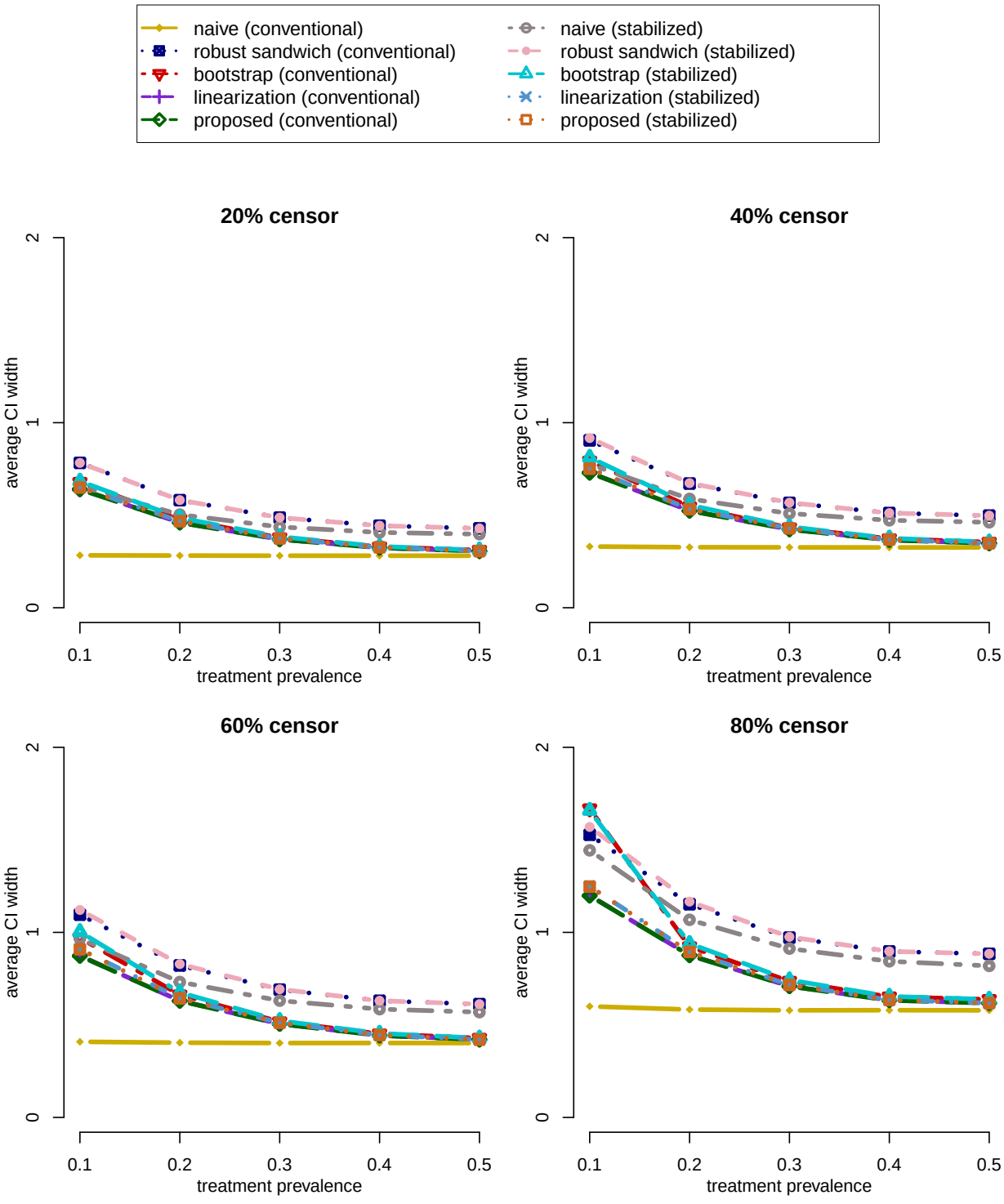

Figure S12: Average widths of 95% confidence intervals with $n = 500$. Total number of failure events is about 400, 300, 200, or 100.

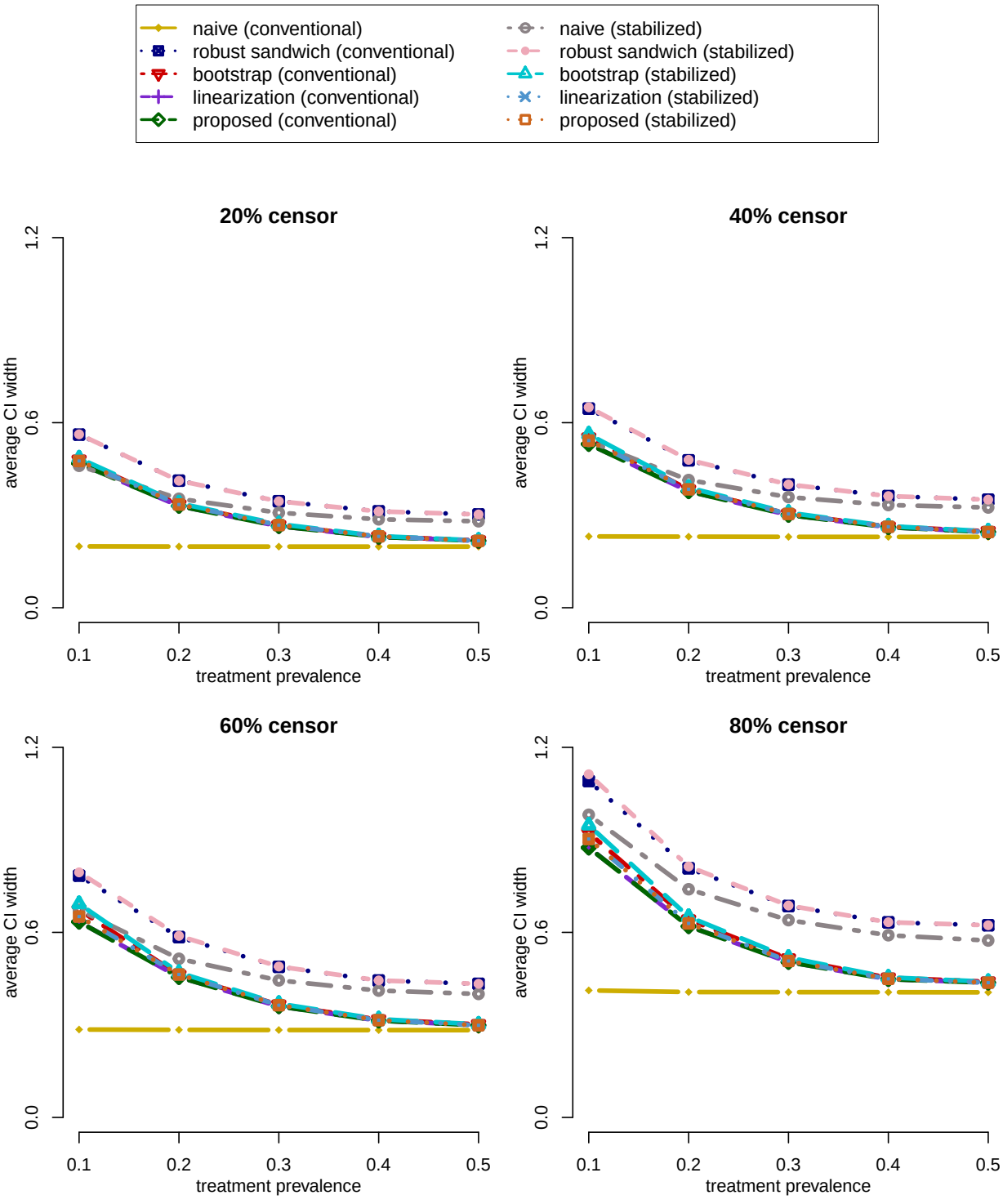

Figure S13: Average widths of 95% confidence intervals with $n = 1000$. Total number of failure events is about 800, 600, 400, or 200.

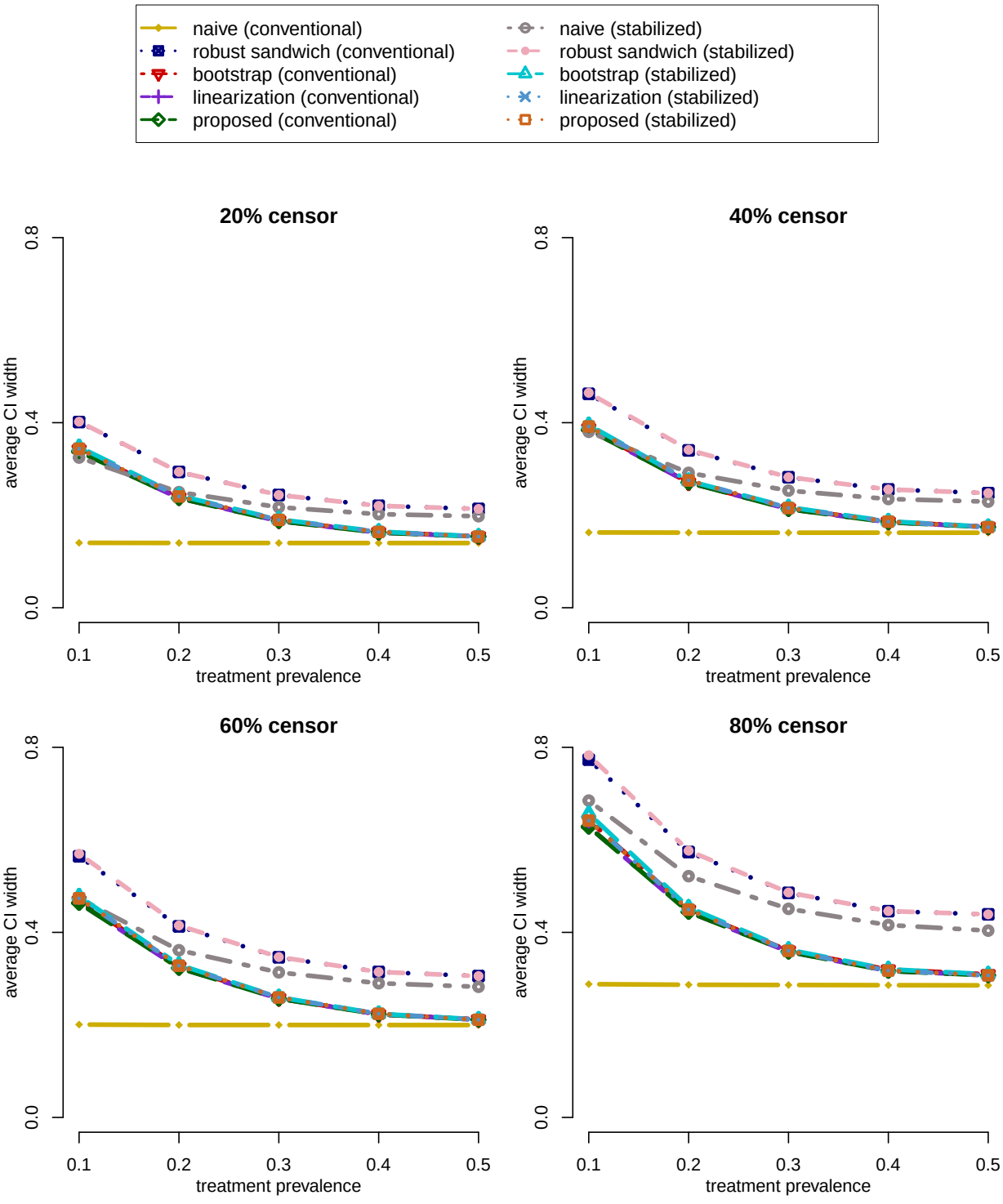

Figure S14: Average widths of 95% confidence intervals with $n = 2000$. Total number of failure events is about 1600, 1200, 800, or 400.

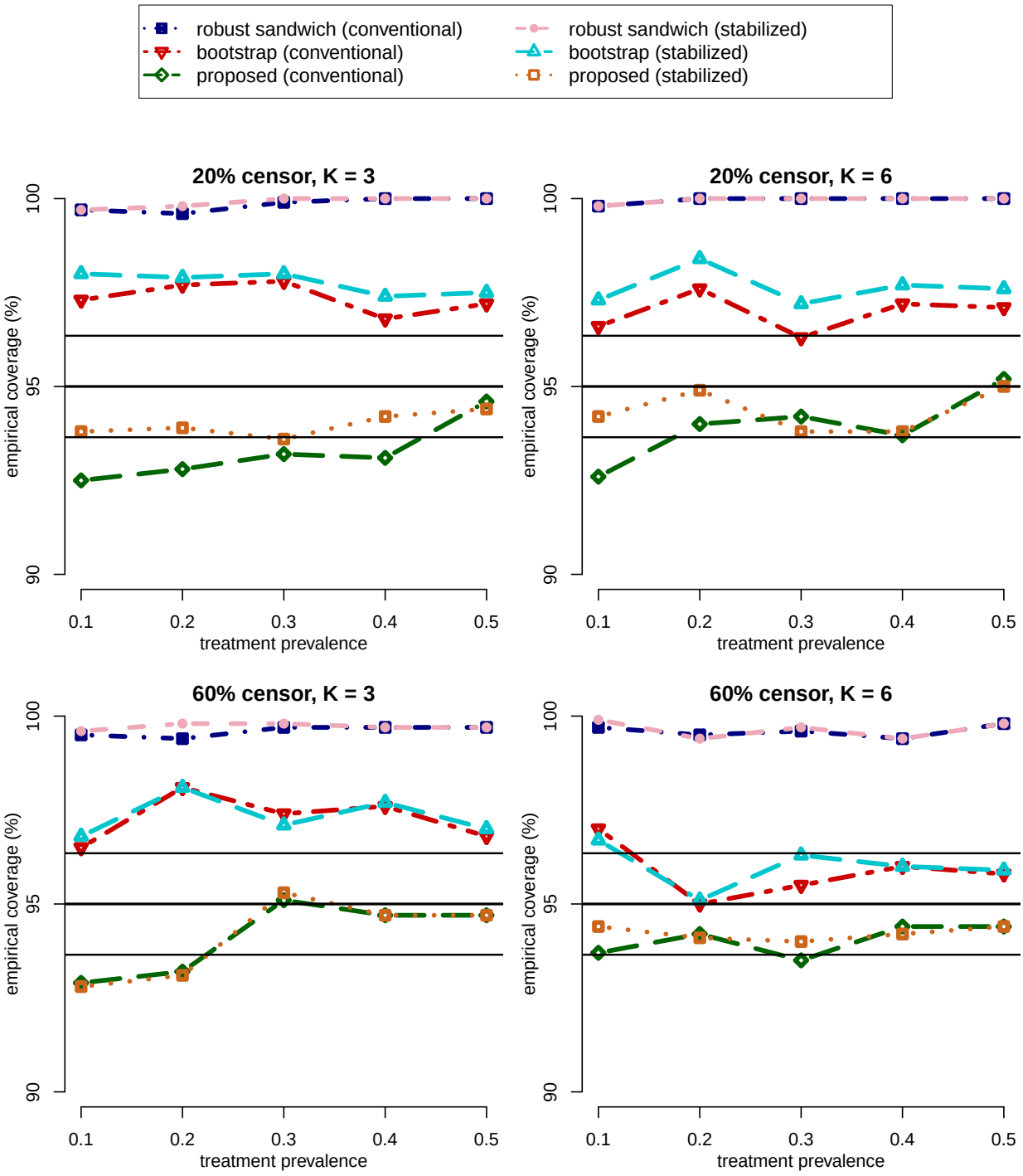

Figure S15: Empirical coverage rates in percent with $n = 80$ clusters each of size K.

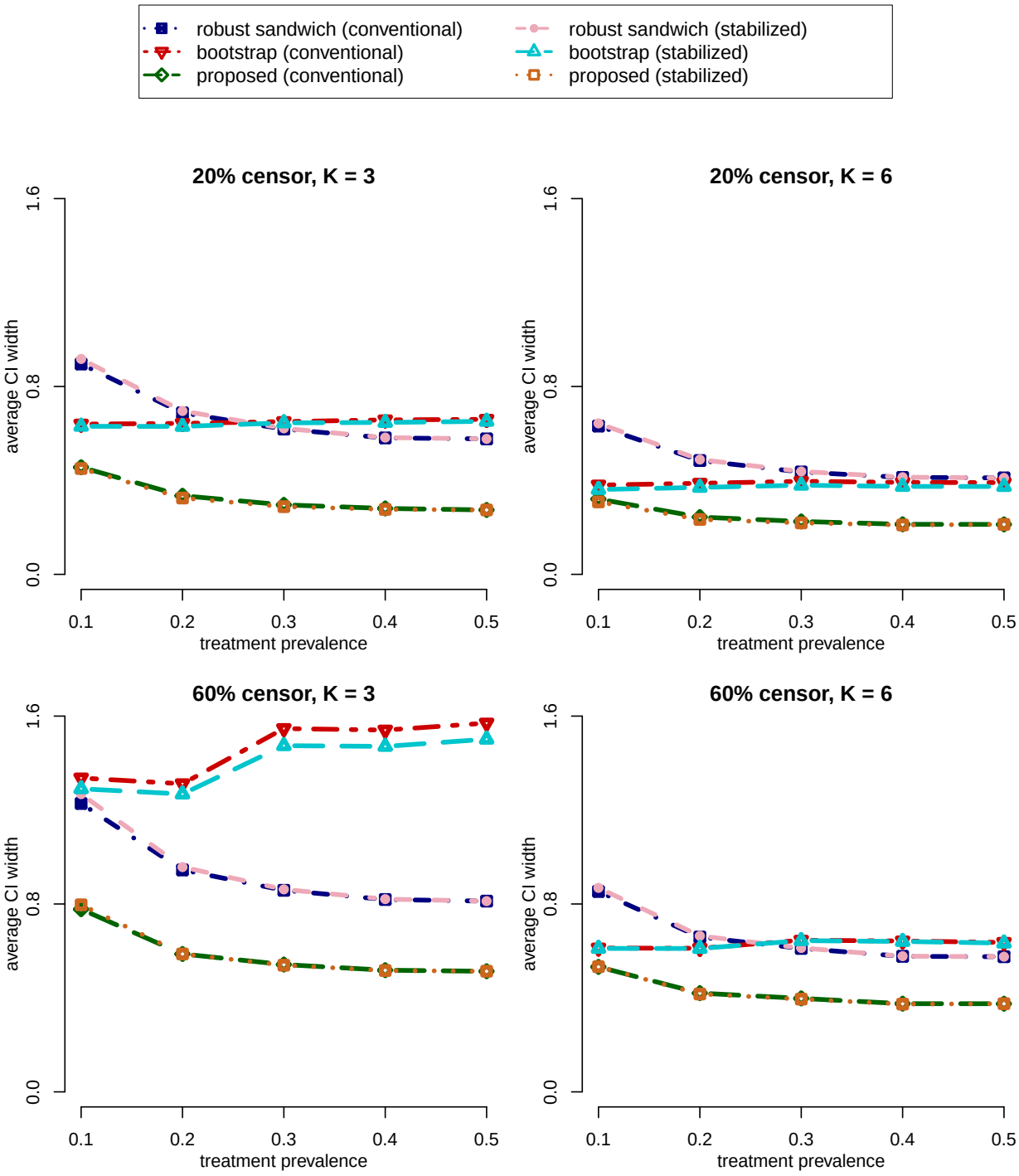

Figure S16: Average widths of 95% confidence intervals with $n = 80$ clusters each of size K.

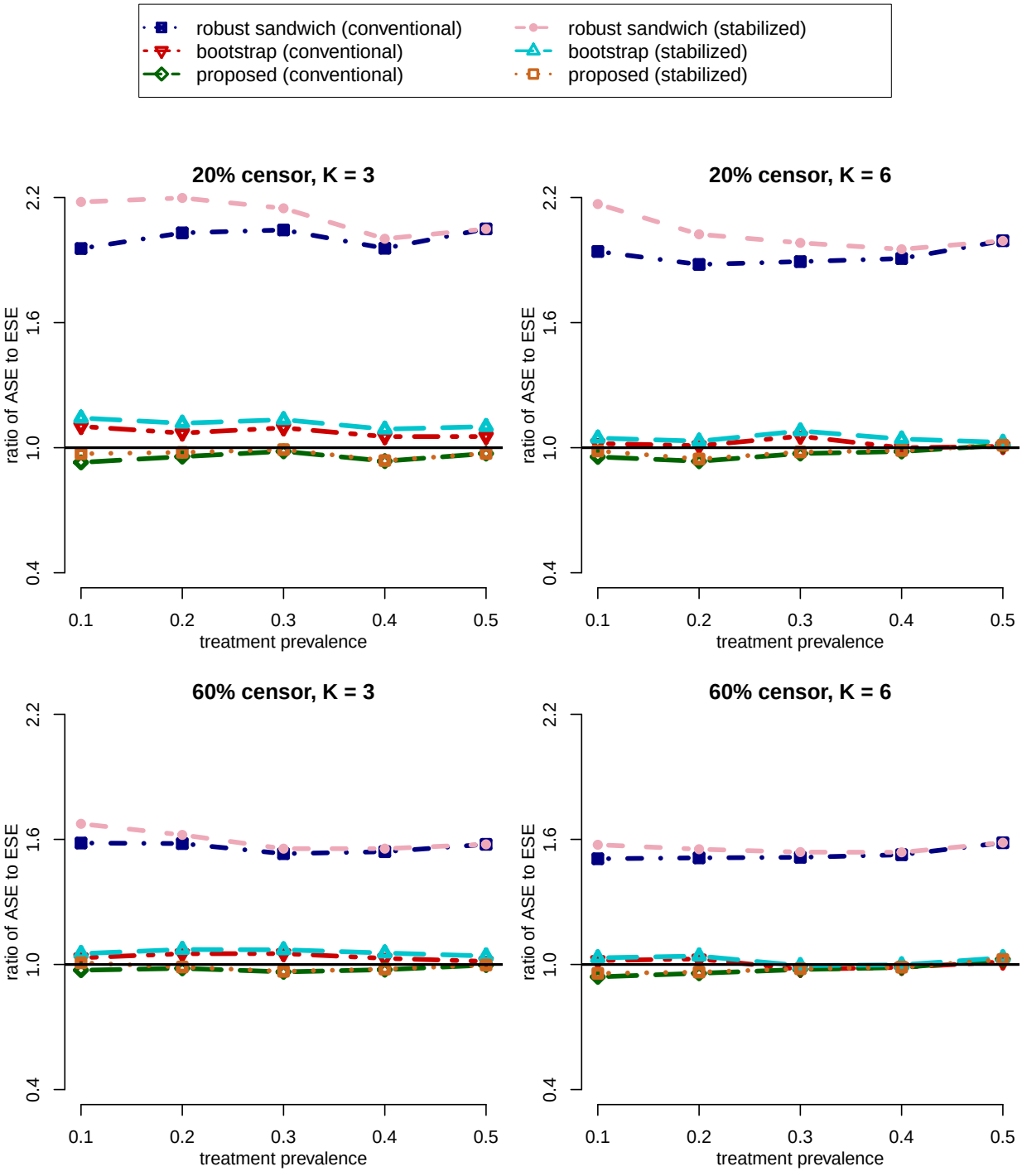

Figure S17: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 200$ clusters each of size K.

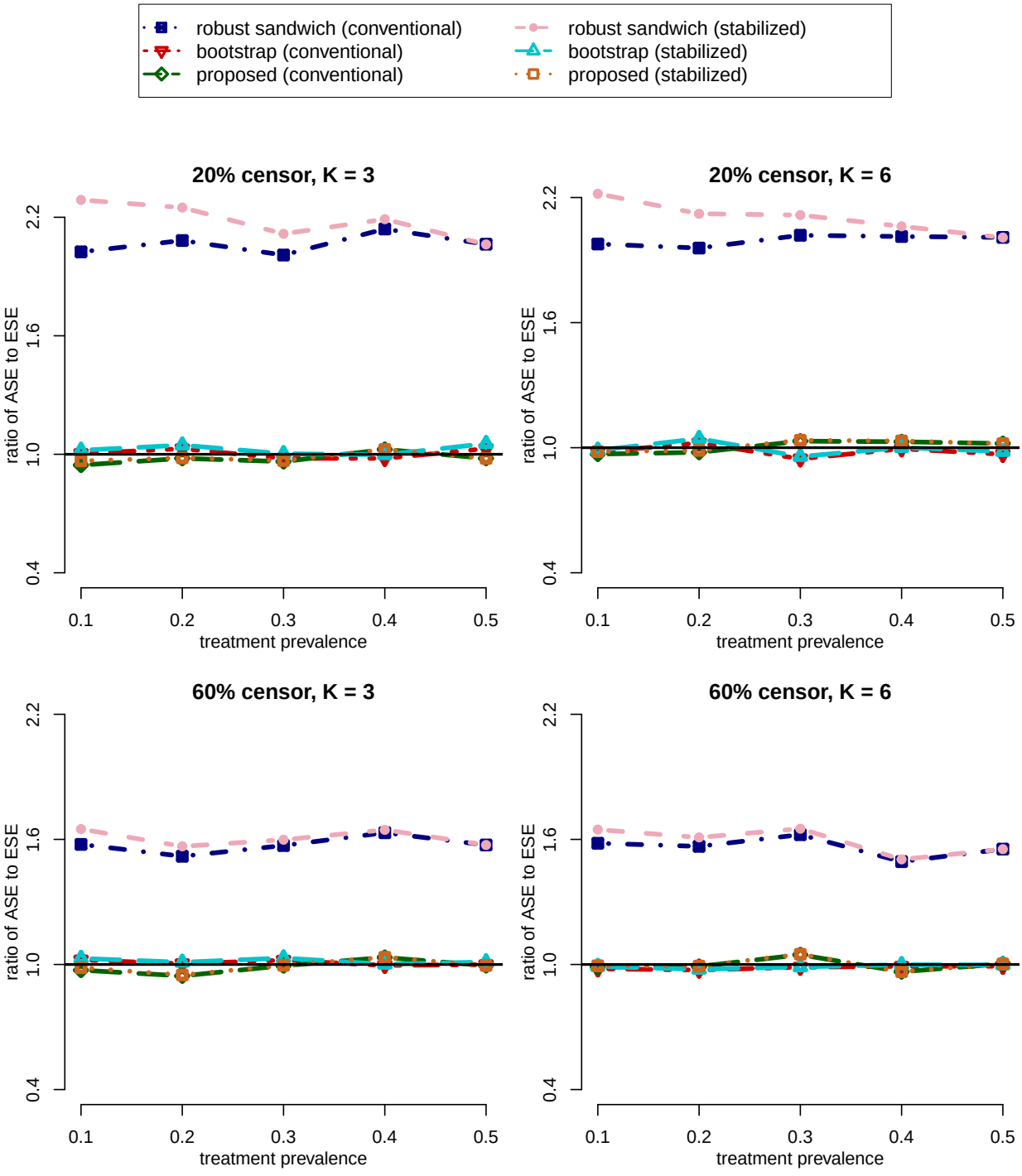

Figure S18: Ratios of average standard error (ASE) to empirical standard error (ESE) with $n = 800$ clusters each of size K.

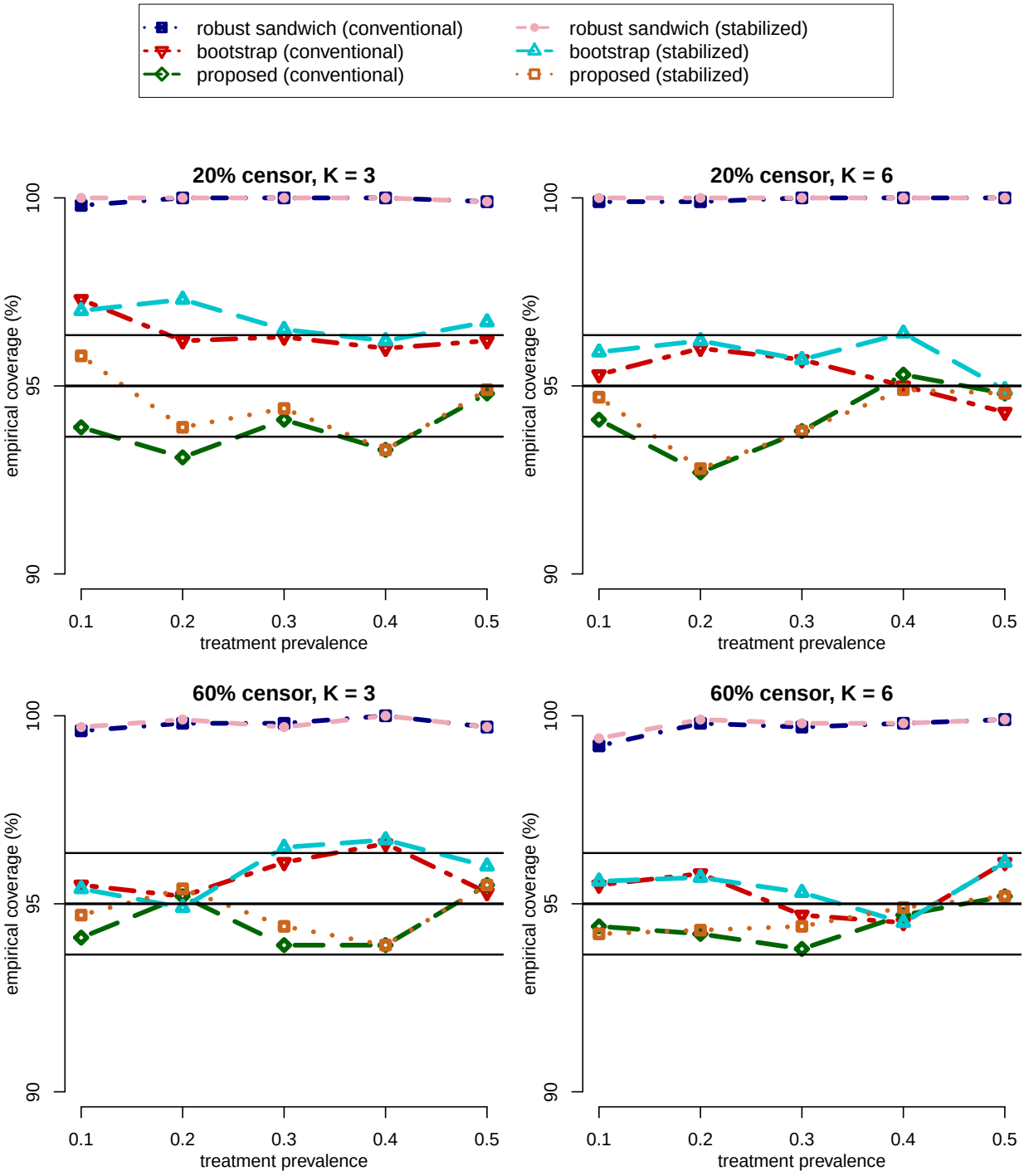

Figure S19: Empirical coverage rates in percent with $n = 200$ clusters each of size K.

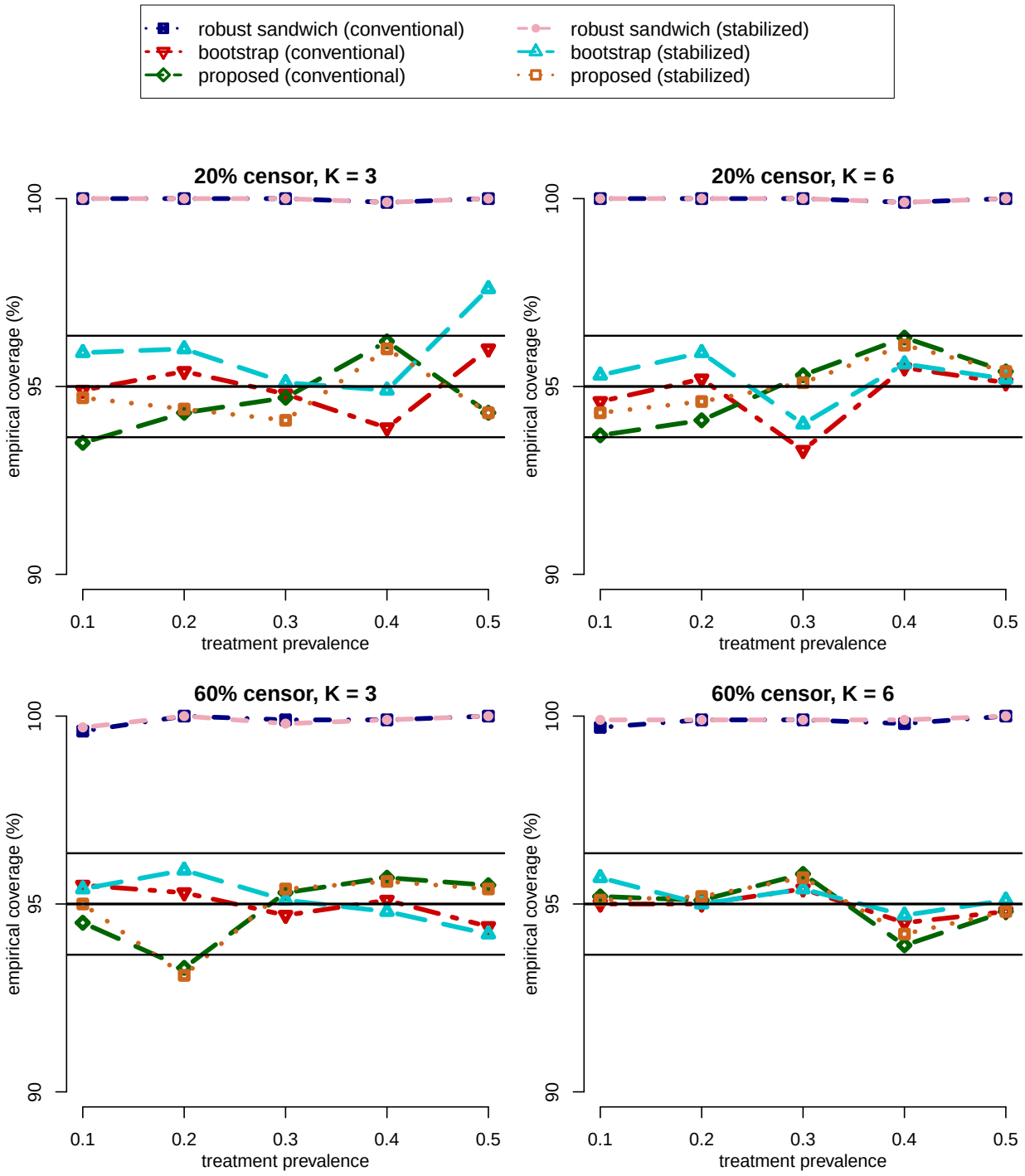

Figure S20: Empirical coverage rates in percent with $n = 800$ clusters each of size K.

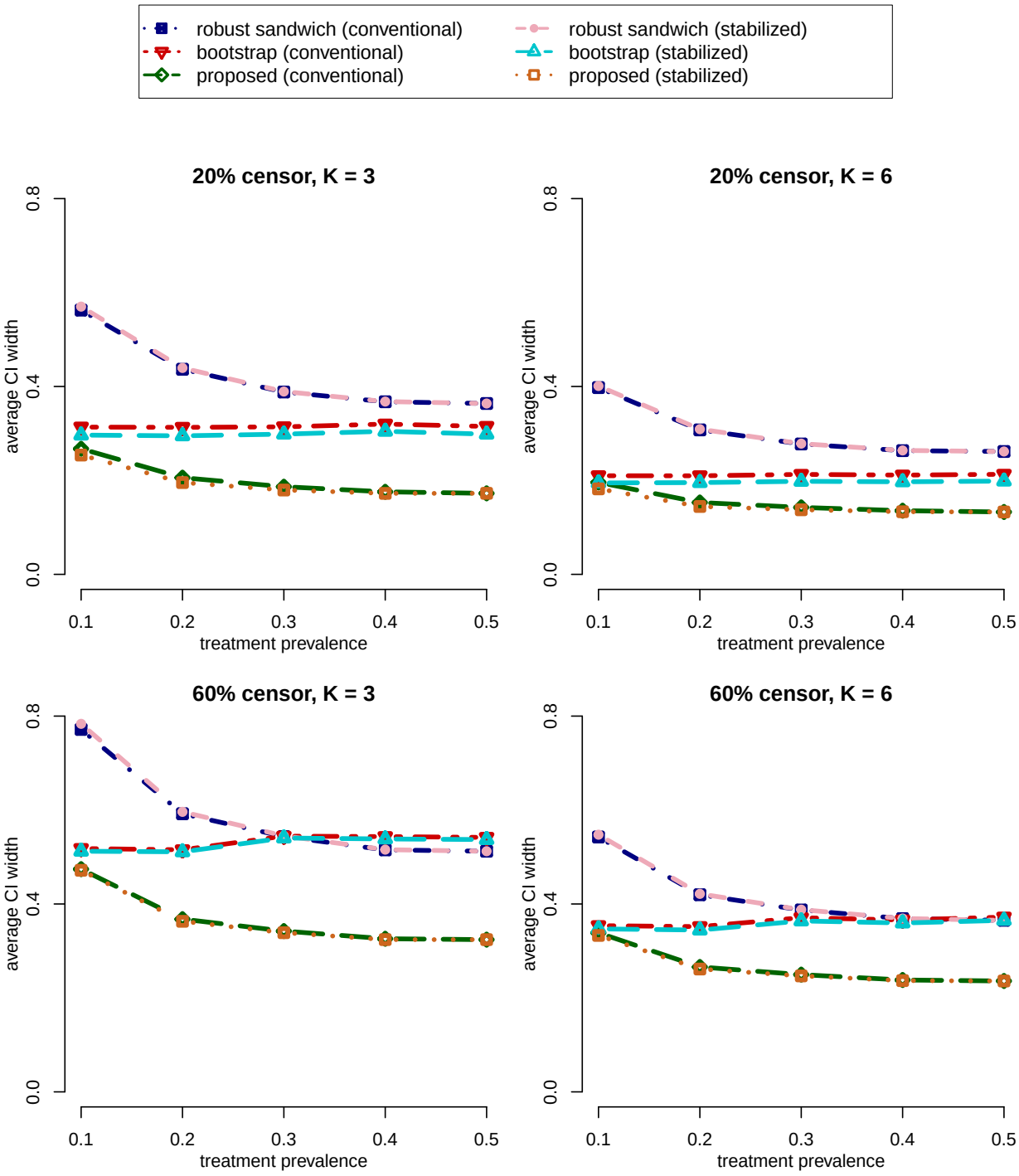

Figure S21: Average widths of 95% confidence intervals with $n = 200$ clusters each of size K.

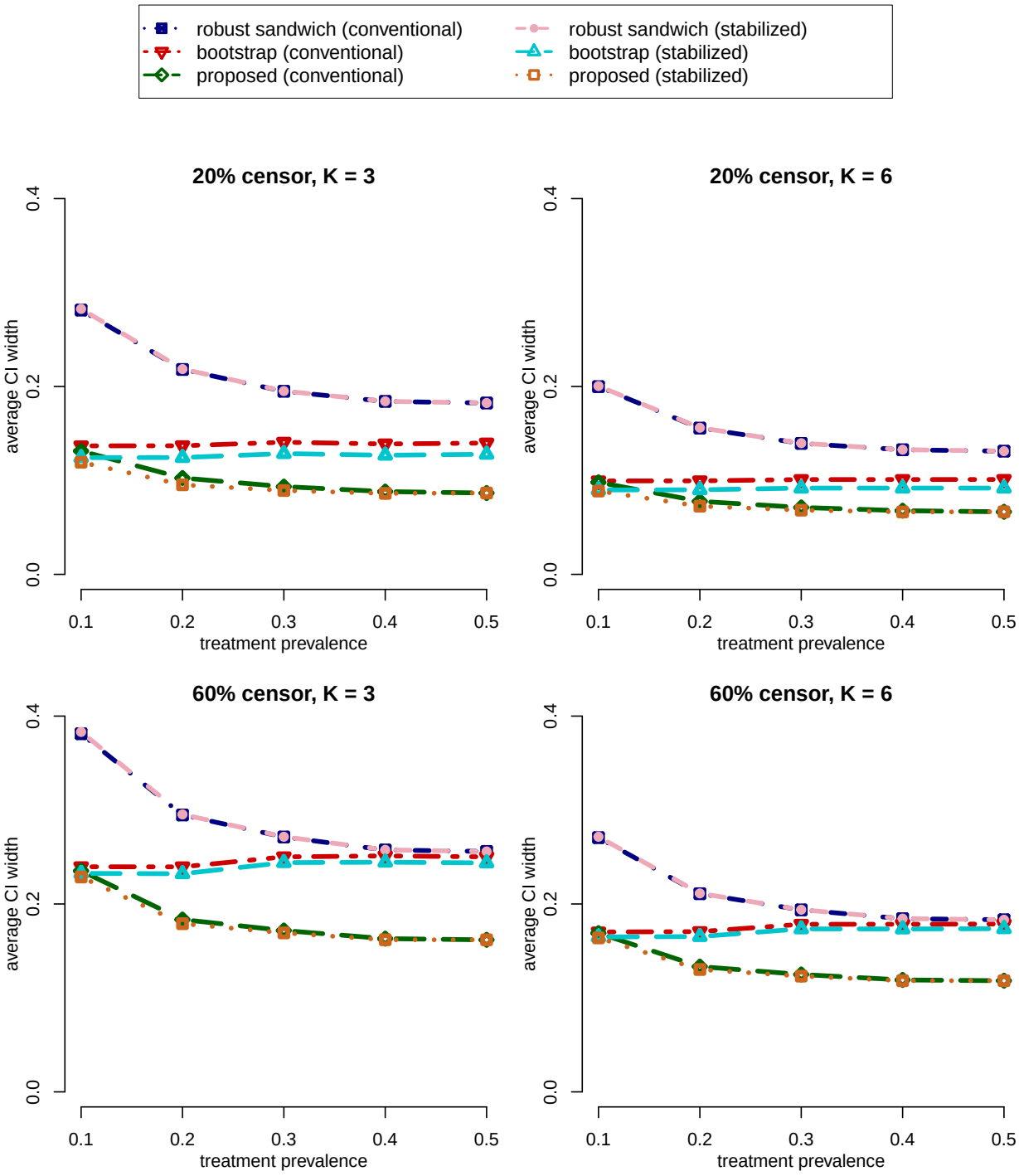

Figure S22: Average widths of 95% confidence intervals with $n = 800$ clusters each of size K.

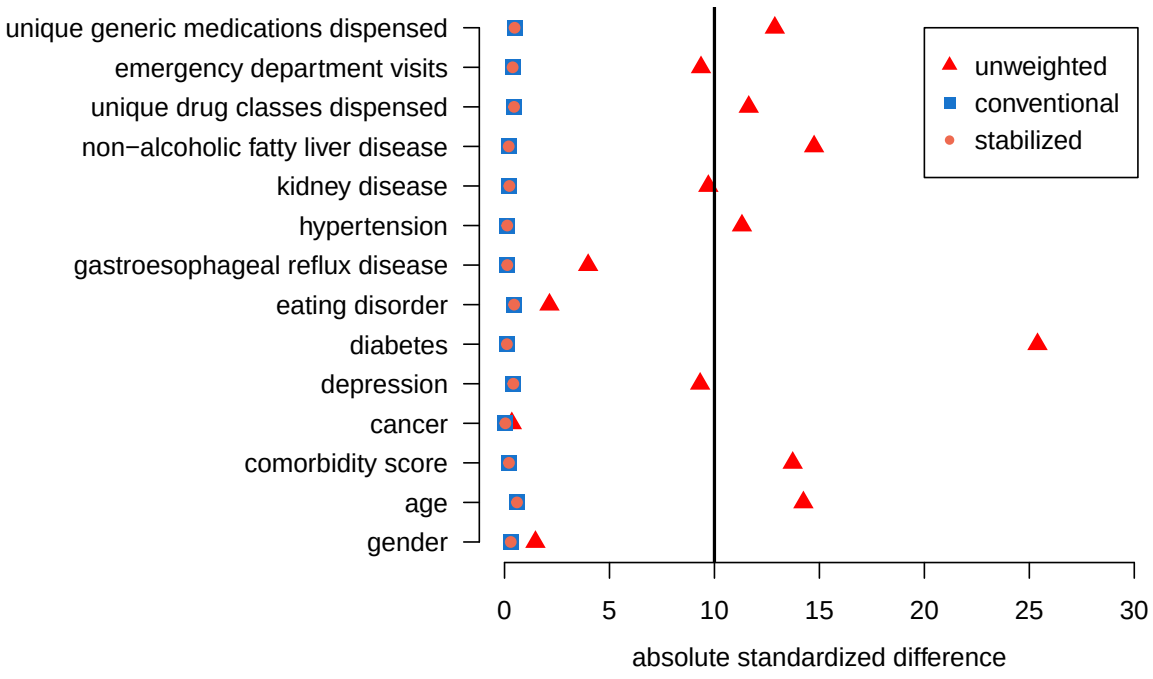

Figure S23: Absolute standardized differences of baseline covariates in unweighted and weighted samples (with conventional inverse probability weights and stabilized weights). The vertical line denotes an absolute standardized difference of 10%, considered by some authors as a threshold below which is indicative of negligible imbalance (Austin and Stuart, 2015).

References

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* **34,** 3661–3679.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79,** 139–147.

Lin, D. Y. and Wei, L.-J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84,** 1074–1078.