American Journal of Applied Mathematics
2022; 10(5): 176-204
http://www.sciencepublishinggroup.com/j/ajam
doi: 10.11648/j.ajam.20221005.11
ISSN: 2330-0043 (Print); ISSN: 2330-006X (Online)

Accounting for Propensity Score Variability in IPTW Weighted Cox Proportional Hazards Regression and Risk Estimation

Michael Richard Crager

Department of Biostatistics, Exact Sciences Corporation, Redwood City, United States

Email address:
mcrager@exactsciences.com

To cite this article:
Michael Richard Crager. Accounting for Propensity Score Variability in IPTW Weighted Cox Proportional Hazards Regression and Risk Estimation. *American Journal of Applied Mathematics*. Vol. 10, No. 5, 2022, pp. 176-204. doi: 10.11648/j.ajam.20221005.11

Received: September 6, 2022; **Accepted**: October 4, 2022; **Published**: October 17, 2022

Abstract: Under the assumption of no unmeasured confounders, Cox proportional hazards regression with inverse probability of treatment (IPTW) weighting based on propensity scores can be used to produce approximately unbiased estimates of treatment effect hazard ratios and event risks using observational cohorts. Often the weights are treated as fixed even though they are random variables, typically derived from a logistic regression analysis applied to the same cohort with treatment use as the outcome. Bootstrapping the entire process of weight-derivation, Cox regression analysis and estimation produces valid confidence intervals that account for the variability in the weights, but this method may be time- and resource-intensive for large cohorts. Here the delta method is used to derive large sample interval estimates of treatment effects and event risks that account for variability in the weights analytically. External time-dependent covariates, left truncation, and cohort sampling study designs are accommodated. Simulation studies show that this method provides confidence interval coverage probabilities at or above nominal level for small and moderate sample sizes. Stabilization of the weights by multiplying them by the overall treatment rate noticeably improves confidence interval coverage probabilities. Software to perform the calculations is freely available.

Keywords: Cox Regression, IPTW, Propensity Score, Risk Estimation, Variability

1. Introduction

Estimation of treatment effects and outcome probabilities from observational cohorts requires adjustment for confounding factors that may have influenced the selection of treatment and the outcome probabilities [1]. Failure to account for these confounding factors leads to biased estimates. When the outcome of interest is the time to a specified event and Cox proportional hazards regression is used for the analysis, inverse probability of treatment weighting (IPTW) based on the propensity for treatment provides approximately unbiased estimates of the average treatment effect in the population under the assumption of no unmeasured confounders [2]. The propensity scores for treatment assignment are typically estimated using a logistic regression analysis of the cohort used for the primary analysis. The resulting weights are often treated as fixed quantities and the variance of the regression parameter estimated using the robust method of Lin and Wei [3]. Using simulation studies, Austin [4] finds that this approach tends to over-estimate the variance of the treatment hazard ratio estimates, giving confidence intervals that have true coverage probability larger than nominal; Austin recommends bootstrapping the entire process of estimating the IPTW weights and using them in the Cox model. This produces valid confidence intervals, but the method may be time- or resource-intensive for large cohorts. An analytic method to account for the weight-estimation variance would therefore be a useful alternative and have the advantage of being fully reproducible.

Estimation of the risk of the event of interest depending on covariate values is often an objective when a proportional hazards regression model is fit. The estimated risk for a specified set of covariate values is a transform of the estimated cumulative hazard function, which depends directly on the weights as well as the number and timing of events and the estimated regression parameters.

Stratified cohort sampling designs are useful when studying relatively rare events and it is prohibitive to obtain the required data on all members of a cohort [5]. Typically, these designs include all cohort members who experienced an event in the sample along with a randomly selected fraction of the members who did not experience an event. The random selection may be stratified by one or more characteristics of the cohort members.

The Cox proportional hazards regression model accommodates time-dependent covariates. These can arise either because a covariate is measured repeatedly over time for each subject ("internal" time-dependence) or from assuming a fixed time-dependence of the effect of a covariate that is assessed at baseline ("external" time-dependence). Examples of external time-dependent covariates are (1) covariates with piece-wise constant hazard ratios, which are constant during specified time-intervals but vary across intervals, (2) covariates with time-dependence determined by a fixed function of time such as a natural cubic spline with parameters determined by baseline characteristics, and (3) current age of a study subject, defined as the age at study entry plus elapsed time since then. Hazard ratio estimates from the Cox model are valid and interpretable for either internal or external time-dependent covariates. The risk of a future event given the covariate values is well defined for external time-dependent covariates since the covariate path over time is determined by the baseline covariate values. For internal time-dependent covariates, however, risk estimation is complex, and requires assuming a distribution of the covariate value path [6-8]. The discussion here will restrict attention to external time-dependent covariates.

Left truncation of observations of a time to event occurs when study subjects enter the study at a time later than time 0. For example, if the time scale utilized is the subject's chronological age, then the data are left-truncated at the subject's age at study entry. Left truncation also occurs when events are structurally impossible during a period after study entry due to definitional constraints. One example is events not being counted during an initial treatment period that varies in duration. It is important to account for left truncation when it occurs to avoid bias in the analysis [9].

Hajage and colleagues [10] propose an analytic variance estimator for the treatment hazard ratio estimator in a Cox regression analysis using IPTW weighting based on propensity scores from a logistic regression. The variance estimator accounts for variability in the propensity score estimates as well as the variability Cox model parameter estimates using influence statistics based on linear approximations to the estimating equations. Simulations with large sample sizes (10,000 virtual subjects) indicate that confidence intervals for the treatment hazard ratio based on this variance estimator have true coverage probabilities that are very close to nominal levels and very similar to what the bootstrap produces. The development is restricted to Cox models with time-invariant covariates and does not accommodate cohort sampling designs or left truncation. Risk estimation is not discussed.

Recently, alternatives to IPTW weighting have been suggested with the goal of reducing the influence of observations with propensity scores close to 0 and the resulting inflation of variance of the estimates [11]. A leading example is the "overlap weight", equal to one minus the propensity score, that is, the probability of *not* receiving the treatment actually received [12, 13]. These weights can be applied to survival analysis [14, 15]. In contrast to IPTW weighting, which creates a simulated population in which the entire cohort has been randomized to treatment, the overlap weights down-weight individuals for which there is little or no overlap in treatments selected, thus creating a simulated population that emphasizes individuals with characteristics associated with clinical equipoise in treatment selection [16]. However, IPTW weighting is still used much more frequently than overlap weights or similar methods [11], perhaps due to the clear interpretation of the population of inference associated with IPTW.

Here the delta method and variance estimation based on "dfbeta" influence statistics are used to derive the asymptotic variance estimates and confidence intervals for hazard ratios and event risk using Cox proportional hazards regression models with IPTW weighting, accounting for the variation in the propensity score-based weight estimates. It is assumed that the propensity scores are derived from a logistic regression model applied to the same cohort. The development accommodates multiple treatments, time-invariant or external time-dependent covariates, stratified cohort sampling study designs, and left truncation of the observations.

2. Methods

Consider an observational cohort including multiple treatments or other interventions that were decided upon, not randomized, and a time-to-event outcome variable. Further suppose it is desired to estimate the risk of an event occurring by a specified time using a Cox proportional hazards regression model with propensity score-based weights estimated from the cohort using a logistic regression model. Denote the regression parameter estimate for the Cox model with covariate vector $\mathrm{z}^{(\beta)}$ by $\hat{\beta}$. The vector $\mathrm{z}^{(\beta)}$ may be time-invariant or externally time-dependent $\mathrm{z}^{(\beta)} = \mathrm{z}^{(\beta)}(t)$, meaning that the time-dependence is fixed rather than deriving from repeated assessments of a subject's covariate values over time. According to the model, the hazard for the event of interest at time t is $\lambda(t) = \lambda_0(t)\exp\left(\mathrm{z}^{(\beta)}(t)\right)$, where $\lambda_0(t)$ is the baseline hazard function.

In conventional inverse-probability-of-treatment weighting (IPTW), the weight for subject $i = 1, 2, \ldots, n$ is the inverse of the probability of (propensity for) use of the treatment received by subject i. A multinomial logistic regression with generalized logit link function can be used to accommodate propensity estimation for $K \geq 2$ treatments. Using treatment K as the reference, let $\hat{\alpha} = (\hat{\alpha}_1^{\mathrm{T}}, \hat{\alpha}_2^{\mathrm{T}}, \ldots, \hat{\alpha}_{K-1}^{\mathrm{T}})^{\mathrm{T}}$ be the vector containing all the maximum likelihood logistic regression parameter estimators, $\hat{\alpha}_k$ being the vector of regression parameter estimators for treatment k. The probability of receiving treatment k when the covariate value is $\mathrm{z}^{(\alpha)}$ is estimated consistently by

$$\hat{p}_k(\mathrm{z}^{(\alpha)}) = \begin{cases} \dfrac{\exp\left(\hat{\alpha}_k^{\mathrm{T}}\mathrm{z}^{(\alpha)}\right)}{1+\sum_{j=1}^{K-1}\exp\left(\hat{\alpha}_j^{\mathrm{T}}\mathrm{z}^{(\alpha)}\right)} & \text{for } k < K \\ \dfrac{1}{1+\sum_{j=1}^{K-1}\exp\left(\hat{\alpha}_j^{\mathrm{T}}\mathrm{z}^{(\alpha)}\right)} & \text{for } k = K. \end{cases}$$

Let k_i represent the treatment received by study subject i and $\mathrm{z}_i^{(\alpha)}$ be the subject's covariate vector value (including the intercept). The estimated weight for subject i is $\hat{w}_i = 1/\hat{p}_{k_i}(\mathrm{z}_i^{(\alpha)})$.

Use of these weights allows the practitioner to estimate the population average treatment effect (ATE) and the average event risk. The estimates will be approximately unbiased under the assumption that there are no unmeasured confounders [1, 2].

Stabilization of the IPTW weights by multiplying them by the proportion π_{k_i} of study subjects who received the same treatment that subject i received is often recommended in order to avoid very high weights and resulting inflation of the variance of model parameter estimates [17, 18]. The stabilized weight used in the Cox regression is $\hat{\omega}_i = \pi_{k_i}\hat{w}_i$. If the data are from a stratified cohort sampling design with sampling weight s_i (the inverse of the stratum-specific sampling ratio) for subject i, then the calculation of the π_{k_i} should reflect the sampling weights and the weight used in the Cox regression is $\hat{\omega}_i = s_i\pi_{k_i}\hat{w}_i$. This form of the weights will be used in the following development. If cohort sampling is not used, set $s_i \equiv 1$. If the propensity weights are not stabilized, set $\pi_{k_i} \equiv 1$.

For subjects with $k_i \le K-1$, the gradient of $\hat{\omega}_i$ with respect to $\hat{\alpha}$ is

$$\begin{aligned} \nabla_{\hat{\alpha}}\hat{\omega}_i &= s_i\pi_{k_i}\nabla_{\hat{\alpha}}\frac{1+\sum_{j=1}^{K-1}\exp\left(\hat{\alpha}_j^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)}{\exp\left(\hat{\alpha}_{k_i}^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)} \\ &= s_i\pi_{k_i}\exp\left(-2\hat{\alpha}_{k_i}^{\mathrm{T}}\mathrm{z}^{(\alpha)}\right)\left[\exp\left(\hat{\alpha}_{k_i}^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\nabla_{\hat{\alpha}}\sum_{j=1}^{K-1}\exp\left(\hat{\alpha}_j^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)-\left\{1+\sum_{j=1}^{K-1}\exp\left(\hat{\alpha}_j^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\right\}\nabla_{\hat{\alpha}}\exp\left(\hat{\alpha}_{k_i}^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\right] \\ &= s_i\pi_{k_i}\exp\left(-\hat{\alpha}_{k_i}^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\left[\left(\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_1^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right),\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_2^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right),\ldots,\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_{K-1}^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\right)^{\mathrm{T}}\right. \\ &\qquad \left.-\left\{1+\sum_{j=1}^{K-1}\exp\left(\hat{\alpha}_j^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\right\}\left(I_{\{k_i=1\}}\mathrm{z}_i^{(\alpha)\mathrm{T}},I_{\{k_i=2\}}\mathrm{z}_i^{(\alpha)\mathrm{T}},\ldots,I_{\{k_i=K-1\}}\mathrm{z}_i^{(\alpha)\mathrm{T}}\right)^{\mathrm{T}}\right], \end{aligned}$$

which, upon recognition of the form of $\hat{p}_{k_i}(\mathrm{z}_i^{(\alpha)})$, gives for $k_i < K$

$$\begin{aligned} \nabla_{\hat{\alpha}}\hat{\omega}_i &= s_i\pi_{k_i}\exp\left(-\hat{\alpha}_{k_i}^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\left(\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_1^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right),\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_2^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right),\ldots,\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_{K-1}^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\right)^{\mathrm{T}} \\ &\quad -s_i\pi_{k_i}\hat{p}_{k_i}^{-1}(\mathrm{z}_i^{(\alpha)})\left(I_{\{k_i=1\}}\mathrm{z}_i^{(\alpha)\mathrm{T}},I_{\{k_i=2\}}\mathrm{z}_i^{(\alpha)\mathrm{T}},\ldots,I_{\{k_i=K-1\}}\mathrm{z}_i^{(\alpha)\mathrm{T}}\right)^{\mathrm{T}}. \end{aligned} \tag{1}$$

For subjects with $k_i = K$,

$$\begin{aligned} \nabla_{\hat{\alpha}}\hat{\omega}_i &= s_i\pi_{k_i}\nabla_{\hat{\alpha}}\left\{1+\sum_{j=1}^{K-1}\exp\left(\hat{\alpha}_j^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\right\} \\ &= s_i\pi_{k_i}\left(\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_1^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right),\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_2^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right),\ldots,\mathrm{z}_i^{(\alpha)\mathrm{T}}\exp\left(\hat{\alpha}_{K-1}^{\mathrm{T}}\mathrm{z}_i^{(\alpha)}\right)\right)^{\mathrm{T}}. \end{aligned} \tag{2}$$

Pugh and colleagues developed an estimator for the covariance matrix of the Cox model regression parameters when a weighted analysis is used to adjust the regression parameter estimates for missing covariate values [19]. To make the adjustment, a logistic regression analysis is used to estimate the probability of missingness and a weight equal to the inverse of the estimated probability for each subject is used in the Cox regression. The estimated covariance matrix of the Cox model parameter accounts for the variability in the logistic regression-based weight estimates. This situation, in which the weights derive from the probability of missingness, is completely analogous to the use of IPTW weights that derive from the probability of receiving treatment. Using the form of Pugh's estimator given by Therneau and Grambsch [20], page 166, the covariance matrix of $\hat{\beta}$ is estimated consistently, accounting for the variance in both the Cox regression and the logistic regression for the propensity score-based weights, by

$$\hat{\mathrm{V}}_{\hat{\beta}} = \mathrm{D}_{\hat{\beta}}^{\mathrm{T}}\left(\mathrm{I}-\mathrm{D}_{\hat{\alpha}}\left(\mathrm{D}_{\hat{\alpha}}^{\mathrm{T}}\mathrm{D}_{\hat{\alpha}}\right)^{-1}\mathrm{D}_{\hat{\alpha}}^{\mathrm{T}}\right)\mathrm{D}_{\hat{\beta}},$$

where $\mathrm{D}_{\hat{\beta}}$ is the matrix of dfbetas for $\hat{\beta}$ and $\mathrm{D}_{\hat{\alpha}}$ is the matrix of dfbetas for $\hat{\alpha}$ from the logistic regression model used to estimate the probability of treatment assignment. The ith row of the dfbeta matrix closely approximates the change in the regression parameter estimate vector that would result from deleting subject i from the analysis set. Note that $\hat{\mathrm{V}}_{\hat\beta}$ is the covariance matrix of the residuals of the linear regression of the $\hat\beta$ dfbeta on the $\hat\alpha$ dfbeta, so the elements of $\hat\beta$ estimated from the Cox regression model using the weights are asymptotically uncorrelated with the elements of $\hat\alpha$.

The dfbetas for the Cox model can be computed from the score residuals and the Fisher information matrix. The score residual for subject i is

$$\mathrm{u}_i = \int_0^\infty \left\{ \mathrm{z}_i^{(\beta)}(t) - \overline{\mathrm{z}^{(\beta)}(t)} \right\} \mathrm{d}\hat{M}_i(t),$$

where

$$\overline{\mathrm{z}^{(\beta)}}(t) = \frac{\sum_{i=1}^{n} \hat\omega_i Y_i(t) \exp\left(\hat\beta^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(t)\right) \mathrm{z}_i^{(\beta)}(t)}{\sum_{i=1}^{n} \hat\omega_i Y_i(t) \exp\left(\hat\beta^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(t)\right)}$$

and where

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) \exp\left(\hat\beta^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(s)\right) \mathrm{d}\hat\Lambda_0(s)$$

is the martingale residual process [20]. Here $N_i(t)$ is the event-counting process for subject i, $Y_i(t)$ is the indicator for whether subject i is in the risk set (having entered the study, accounting for left truncation, if any, and still being followed) at time t, and

$$\hat\Lambda_0(t) = \int_0^t \frac{\sum_{i=1}^{n} \hat\omega_i \,\mathrm{d} N_i(s)}{\sum_{i=1}^{n} \hat\omega_i Y_i(s) \exp\left(\hat\beta^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(s)\right)}$$

is the baseline cumulative hazard estimator. The ith row of the dfbeta matrix $\mathrm{D}_{\hat\beta}$ is $\hat\omega_i \mathrm{u}_i^{\mathrm{T}} \mathrm{I}^{-1}(\hat\beta)$, where $\mathrm{I}(\hat\beta)$ is the Fisher information matrix evaluated at the maximum partial likelihood estimate. $\mathrm{I}(\hat\beta)$ is the matrix of the negatives of the second derivatives of the log partial likelihood with respect to the regression parameter estimates. Its inverse is output by standard proportional hazard regression packages as the (naïve) model-based estimate of the covariance matrix of $\hat\beta$.

Using results from Czepiel [21] (see equation 32), the ith row of the dfbeta matrix D_{α} for the multinomial logistic regression model is given by

$$\left(\left\{ I_{\{k_i=1\}} - \hat{p}_1(\mathrm{z}_i^{(\alpha)}) \right\} \mathrm{z}_i^{(\alpha)\mathrm{T}}, \left\{ I_{\{k_i=2\}} - \hat{p}_2(\mathrm{z}_i^{(\alpha)}) \right\} \mathrm{z}_i^{(\alpha)\mathrm{T}}, \ldots, \left\{ I_{\{k_i=K-1\}} - \hat{p}_{K-1}(\mathrm{z}_i^{(\alpha)}) \right\} \mathrm{z}_i^{(\alpha)\mathrm{T}} \right) \hat{\mathrm{V}}_{\hat\alpha},$$

where $I_{\{k_i=k\}}$ is the indicator for whether subject i received treatment k and $\hat{\mathrm{V}}_{\hat\alpha}$ is the estimated covariance matrix of the maximum likelihood logistic regression parameter estimator $\hat\alpha$, that is, the inverse of the Fisher information matrix. With stratified cohort sampling weighting, in which subject i is weighted by s_i, the inverse of the sampling fraction for that subject's stratum, the ith row becomes

$$s_i \left(\left\{ I_{\{k_i=1\}} - \hat{p}_1(\mathrm{z}_i^{(\alpha)}) \right\} \mathrm{z}_i^{(\alpha)\mathrm{T}}, \left\{ I_{\{k_i=2\}} - \hat{p}_2(\mathrm{z}_i^{(\alpha)}) \right\} \mathrm{z}_i^{(\alpha)\mathrm{T}}, \ldots, \left\{ I_{\{k_i=K-1\}} - \hat{p}_{K-1}(\mathrm{z}_i^{(\alpha)}) \right\} \mathrm{z}_i^{(\alpha)\mathrm{T}} \right) \hat{\mathrm{V}}_{\hat\alpha},$$

where $\hat{\mathrm{V}}_{\hat\alpha}$ is now the (naïve) model-based estimate of the covariance matrix of $\hat\alpha$ in the logistic regression using the cohort sampling weights.

The estimated cumulative hazard at time t for a patient with covariate vector $\mathrm{z}^{(\beta)}$ is

$$\hat\Lambda\left(t; \mathrm{z}^{(\beta)}\right) = \int_0^t \exp\left(\hat\beta^{\mathrm{T}} \mathrm{z}^{(\beta)}(s)\right) \mathrm{d}\hat\Lambda_0(s).$$

The gradient of the increment in the baseline hazard at time t estimator with respect to the Cox regression parameter estimate vector is

$$\nabla_{\hat\beta}\, \mathrm{d}\hat\Lambda_0(t) = -\overline{\mathrm{z}^{(\beta)}}(t)\, \mathrm{d}\hat\Lambda_0(t).$$

Thus, since the Stieltjes integral represents a finite sum,

$$\begin{aligned} \nabla_{\hat\beta} \hat\Lambda\left(t; \mathrm{z}^{(\beta)}\right) &= \int_0^t \left\{ \nabla_{\hat\beta} \exp\left(\hat\beta^{\mathrm{T}} \mathrm{z}^{(\beta)}(s)\right) \right\} \mathrm{d}\hat\Lambda_0(s) + \int_0^t \exp\left(\hat\beta^{\mathrm{T}} \mathrm{z}^{(\beta)}(s)\right) \nabla_{\hat\beta}\, \mathrm{d}\hat\Lambda_0(s) \\ &= \int_0^t \left\{ \mathrm{z}^{(\beta)}(s) - \overline{\mathrm{z}^{(\beta)}}(s) \right\} \exp\left(\hat\beta^{\mathrm{T}} \mathrm{z}^{(\beta)}(s)\right) \mathrm{d}\hat\Lambda_0(s). \end{aligned}$$

The gradient of the increment in baseline hazard estimator at time t with respect to the logistic regression parameter estimate vector $\hat\alpha$ is

$$\nabla_{\hat{\alpha}} \,\mathrm{d}\,\hat{\Lambda}_0(t) = \nabla_{\hat{\alpha}} \left\{ \frac{\sum_{i=1}^{n} \hat{\omega}_i \,\mathrm{d}\, N_i(t)}{\sum_{i=1}^{n} \hat{\omega}_i Y_i(t) \exp\left(\hat{\beta}^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(t)\right)} \right\}$$

$$= \frac{\sum_{i=1}^{n} \left(\nabla_{\hat{\alpha}} \hat{\omega}_i\right) \mathrm{d}\, N_i(t)}{\sum_{i=1}^{n} \hat{\omega}_i Y_i(t) \exp\left(\hat{\beta}^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(t)\right)} - \left\{\sum_{i=1}^{n} \hat{\omega}_i \,\mathrm{d}\, N_i(t)\right\} \left\{\sum_{i=1}^{n} \left(\nabla_{\hat{\alpha}} \hat{\omega}_i\right) Y_i(t) \exp\left(\hat{\beta}^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(t)\right)\right\} \left\{\sum_{i=1}^{n} \hat{\omega}_i Y_i(t) \exp\left(\hat{\beta}^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(t)\right)\right\}^{-2},$$

where $\nabla_{\hat{\alpha}} \hat{\omega}_i$ is given in equations (1) and (2). This permits the computation of

$$\nabla_{\hat{\alpha}} \hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right) = \int_0^t \exp\left(\hat{\beta}^{\mathrm{T}} \mathrm{z}^{(\beta)}(s)\right) \nabla_{\hat{\alpha}} \,\mathrm{d}\,\hat{\Lambda}_0(s).$$

It is also necessary to account for the variability in the number and timing of events (the jumps in the event counting process). Let T_i be the end of follow-up for subject i, at which time either an event occurred $\left(\mathrm{d}\, N_i(T_i) = 1\right)$ or the subject's time to event was censored $\left(\mathrm{d}\, N_i(T_i) = 0\right)$. Then

$$\frac{\partial \hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)}{\partial \,\mathrm{d}\, N_i(T_i)} = I_{\{T_i \le t\}} \exp\left\{\hat{\beta}^{\mathrm{T}} \mathrm{z}^{(\beta)}(T_i)\right\} \mathrm{d}\,\hat{\Lambda}_0(T_i).$$

We saw previously that the elements of $\hat{\beta}$ are asymptotically uncorrelated with the elements of $\hat{\alpha}$. It is also known that $\hat{\beta}$ is asymptotically uncorrelated with the number and timing of events [22]. Therefore, the variance of $\hat{\Lambda}(t; \mathrm{z})$ can be consistently estimated by

$$\widehat{\mathrm{Var}}\left\{\hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)\right\} = \left\{\nabla_{\hat{\beta}} \hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)\right\}^{\mathrm{T}} \hat{\mathrm{V}}_{\hat{\beta}} \left\{\nabla_{\hat{\beta}} \hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)\right\} + \left\{\nabla_{\hat{\alpha}} \hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)\right\}^{\mathrm{T}} \hat{\mathrm{V}}_{\hat{\alpha}} \left\{\nabla_{\hat{\alpha}} \hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)\right\} + \sum_{i=1}^{n} \left\{\frac{\partial \hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)}{\partial \,\mathrm{d}\, N_i(T_i)}\right\}^2 \mathrm{d}\, N_i(T_i).$$

Using the delta method, the variance of the log cumulative hazard estimator $\hat{\rho}(t; \mathrm{z}) = \ln \hat{\Lambda}(t; \mathrm{z})$ is estimated consistently by

$$\widehat{\mathrm{Var}}\left\{\hat{\rho}(t; \mathrm{z}^{(\beta)})\right\} = \frac{\widehat{\mathrm{Var}}\left\{\hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)\right\}}{\left\{\hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)\right\}^2}.$$

Transforming this estimator, the event risk at time *t* is estimated consistently by

$$\begin{aligned} \hat{r}\left(t; \mathrm{z}^{(\beta)}\right) &= 1 - \exp\left(-\hat{\Lambda}\left(t; \mathrm{z}^{(\beta)}\right)\right) \\ &= 1 - \exp\left[-\exp\left\{\hat{\rho}\left(t; \mathrm{z}^{(\beta)}\right)\right\}\right], \end{aligned}$$

and a level $1-\alpha$ confidence interval for the risk at time *t* has endpoints

$$1 - \exp\left[-\exp\left\{\hat{\rho}\left(t; \mathrm{z}^{(\beta)}\right) \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \widehat{\mathrm{Var}}\left(\hat{\rho}\left(t; \mathrm{z}^{(\beta)}\right)\right)\right\}\right],$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution.

In some situations, it may be appropriate to fit a Cox proportional hazards regression allowing a distinct baseline hazard function in each stratum of the population. If S_k is the subset of patients in stratum k, the baseline cumulative hazard function estimate for stratum k is

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^{n} I_{\{i \in S_k\}} \hat{\omega}_i \,\mathrm{d}\, N_i(s)}{\sum_{i=1}^{n} I_{\{i \in S_k\}} \hat{\omega}_i Y_i(s) \exp\left(\hat{\beta}^{\mathrm{T}} \mathrm{z}_i^{(\beta)}(s)\right)}.$$

3. Simulation Studies

The confidence intervals developed above are valid for large samples. To evaluate their performance in small and moderate samples, data sets were simulated with three treatments, a normally distributed score S and 5 other covariates $z_1, z_2, \ldots, z_5$. The probability of treatment was related to S with odds ratios 0.95 (treatment 1 vs. 3) and 1.05 (treatment 2 vs. 3) and the 5 additional covariates with log odds ratios $\gamma_i^{(1vs.3)} = (-1)^{i+1}\ln(0.9-(i-3)/40)$ and $\gamma_i^{(1vs.3)} = (-1)^{i}\ln(0.9-(i-3)/40)$. These parameters resulted in approximately 20% of simulated subjects with treatment 1, 20% with treatment 2 and 60% with treatment 3.

In relation to the event risk, treatments 1 and 2 were given hazard ratios of 0.5 and 0.75 relative to treatment 3. The score S was associated with the risk of an event with a standardized hazard ratio of 2. The other covariates were given hazard ratios $\beta_{z_1} = \ln 1.5$, $\beta_{z_2} = \ln 1.4$, $\beta_{z_3} = \ln 1.2$, $\beta_{z_4} = \ln 1.05$ and $\beta_{z_5} = \ln 1.06$.

The covariate values for S and $z_1, z_2, \ldots, z_5$ were simulated using a multivariate normal distribution with correlations of 0.5 between S and z_1, 0.6 between z_2 and z_3, 0.7 between z_4 and z_5, and 0 between the other covariate pairs.

Times to event were generated using the exponential distribution with hazard rate determined by the proportional hazards model. Random censoring was simulated using an independent exponential distribution, targeting a censoring rate of 75%.

Data sets were simulated having specified numbers of events ranging from 40 to 160.

For each data set, a multinomial logistic regression model with generalized logit link function was fit using the covariates $z_1, z_2, \ldots, z_5$, and the propensity score for each subject was calculated from this model. Weighted Cox models were then fit (1) using effects for treatment, score and z_1, and (2) using effects for treatment, score and $z_1, z_2, \ldots, z_5$. For each model, 95% confidence intervals were computed for the hazard ratios for the score and each treatment. Confidence intervals were also computed for the event risk under scenarios reflecting the use of each treatment and a score value of -2, -1, 0, 1 or 2, with the other covariates fixed at 0. For each scenario, the event risk was estimated at a fixed time at approximately the 75th percentile of the time-to-event in the population. The entire simulation process was replicated 1000 times. True coverage probabilities were estimated for each of the confidence intervals as the proportion of intervals calculated from the simulated data sets that included the true value. The 1000 replicates give a precision of ±1.3% (half-width of a 95% confidence interval) for true coverage probabilities near the nominal level of 95%. Coverage probabilities were assessed using (1) conventional IPTW weights, (2) stabilized weights, (3) stabilized weights truncated at the 5th and 95th percentiles. For comparison, risk and hazard ratio estimates were computed without accounting for the variation in weights (treating the weights as fixed) and the geometric means of the ratio of the widths of confidence intervals with and without accounting for variation in the weights was computed.

The weighted Cox regression analysis estimates the ATE, which is the marginal hazard ratio for treatment in the population, and, similarly, the marginal event risk in the population. Accordingly, the true marginal values were generated using methods similar to those described by Austin [4]: a very large sample of one million subjects was simulated using the methods described above and for each simulated subject, three independent time-to-event observations were generated, one for each treatment. A Cox proportional hazards regression model was applied to the resulting three million-subject simulated data set, and the resulting estimated hazard ratios and event risks were used as the true values.

Example true coverage probabilities for nominal 95% confidence intervals for risk using the stabilized weights are shown in Figure 1 (40 events), Figure 2 (80 events) and Figure 3 (160 events). In each of these simulations, the Cox model included terms for treatment indicators (1 vs. 3 and 2 vs. 3), score and z_1. Covariates z_2, z_3, z_4 and z_5 were omitted from the Cox model but included in the logistic regression model for treatment propensity. True coverage probability estimates are shown for the method accounting for variability in the weights and for the method treating the weights as fixed. With a small data set of 40 events, the bare minimum necessary to meet the threshold of 10 events per covariate [23], the method accounting for weight variability gives coverage probabilities at or above the nominal level of 95% whereas the method treating the weights as fixed generally gives substantially sub-nominal coverage levels. With 80 events, accounting for weight variability gives intervals at or above nominal level; confidence intervals treating weights as fixed have true coverage probabilities closer to nominal level than with 40 events but still noticeably anti-conservative in some cases. With 160 events, accounting for the weight variability gives intervals with true coverage levels higher than nominal (approximately 97%) and treating the weights as fixed gives coverage probabilities at the nominal level. Accounting for the variability in the weights, true coverage probabilities for the marginal hazard ratios for the score and treatment were at or above nominal levels except for a few instances with very small samples, whereas treating the weights as fixed gave true coverage probabilities that were noticeably below nominal level even with larger samples (Figure 4).

True coverage probabilities from simulations including treatment, the score and all five prognostic variables $z_1, z_2, \ldots, z_5$ in the Cox model and including the score and $z_1, z_2, \ldots, z_5$ in the logistic regression model are shown in Figure 5 (80 events, the minimum necessary to attain 10 events per covariate), Figure 6 (120 events) and Figure 7 (160 events). Here the coverage probabilities tend to be somewhat conservative (higher than nominal) for the intervals accounting for the variability in the weights and somewhat anti-conservative (lower than nominal) for the intervals treating the weights as fixed. Coverage probabilities for confidence intervals for hazard ratios for this model are shown in Figure 8. The method accounting for variability in the weights generally had coverage probabilities at or above nominal level except for one treatment, which had anti-conservative coverage probabilities for most sample sizes. Coverage probabilities for confidence intervals not accounting for weight variability were generally anti-conservative.

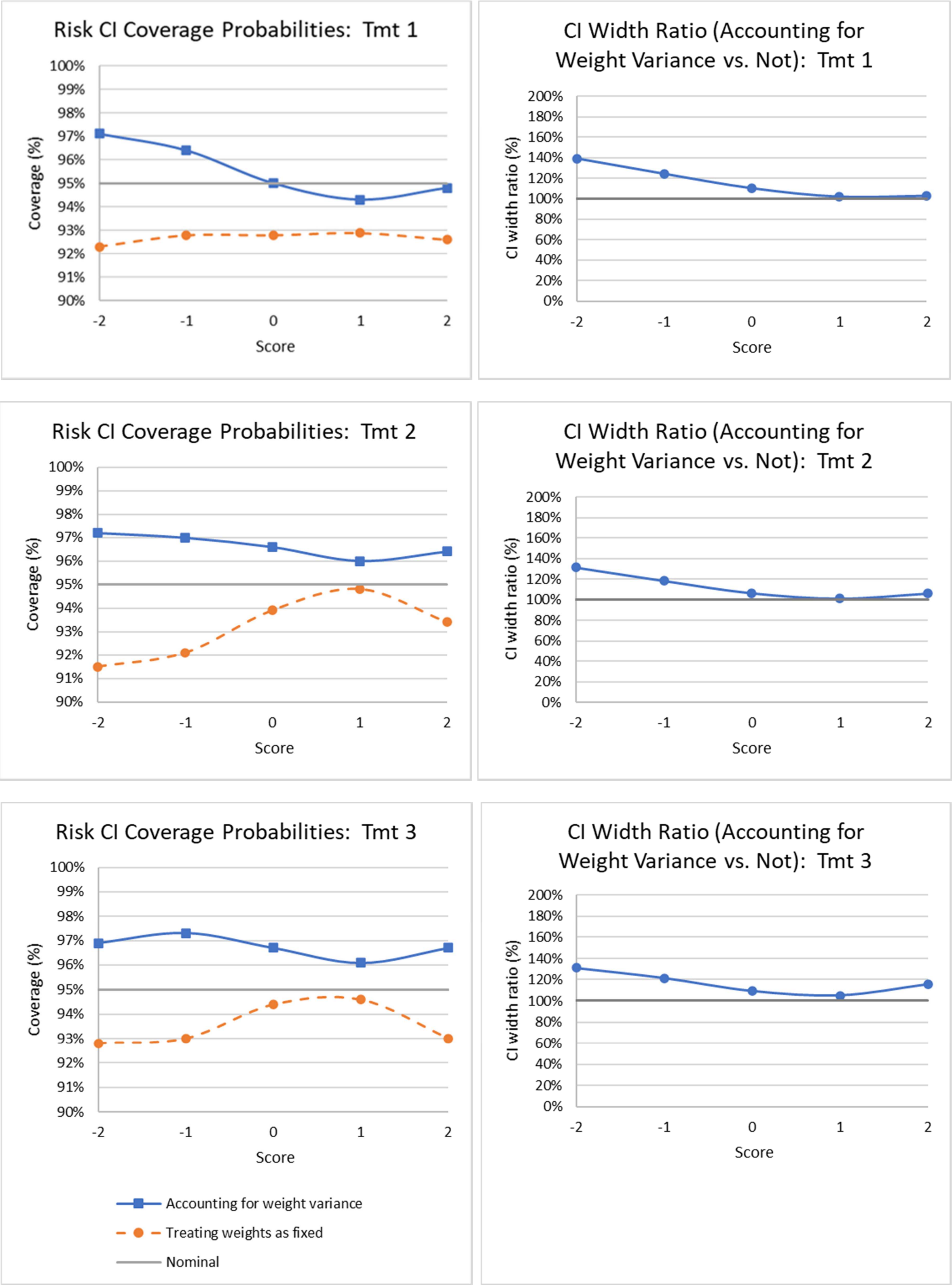


Figure 1. *Simulation results for 40 events. Cox model with effects for treatment indicators, score S and covariate z_1. Propensity scores from multinomial logistic regression model with effects for $z_1, z_2, \ldots, z_5$. Stabilized weights. Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

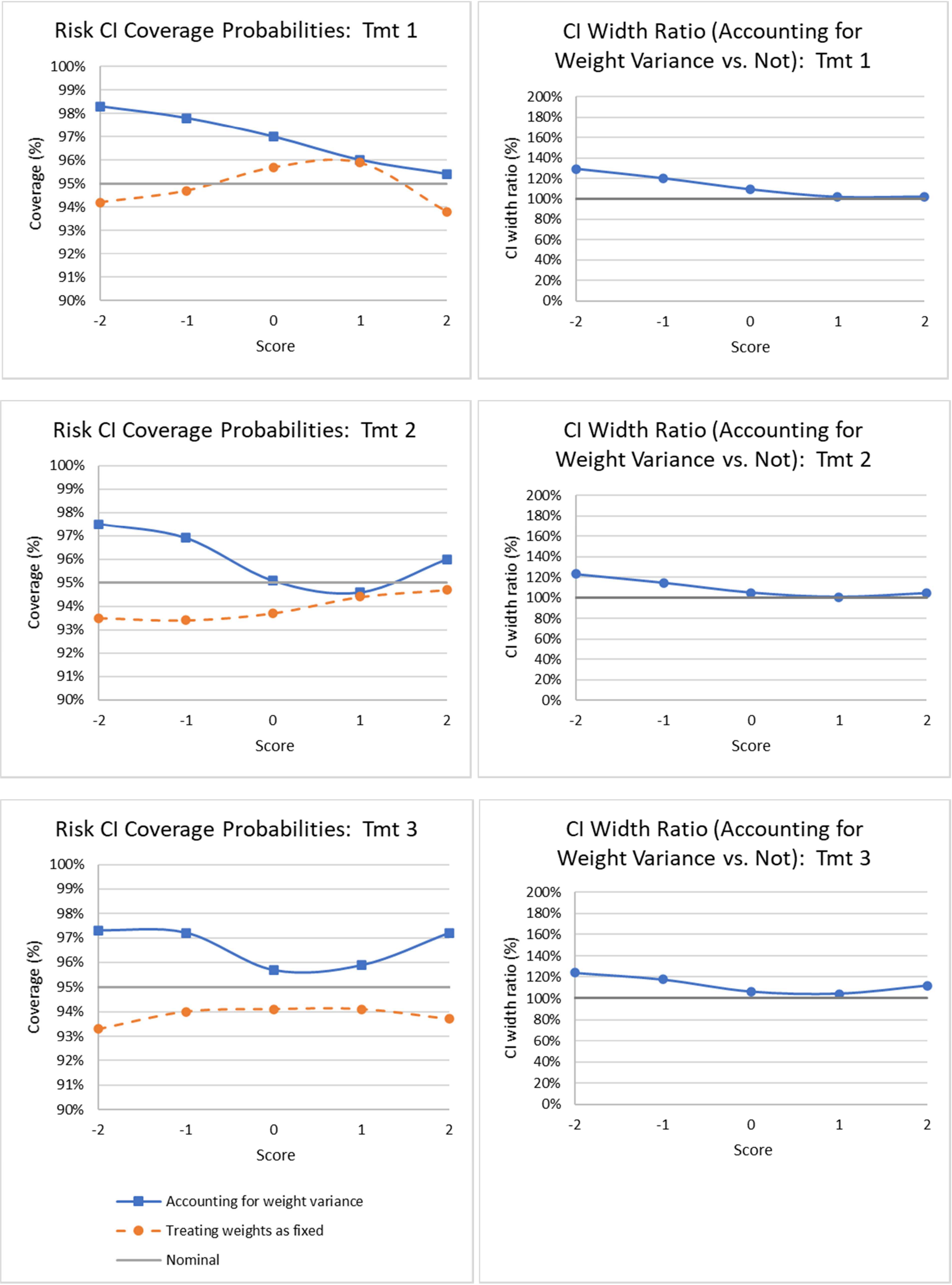


Figure 2. *Simulation results for 80 events. Cox model with effects for treatment indicators, score S and covariate z_1. Propensity scores from multinomial logistic regression model with effects for $z_1, z_2, \ldots, z_5$. Stabilized weights. Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

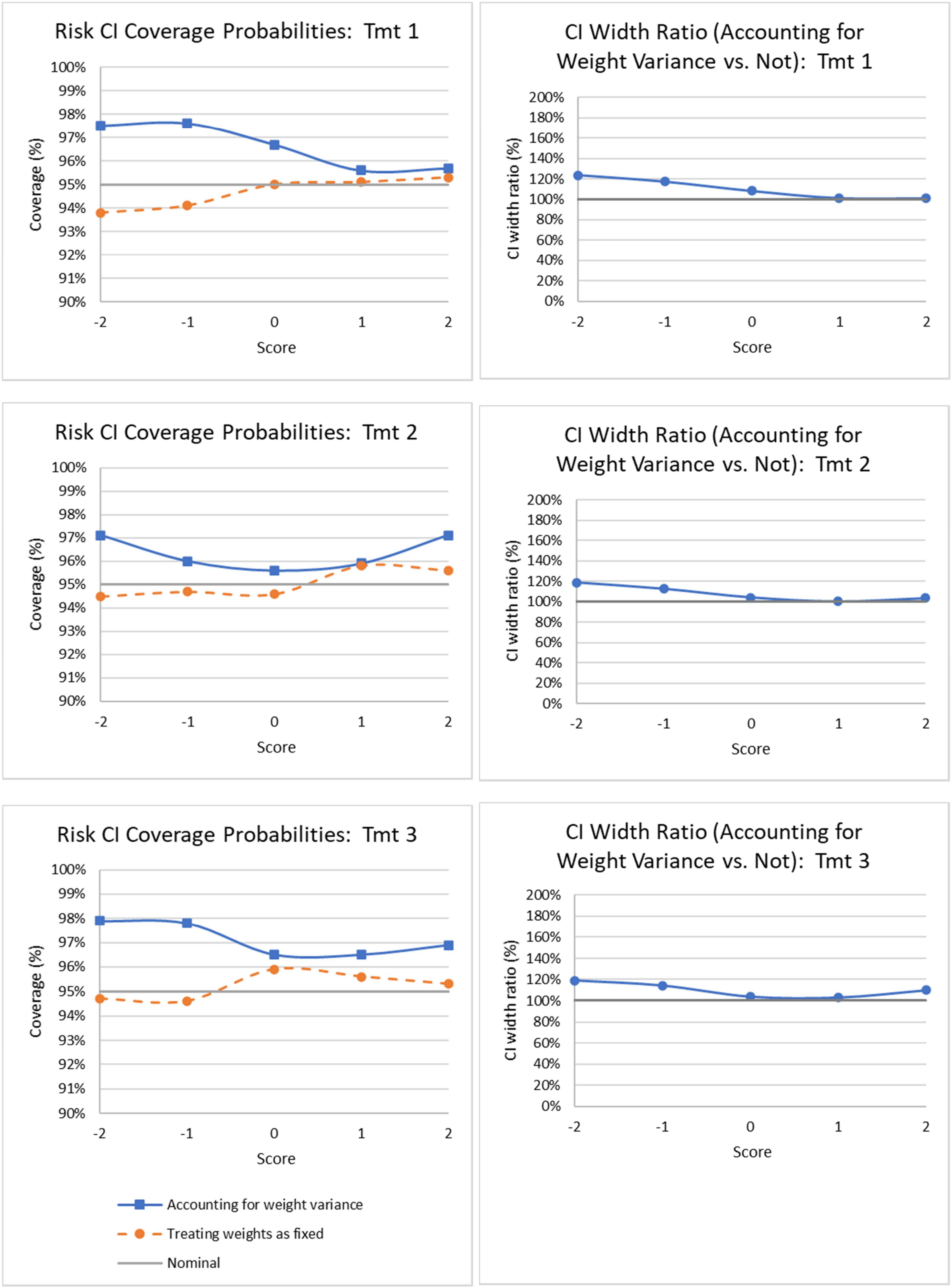


Figure 3. *Simulation results for 160 events. Cox model with effects for treatment indicators, score S and covariate z_1. Propensity scores from multinomial logistic regression model with effects for $z_1, z_2, \ldots, z_5$. Stabilized weights. Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

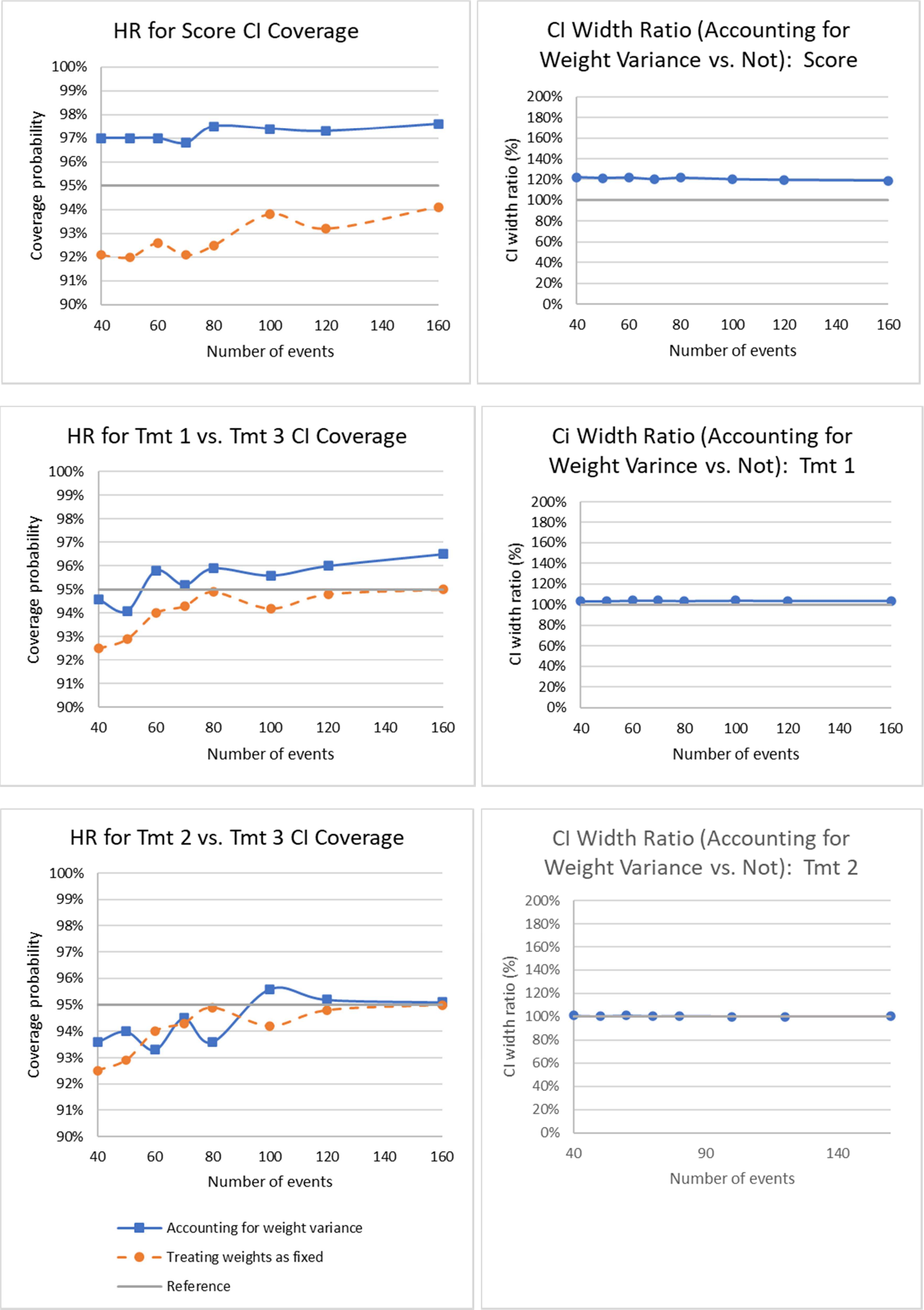


Figure 4. *Simulation results for hazard ratio CI coverage probabilities. Cox model with effects for treatment indicators, score S and z_1. Propensity scores from multinomial logistic regression model with effects for $z_1, z_2, \ldots, z_5$. . Stabilized weights.*

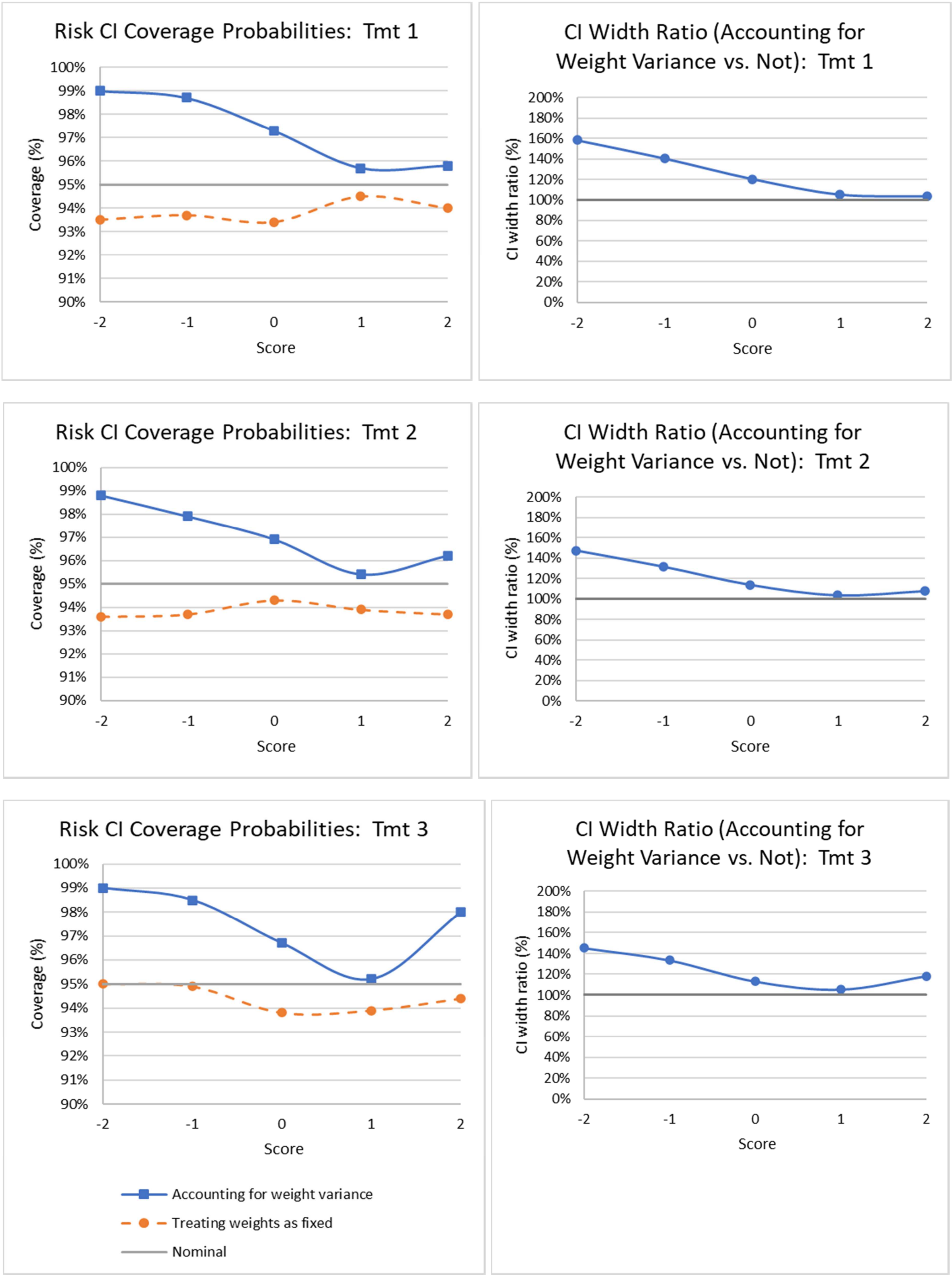
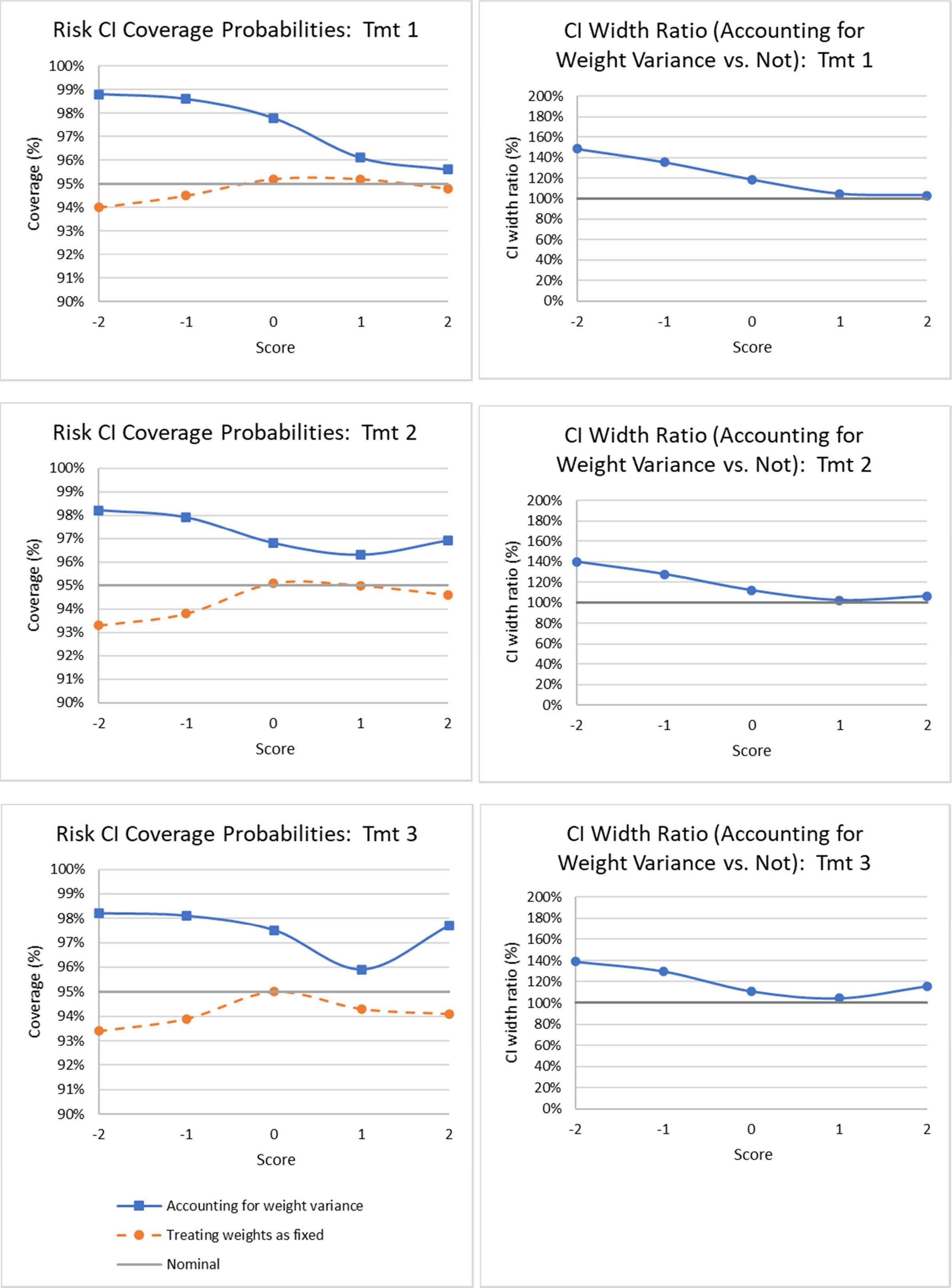


Figure 5. *Simulation results for 80 events. Cox model with effects for treatment indicators, score S and covariates $z_1, z_2, \ldots, z_5$. Propensity scores from multinomial logistic regression model with effects for $z_1, z_2, \ldots, z_5$. Stabilized weights. Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*



Risk CI Coverage Probabilities: Tmt 1

CI Width Ratio (Accounting for Weight Variance vs. Not): Tmt 1

Risk CI Coverage Probabilities: Tmt 2

CI Width Ratio (Accounting for Weight Variance vs. Not): Tmt 2

Risk CI Coverage Probabilities: Tmt 3

CI Width Ratio (Accounting for Weight Variance vs. Not): Tmt 3

Figure 6. *Simulation results for 120 events. Cox model with effects for treatment indicators, score* S *and covariates* $z_1, z_2, \ldots, z_5$*. Propensity scores from multinomial logistic regression model with effects for* $z_1, z_2, \ldots, z_5$*. Stabilized weights. Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

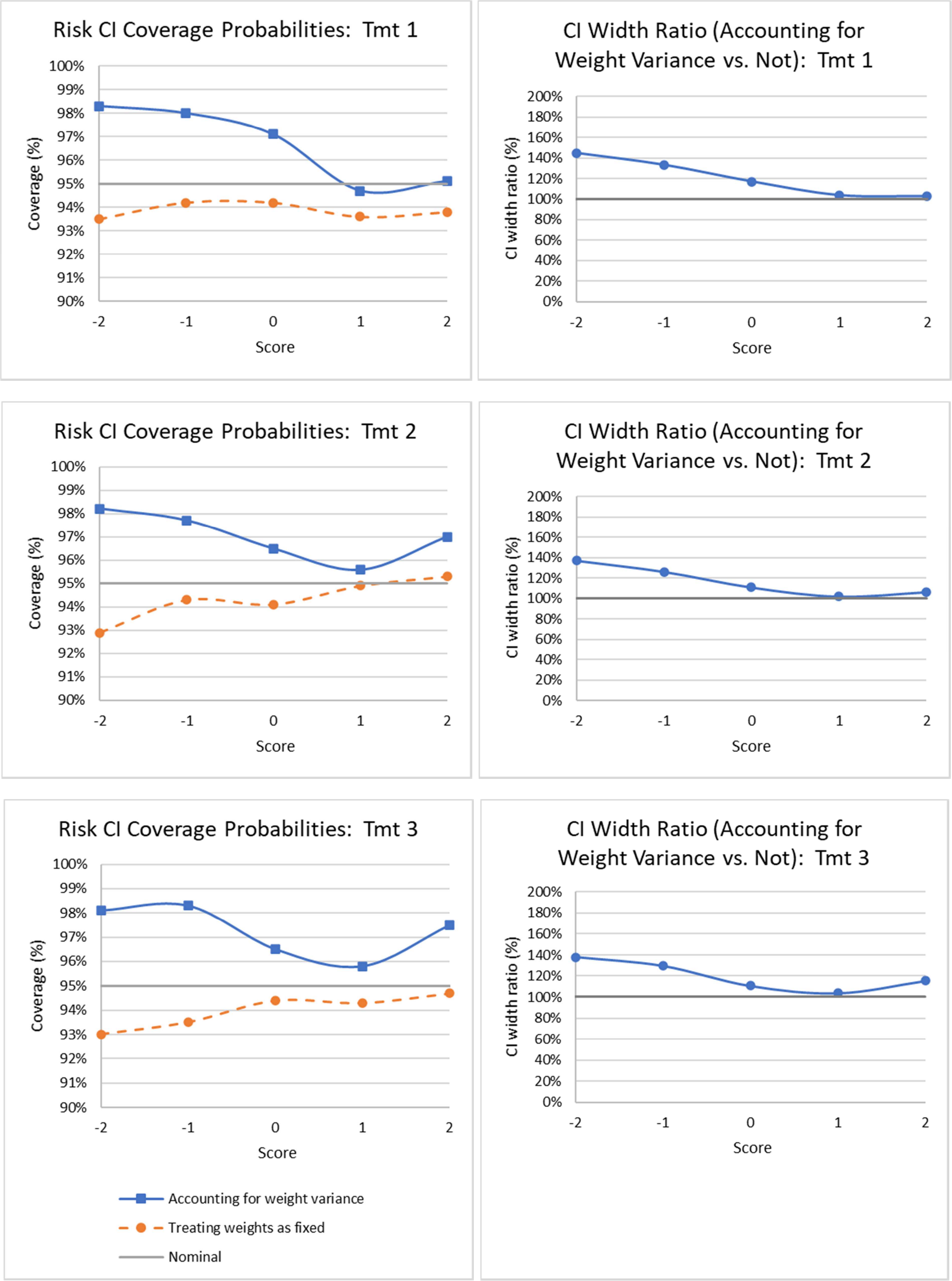


Figure 7. *Simulation results for 160 events, Cox model with effects for treatment indicators, score S and covariates $z_1,z_2,\ldots,z_5$. Propensity scores from multinomial logistic regression model with effects for $z_1,z_2,\ldots,z_5$. Stabilized weights. Coverage probabilities of 95% confidence intervals for risk and confidence interval width ratios (accounting for weight variance versus not) for specified covariate values.*

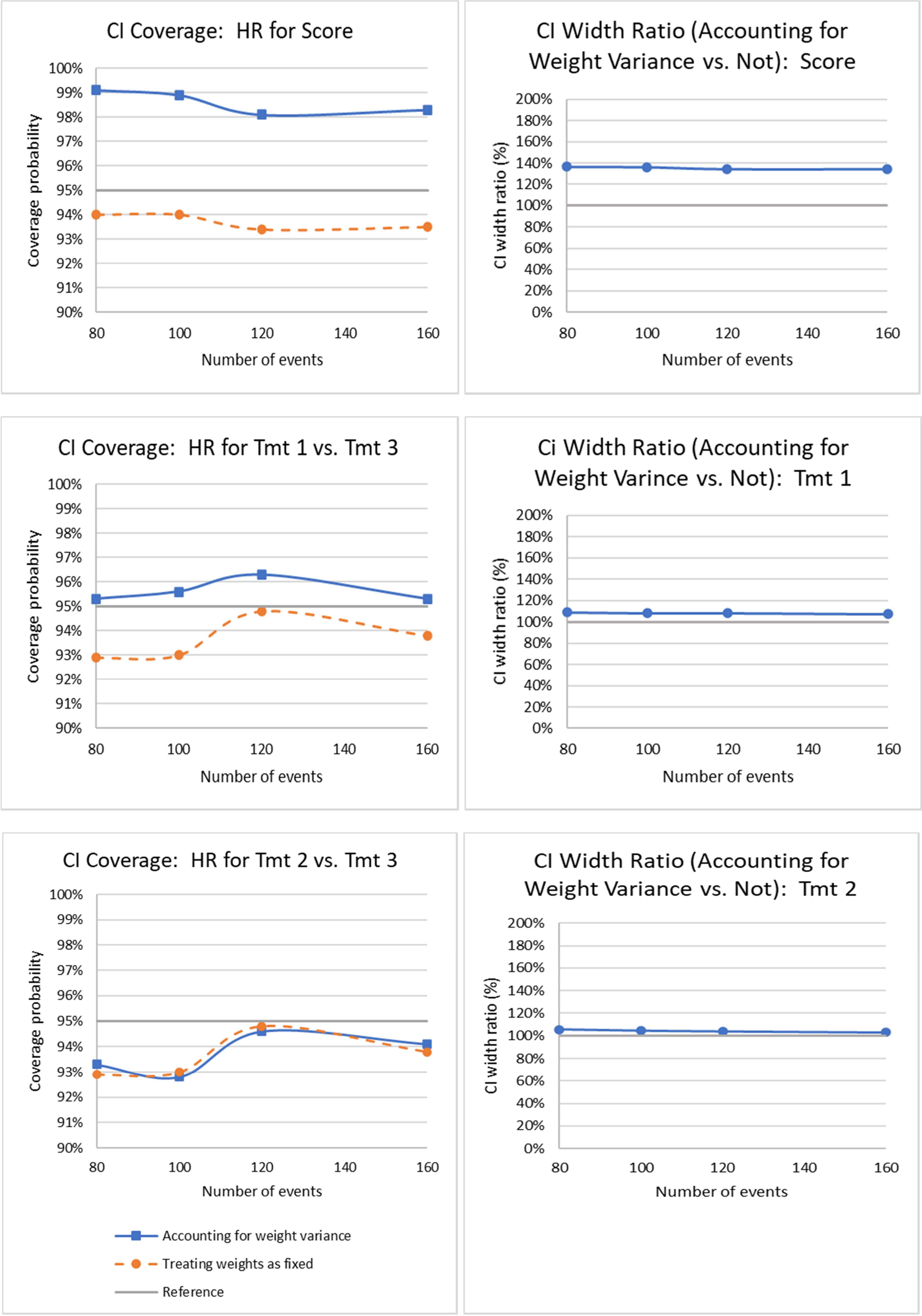


Figure 8. *Simulation results for hazard ratio CI coverage probabilities. Cox model with effects for treatment indicators, score S and covariates $z_1, z_2, \ldots, z_5$. Propensity scores from multinomial logistic regression model with effects for $z_1, z_2, \ldots, z_5$. Stabilized weights.*

Simulations were also conducted for analyses that did not stabilize the weights. In these analyses, the weight was simply the inverse propensity for selection of the treatment received without multiplying by the overall frequency of that treatment. Consequently, simulated subjects with treatments 1 and 2 tended to have higher weights in the analysis than those with treatment 3. True coverage probabilities for confidence intervals from analyses conducted without stabilizing the weights are shown in the Appendix, Figures A1 through A4. They are less consistent than the coverage probabilities when stabilized weights were used, sometimes being more conservative and sometimes more anti-conservative. Without accounting for the variability of the weights, the coverage probabilit (ies of confidence intervals were often well below nominal level, especially for small samples.

Simulations that used stabilized weights that were truncated at the 5th and 95th percentiles are shown in the Appendix, Figures A5 through A8. True coverage probabilities for confidence intervals computed using these weights were at or above nominal level and tended to be slightly less conservative than the intervals computed without truncation.

The logistic regression models for treatment propensity used in the primary simulations did not include covariates related to treatment selection but not to the event of interest. It is generally not helpful to include such covariates in propensity score models, as they tend to decrease the precision of the subsequent estimation (here, using the weighted Cox regression) without decreasing their bias [18, 24]. However, it is not uncommon for practitioners to include covariates in the logistic regression model that are associated with the probability of receiving treatment but not associated with event risk. To cover this situation, simulations were conducted using the same configuration as above but in which covariates z_2, z_3, z_4 and z_5 were associated only with the probability of treatment, not event risk. The results are shown in the Appendix Figures A9 through A13. The true coverage probabilities for the confidence intervals for risk are still at or above the nominal level for all but the smallest data set. True coverage probabilities for the confidence intervals for the hazard ratios were noticeably anti-conservative for some of the lower sample sizes.

Accounting for the variation in weights substantially increased the width of the confidence interval widths relative to treating weights as fixed with width ratios as high as 140% when estimating risks for low scores, that is, when the risk estimate itself tended to be low. When the estimated risk is low, the confidence interval widths also tend to be low, so the net effect on the absolute confidence interval width will not be large. Interval width ratios were much closer to 100% (little change in width from accounting for variation in weights) for estimates associated with higher scores.

4. Discussion

Overall, the simulation studies indicate that the confidence intervals accounting for the variation in weight estimation have true coverage probabilities at or somewhat above the nominal level. Treating the weights as fixed led to substantially anti-conservative coverage probabilities when sample sizes were small. Stabilization of the IPTW weights by multiplying them by the overall average probability of treatment substantially improved the coverage probability of confidence intervals accounting for the variation in the weight estimates. Truncation of the weights at the tails does not adversely affect the true coverage probabilities and may make them less conservative in some cases.

The results of previous work using simulations suggested that treating the weights as fixed and using the robust variance estimator of Lin and Wei [3] gives hazard ratio confidence intervals with higher than nominal coverage probability [4, 10]. The results presented here suggest the opposite: at least for smaller samples, treating the weights as fixed leads to confidence intervals with anti-conservative true coverage probabilities, sometimes substantially so. One factor that may explain the difference in simulation results is sample size. Austin [4] simulated data sets with 10 covariates and sample sizes of over 9000; Hajage and colleagues [10] simulated data sets with just one covariate (treatment indicator) and sample sizes of 10,000. The primary simulations presented here used much smaller data sets so as to evaluate the small and moderate-sample properties of the risk and hazard ratio confidence intervals. The method for simulating treatment assignment may also have led to differences in this work versus previous work in the hazard ratio confidence interval coverage probabilities. Austin [4] used simulated data sets in which each subject received treatment or no treatment and fixed the percentage of subjects who received treatment at various values ranging from less than 10% to 50%. Hajage [10] allowed the proportion of simulated subjects receiving treatment to vary as a Bernoulli random variable. In those simulations, the coverage probabilities for hazard ratio confidence intervals treating the weights as fixed were conservative when probability of treatment was 50% but at or below nominal levels when probability of treatment was 25% or 10%. The simulations conducted here considered 3 treatments with assignment modeled as a multinomial random variable with probabilities 20%, 20% and 60%. Finally, the current simulations used a higher censoring rate (approximately 75%) than those described by Hajage (0% or 50%) [10]. Austin [4] does not describe censoring in the simulations. Considering all the simulation work together, it seems fair to say that treating weights as fixed and using the robust variance estimator may result in either conservative or anti-conservative confidence intervals for hazard ratios.

Austin [4] suggested using the bootstrap, with replication of the entire process of fitting the logistic regression model, deriving the weights, and fitting the weighted Cox model, and showed that this gives confidence intervals for hazard ratios with coverage probabilities at nominal level. The bootstrap may not be practical, however, if the data set to be analyzed is very large since each model fit may take substantial compute time and resource. Also, bootstrap confidence intervals are not in general fully reproducible since they are based on random resampling and since random number generators differ across computer operating systems. For pre-planned analyses using the bootstrap, reproducibility of the results can be obtained by drawing the bootstrap sample in advance and archiving it. This option may not be available for post-hoc or exploratory analyses. If the only goal of fitting the Cox model is the estimation of the treatment effect hazard ratio, time-invariant covariates are used and stratified cohort sampling is not used, then the method of Hajage and colleagues [10] provides a viable analytic alternative. When risk estimation is a goal of the analysis, stratified cohort sampling or external time-dependent covariates are used in the Cox model, or left truncation is a feature of the data, the closed form confidence intervals described here provide a useful alternative to the bootstrap that is practical and always fully reproducible.

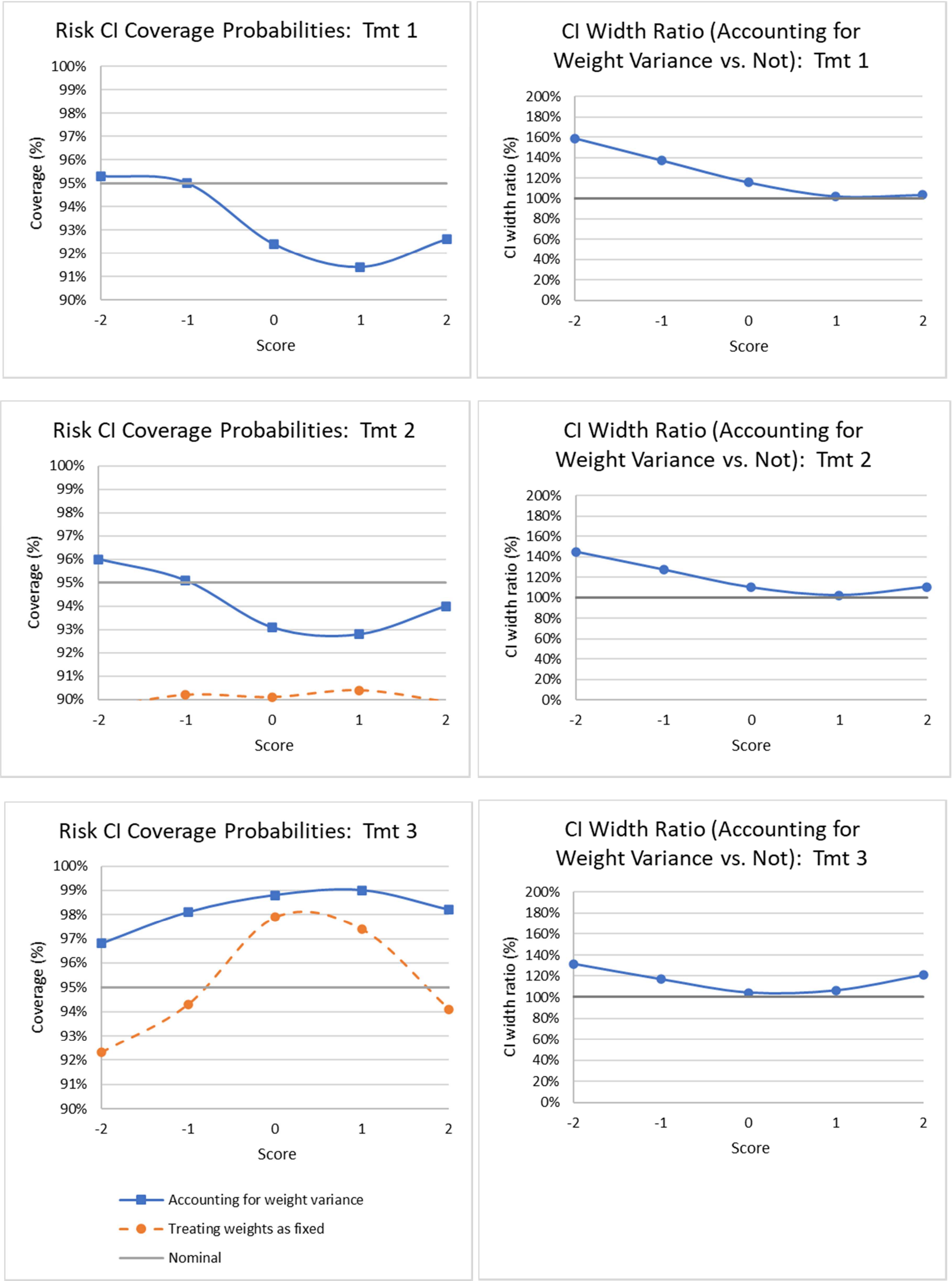
5. Software

A SAS macro that carries out the methods described here is available in the repository https://mcrager.github.io/SAS-macros/. This macro fits the specified logistic regression and weighted Cox models and provides estimates of the risk of an event on or before a specified time for specified covariate values as well as the Cox regression parameter (log hazard ratio) estimates, their standard errors and p-values from Wald tests of the null hypothesis that the log hazard ratio is zero, all accounting for the variability of the estimated weights. In addition, the macro provides assessments of the balance of covariates across treatments (or other interventions as specified) using standardized differences.

6. Conclusion

When using propensity scores and IPTW weighting in a Cox proportional hazards regression, failure to account for the variability in the propensity score estimation can lead to hazard ratio and event risk confidence intervals with true coverage probabilities substantially below nominal levels. The methods described here account for the propensity score estimation variation in closed form and provide confidence intervals with coverage probabilities at or above nominal levels. The methods accommodate weighted analyses from stratified cohort sampling studies, external time-dependent covariates and left truncation. Software to carry out the calculations is freely available.

Appendix



Risk CI Coverage Probabilities: Tmt 1

CI Width Ratio (Accounting for Weight Variance vs. Not): Tmt 1

Risk CI Coverage Probabilities: Tmt 2

CI Width Ratio (Accounting for Weight Variance vs. Not): Tmt 2

Risk CI Coverage Probabilities: Tmt 3

CI Width Ratio (Accounting for Weight Variance vs. Not): Tmt 3

Figure A1. *Simulation results using conventional IPTW weights (not stabilized) for 40 events. Cox model with effects for treatment indicator, score* S *and* z_1*. Propensity scores from multinomial logistic regression model with effects for* z_1, z_2, z_3, z_4, z_5*. . Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

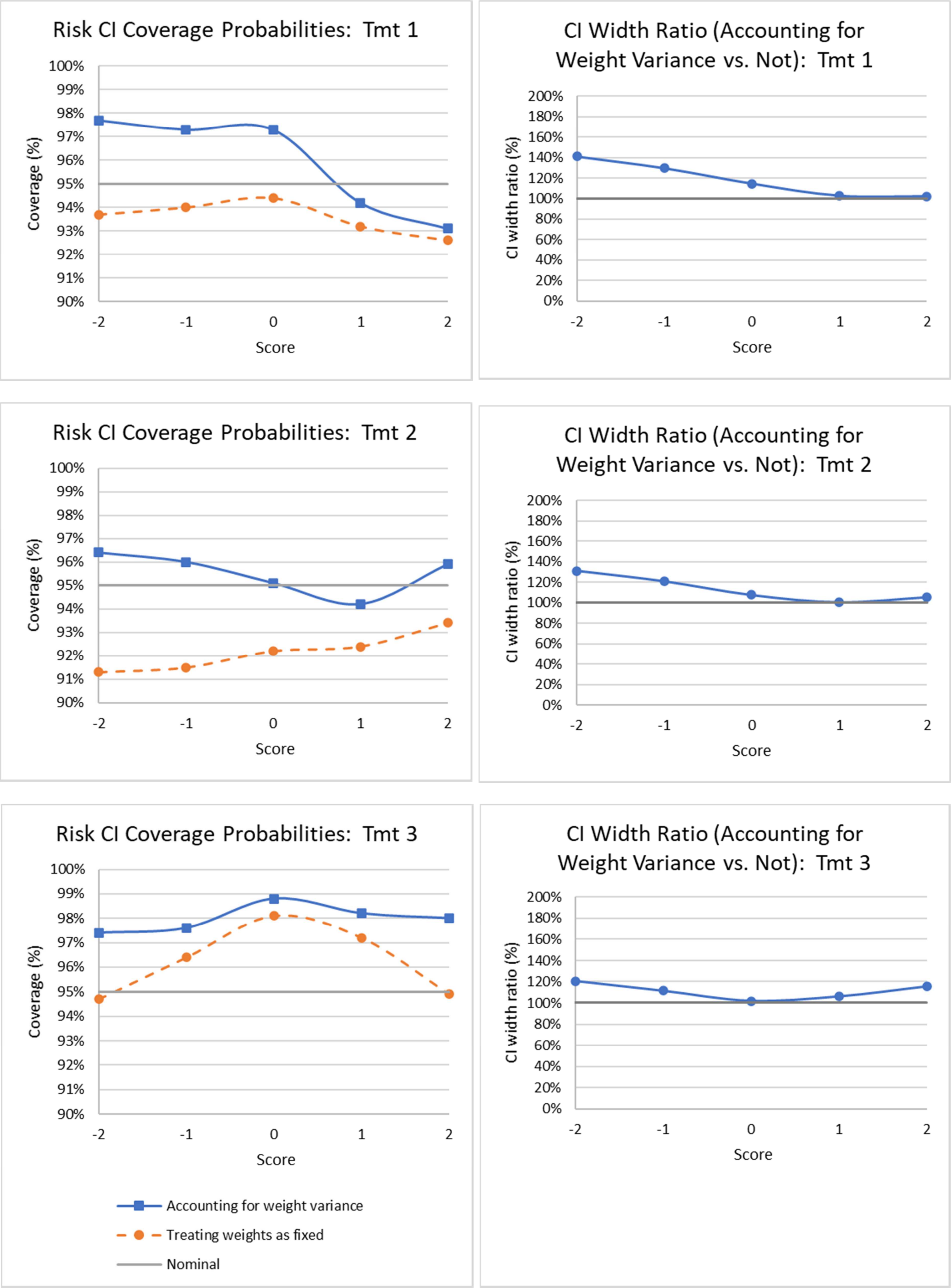


Figure A2. *Simulation results using conventional IPTW weights (not stabilized) for 80 events. Cox model with effects for treatment indicators, score S and z_1. Propensity scores from multinomial logistic regression model with effects for $z_1, z_2, \ldots, z_5$. Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

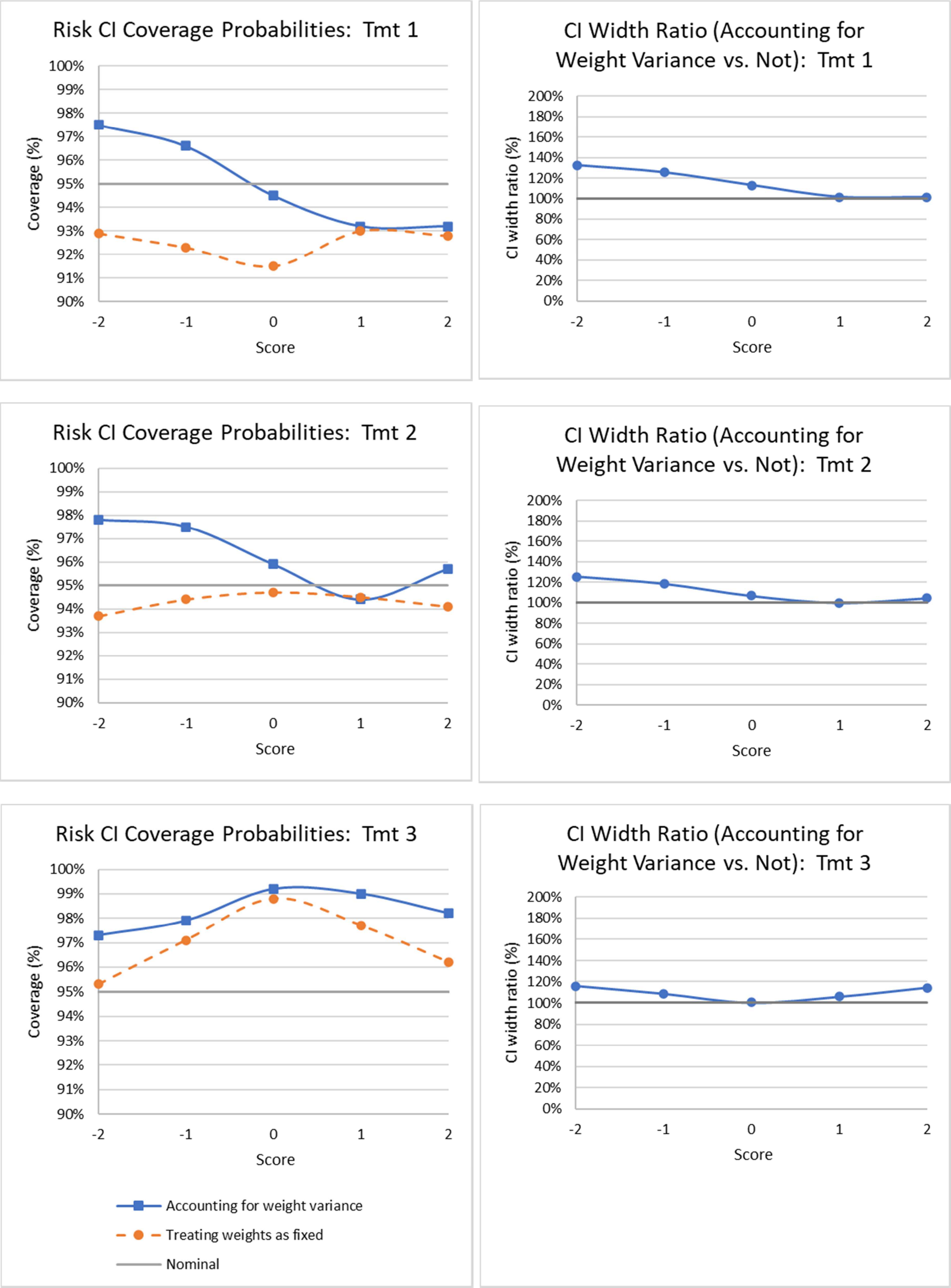


Figure A3. *Simulation results using conventional IPTW weights (not stabilized) for 160 events. Cox model with effects for treatment indicators, score* S *and* z_1. *Propensity scores from multinomial logistic regression model with effects for* $z_1, z_2, \ldots, z_5$. *Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

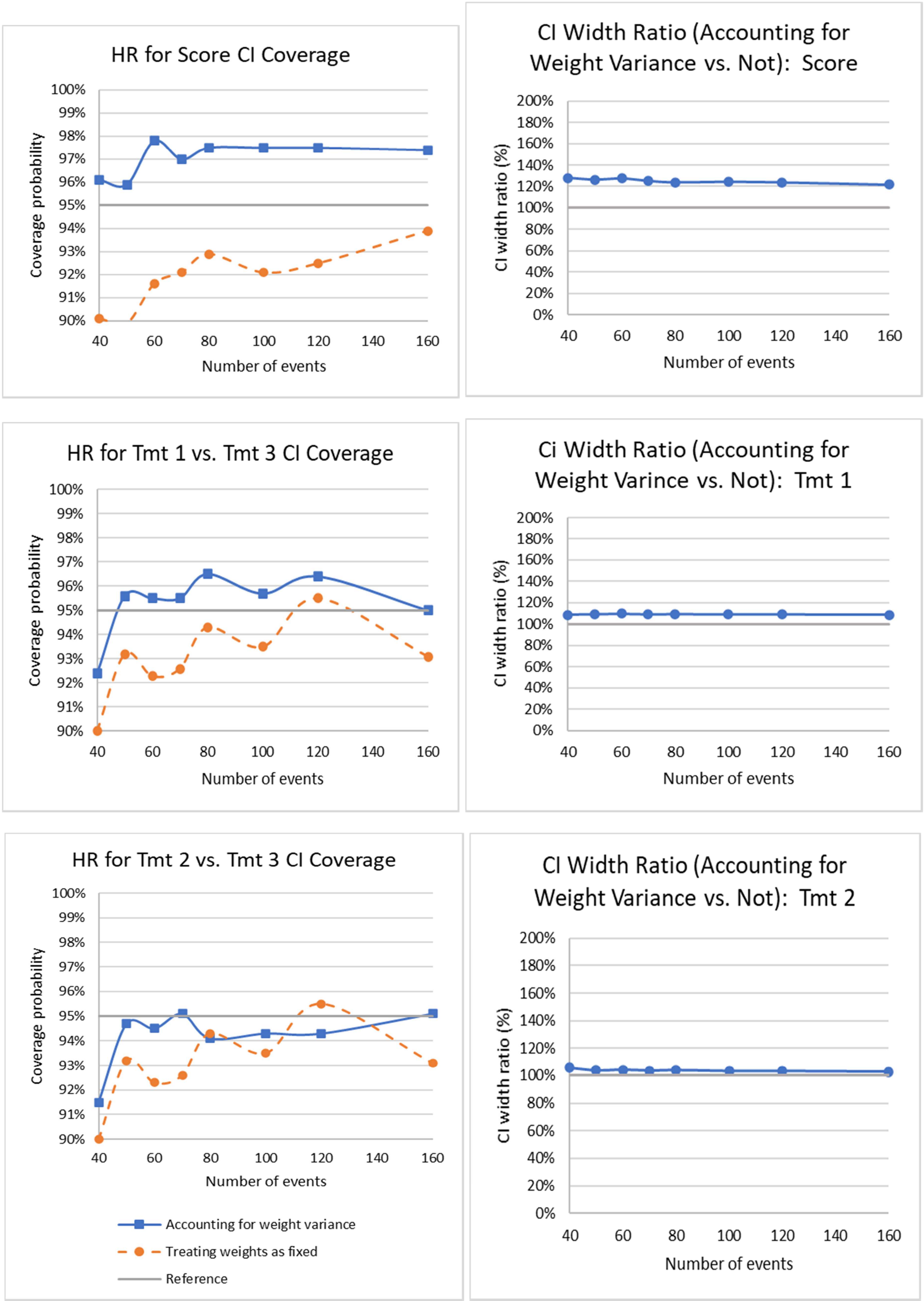


Figure A4. *Simulation results using conventional IPTW weights (not stabilized) for hazard ratio CI coverage probabilities. Cox model with effects for treatment indicators, score and za1. Propensity scores from multinomial logistic regression model with effects for* $z_1, z_2, \ldots, z_5$.

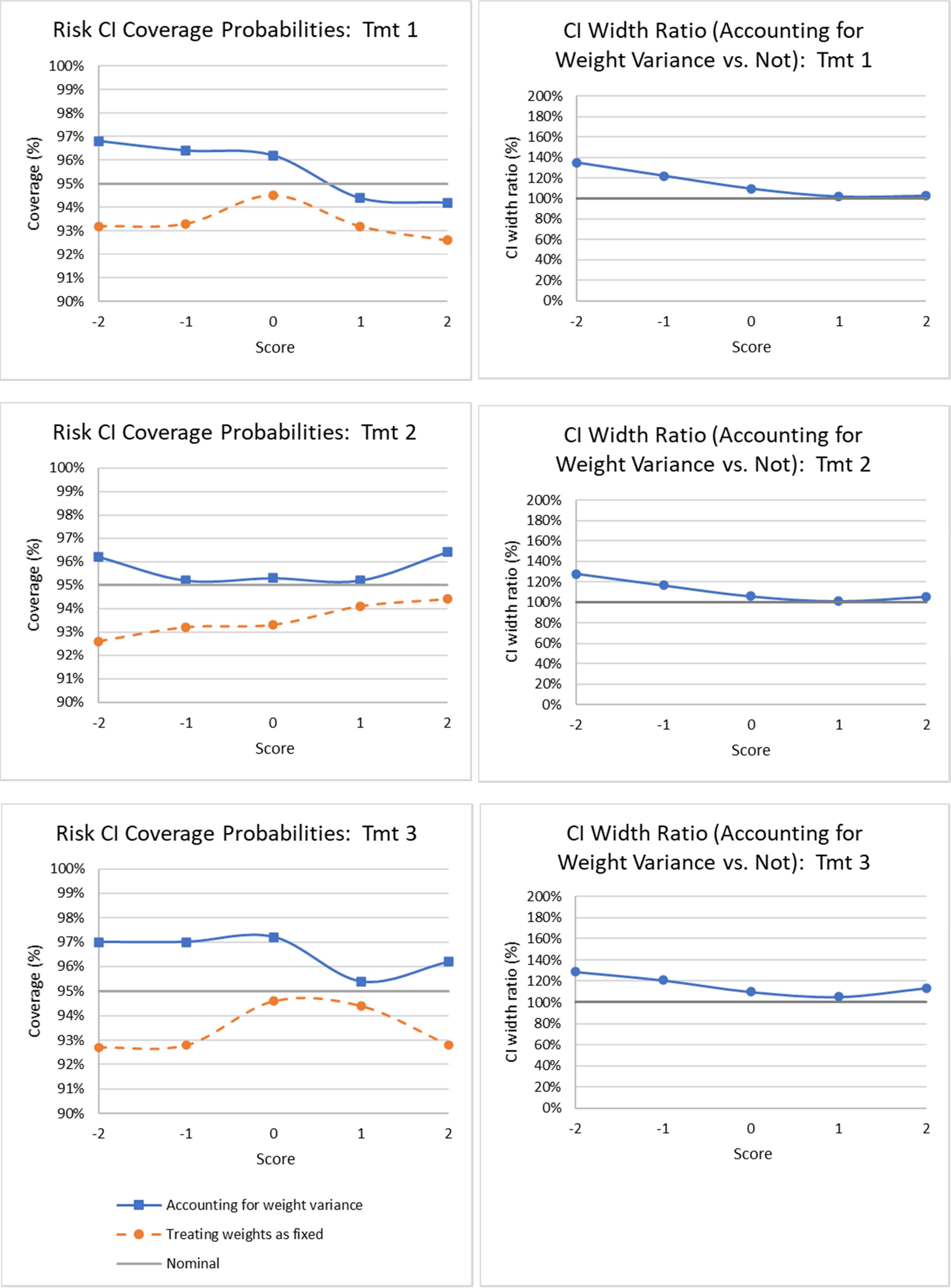


Figure A5. *Simulation results using stabilized weights truncated at 5th, 95th percentiles, for 40 events. Cox model with effects for treatment indicators, score* S *and* z_1. *Propensity scores from multinomial logistic regression model with effects for* z_1, z_2, z_3, z_4, z_5. *Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

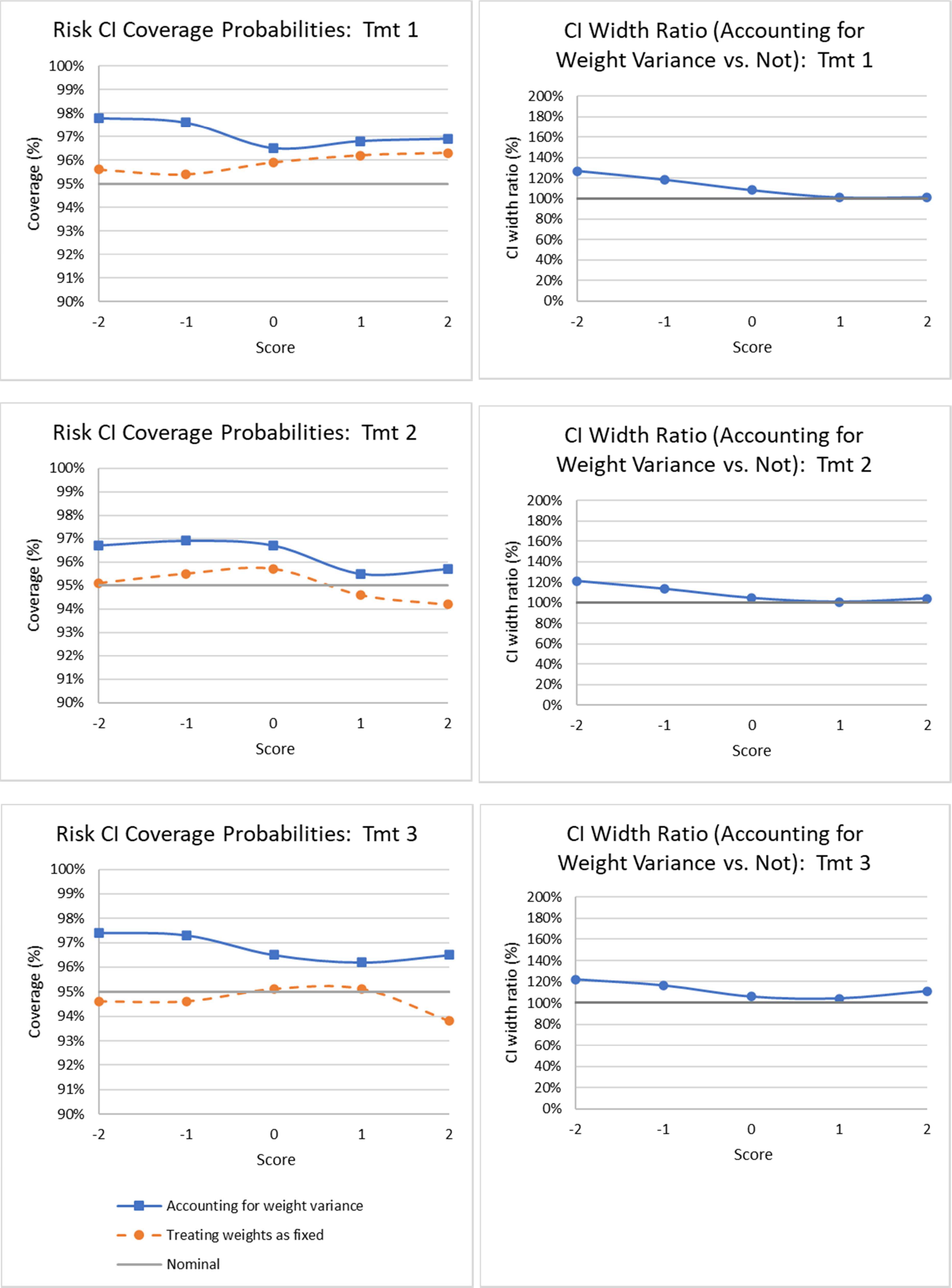


Figure A6. *Simulation results using stabilized weights truncated at 5th, 95th percentiles, for 80 events. Cox model with effects for treatment indicators, score* S *and* z_1*. Propensity scores from multinomial logistic regression model with effects for* $z_1, z_2, \ldots, z_5$*. Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

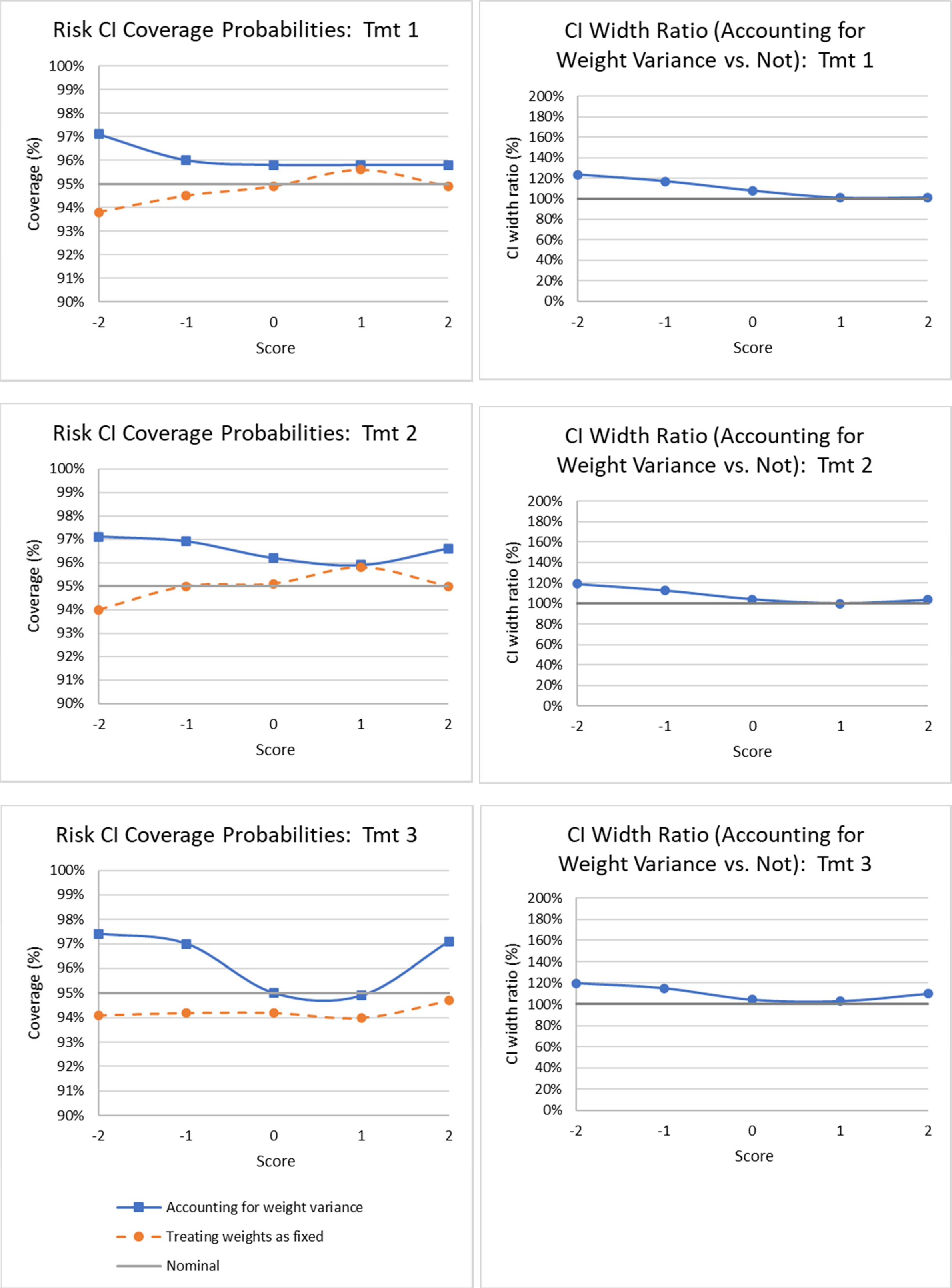


Figure A7. *Simulation results using stabilized weights truncated at 5th, 95th percentiles, for 160 events. Cox model with effects for treatment indicators, score* S *and* z_1. *Propensity scores from multinomial logistic regression model with effects for* $z_1, z_2, \ldots, z_5$. *Coverage probabilities of 95% confidence intervals (CI) for risk and CI width ratios for specified covariate values.*

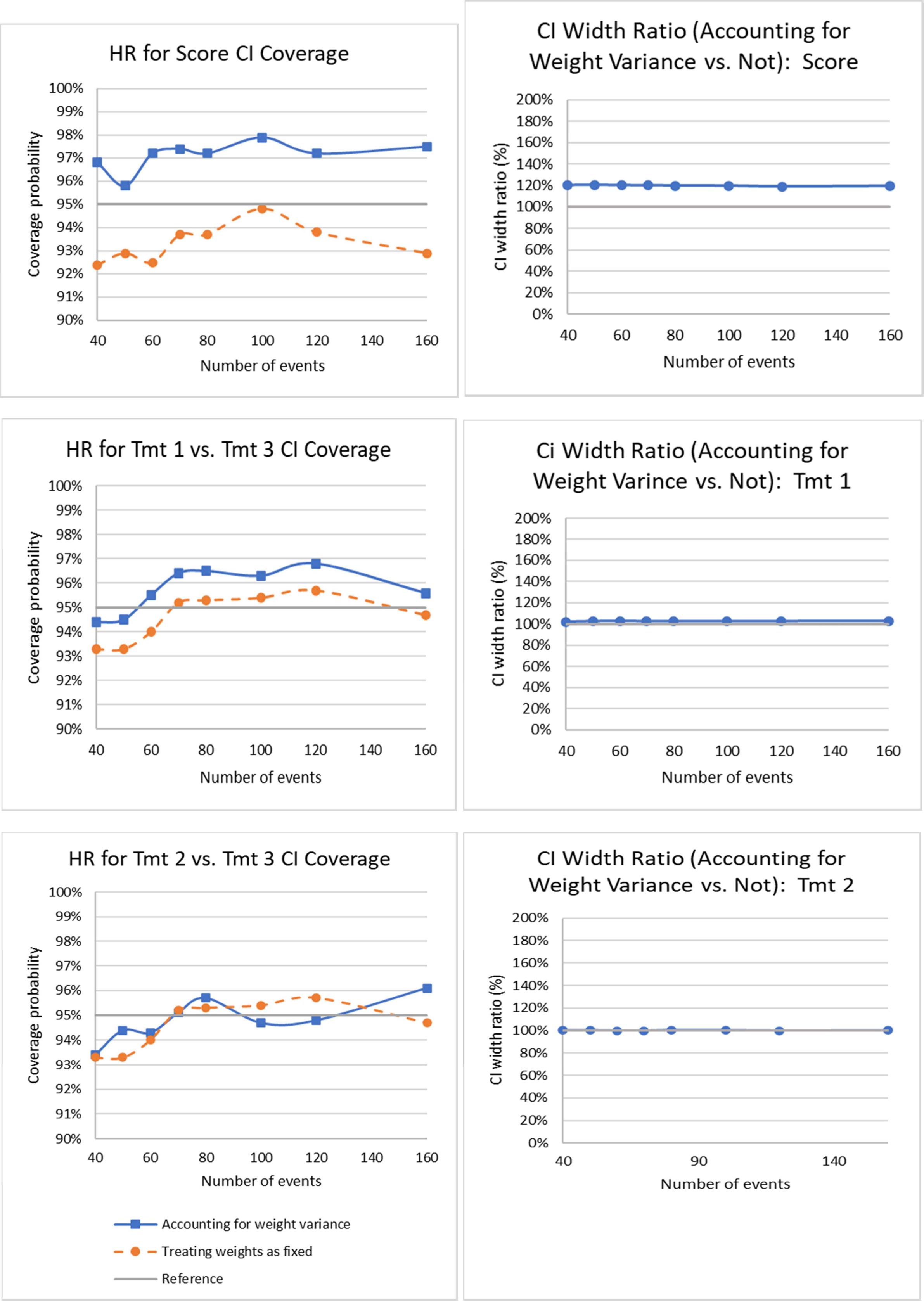


Figure A8. *Simulation results using stabilized weights truncated at 5th, 95th percentiles for hazard ratio CI coverage probabilities. Cox model with effects for treatment indicators, score and za1. Propensity scores from multinomial logistic regression model with effects for* $z_1, z_2, \ldots, z_5$.

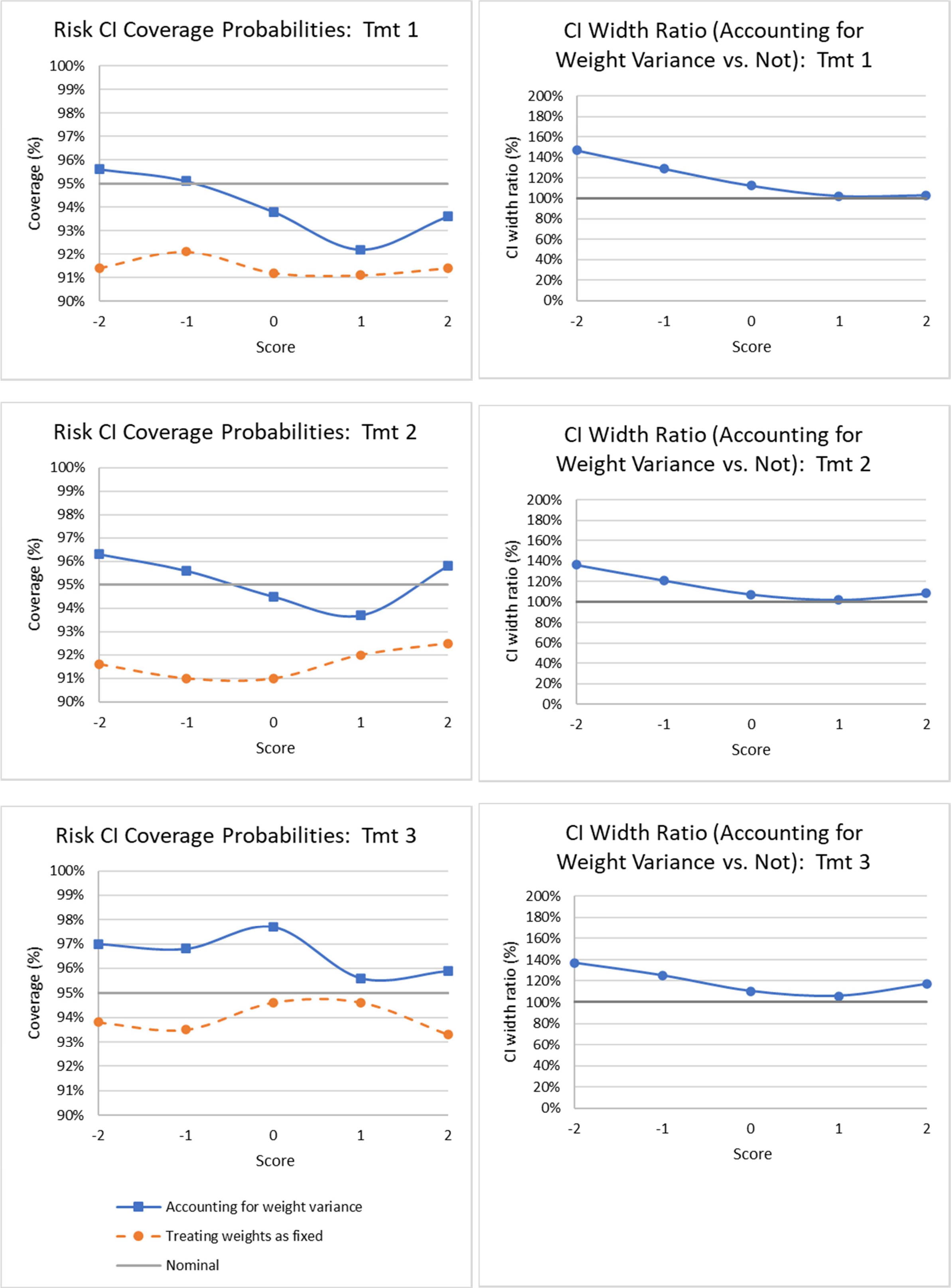


Figure A9. *Simulation results using covariates in the logistic regression that are not associated with risk, 40 events. Cox model with effects for treatment indicators, score and z_1. Propensity scores from multinomial logistic regression model with effects for $z_1, z_2, \ldots, z_5$, with only z_1 associated with risk. Stabilized weights.*

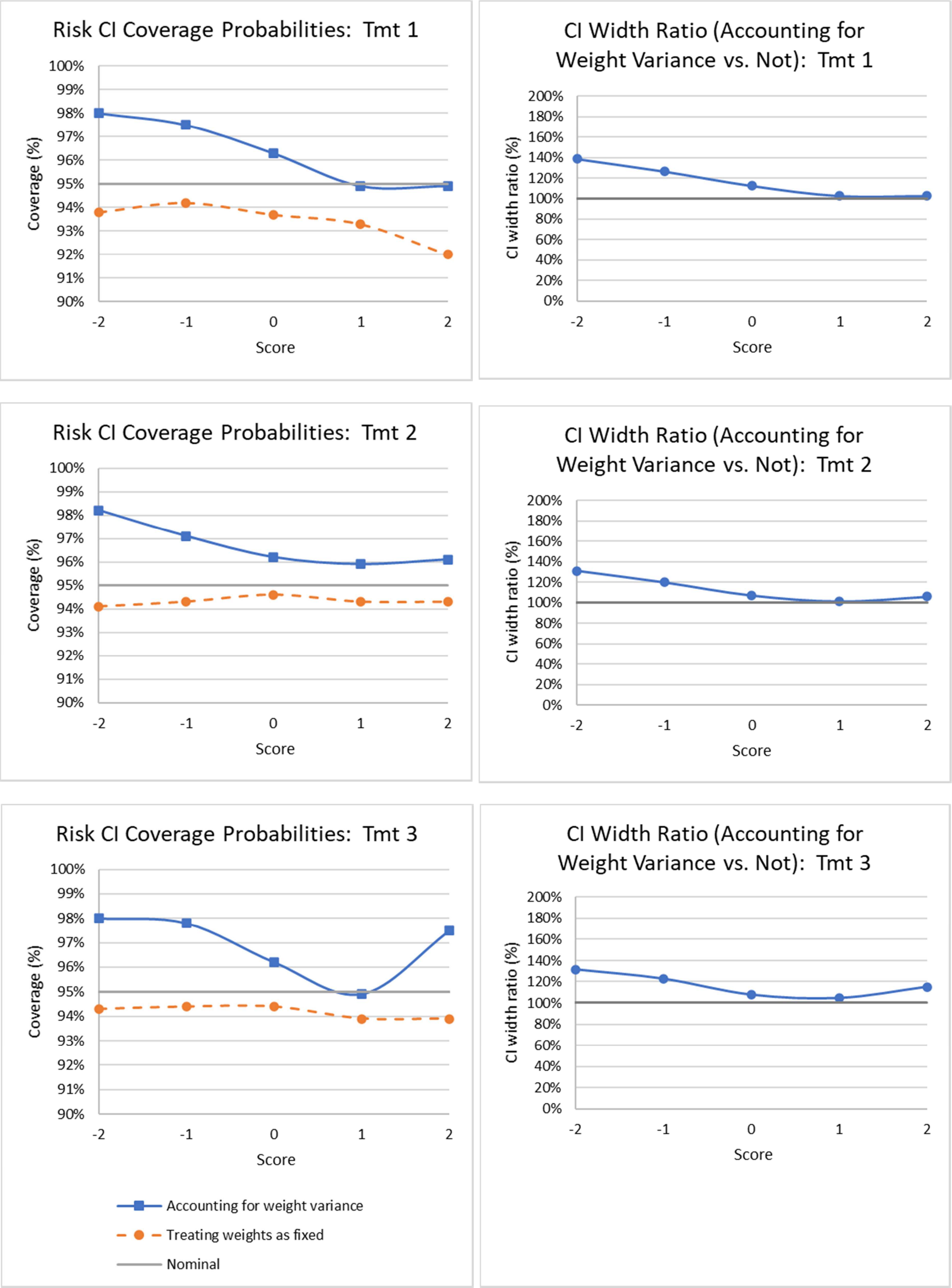


Figure A10. *Simulation results using covariates in the logistic regression that are not associated with risk, 80 events. Cox model with effects for treatment indicators, score and z_1. Propensity scores from multinomial logistic regression model with effects for $z_1,z_2,\ldots,z_5$, with only z_1 associated with risk. Stabilized weights.*

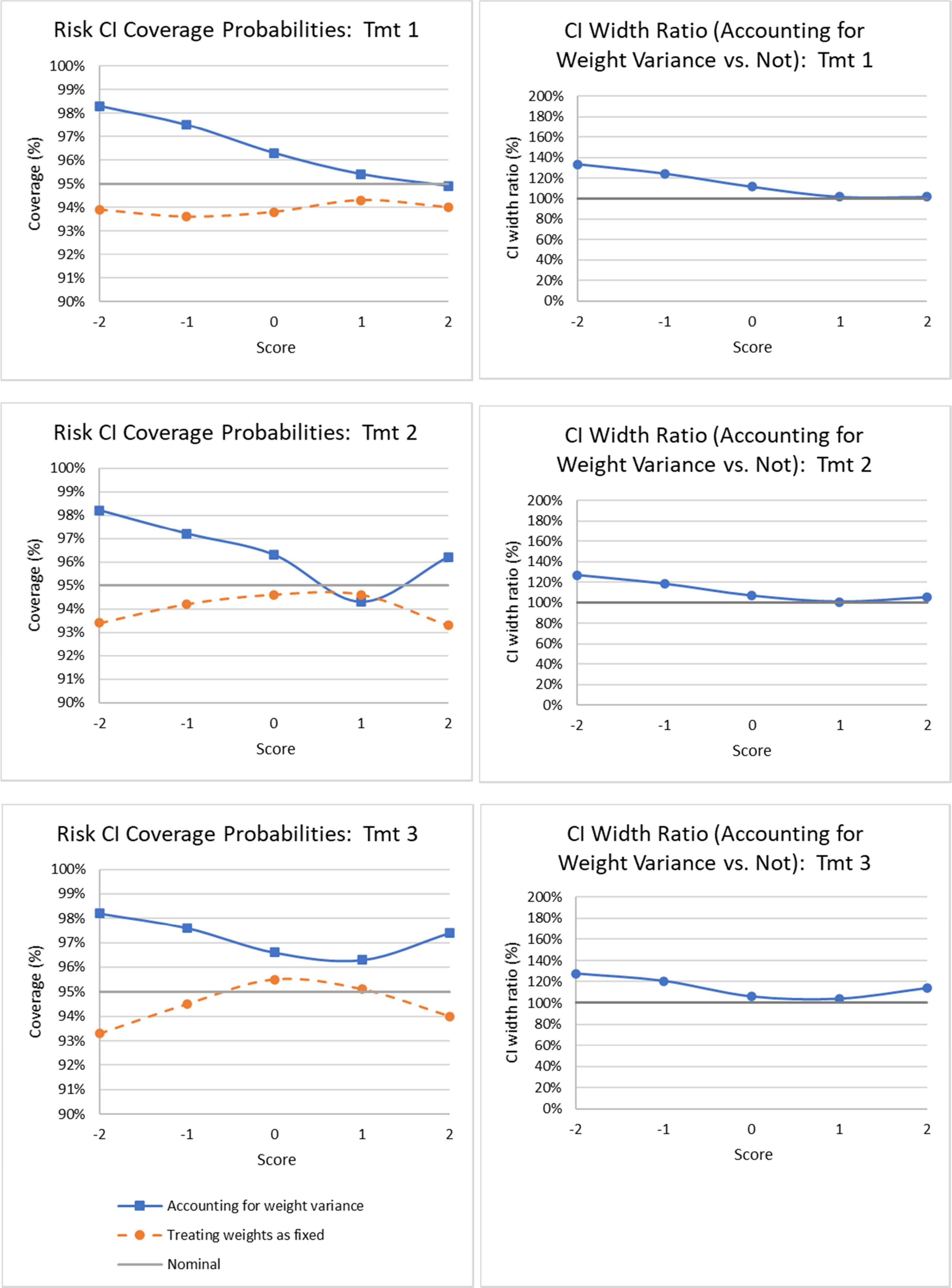


Figure A11. *Simulation results using covariates in the logistic regression that are not associated with risk, 160 events. Cox model with effects for treatment indicators, score and* z_1*. Propensity scores from multinomial logistic regression model with effects for* $z_1, z_2, \ldots, z_5$*, with only* z_1 *associated with risk. Stabilized weights.*

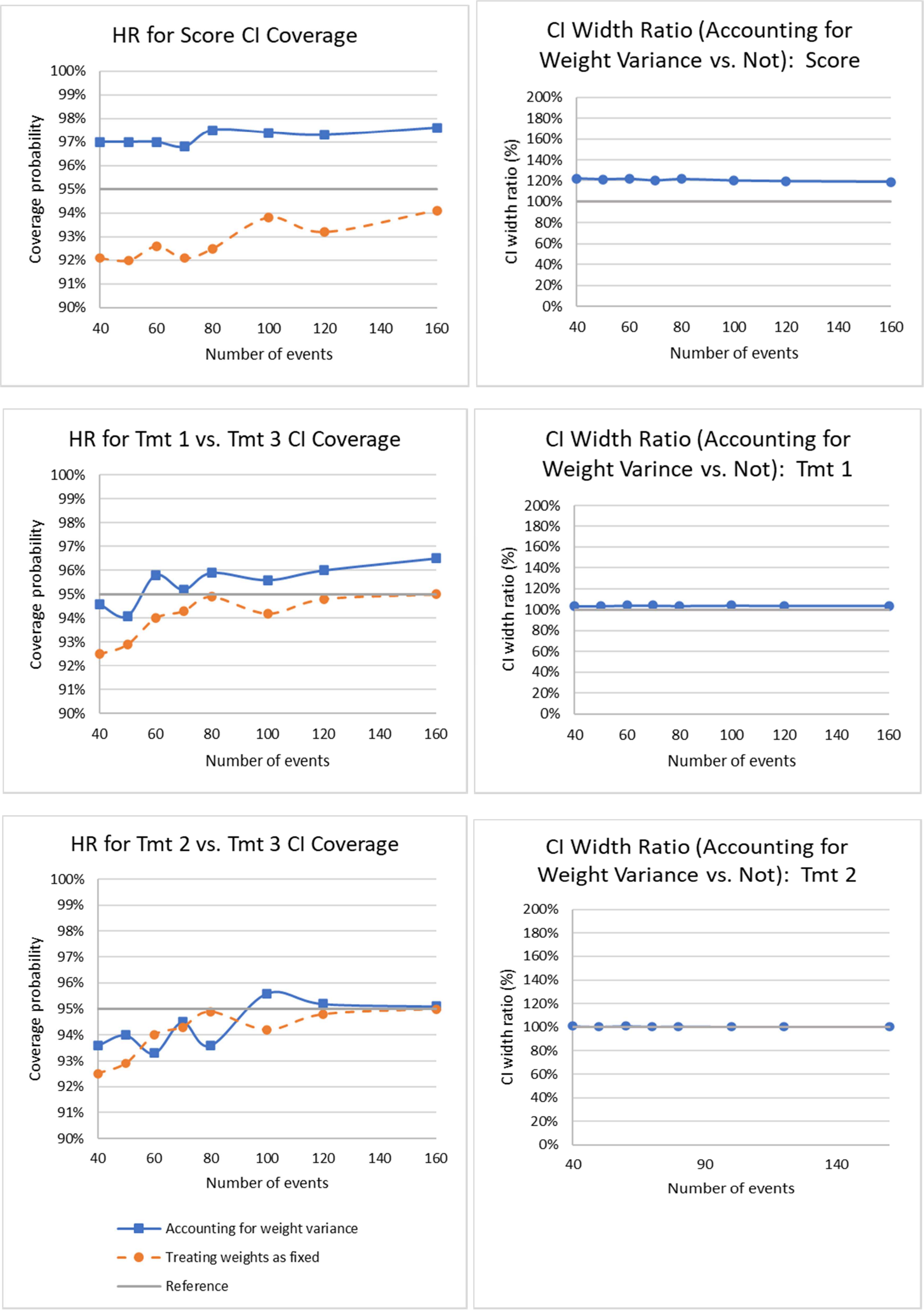


Figure A12. *Simulation results using covariates in the logistic regression that are not associated with risk, hazard ratio CI coverage probabilities. Cox model with effects for treatment indicators, score and* z_1. *Propensity scores from multinomial logistic regression model with effects for* $z_1, z_2, \ldots, z_5$, *with only* z_1 *associated with risk. Stabilized weights.*

References

[1] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 76, 41–55.

[2] Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. Statistics in Medicine 32, 2837–2849.

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

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

[5] Gray, R. (2009). Weighted analyses for cohort sampling designs. Lifetime Data Analysis 15, 24–40.

[6] Kalbfleisch, J. D., & Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data, 2nd ed. Wiley, New York.

[7] Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). Statistical Models Based on Counting Processes, 1st ed. Spring-Verlag, New York.

[8] Cortese, G., & Andersen, P. K. (2009). Competing risks and time-dependent covariates. Biometrical Journal 51, 138–158.

[9] Cain, K., C., Harlow, S. D., Little, R., J., Nan, B., Yosef, M., Taffe, J. R., & Elliott MR (2011). Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. American Journal of Epidemiology 173, 1078–1084.

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

[11] Lee, M.-J., & Lee S. (2022). Review and comparison of treatment effect estimators using propensity and prognostic scores. International Journal of Biostatistics. doi: 10.1515/ijb-2021-0005.

[12] Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. Journal of the American Statistical Association 113, 390-400. https:/doi.org/10.1080/02621459.2016.1260466

[13] Li, F., Thomas, L. E., & Li, F. (2019). Addressing extreme propensity scores via the overlap weights. American Journal of Epidemiology 188, 250–257.

[14] Mao, H., Li, L., Wei, Y., & Shen, Y. (2018). On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. Statistics in Medicine 36, 3745-3763.

[15] Cheng, C., Li, F., Thomas, L. E., & Li, F. F. (2022). Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. American Journal of Epidemiology 191, 1140-1151.

[16] Thomas, L. E., Li, F., & Pencina, M. J. (2020). Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. Journal of the American Medical Association 323, 2417–2418.

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

[18] Austin, P. C., & 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.

[19] Pugh, M., Robbins, J., Lipsitz, S., & Harrington, D. (1993). Inference in the Cox Proportional Hazards Model with Missing Covariates. Technical Report 758Z. Department of Biostatistics, Harvard School of Public Health.

[20] Therneau, T. M., & Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model. Springer, New York.

[21] Czepiel, S. A. (2002). Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. http://czep.net

[22] Tsiatis, A. (1981). A large sample study of the estimates for the integrated hazard function in Cox's regression model for survival data. The Annals of Statistics 9, 93–108.

[23] Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of Clinical Epidemiology 48, 1503–1510.

[24] Brookhart, A. M., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. American Journal of Epidemiology 163, 1149–1156.