Received: 20 September 2018 | Revised: 11 August 2019 | Accepted: 27 September 2019

DOI: 10.1002/sim.8404

RESEARCH ARTICLE

WILEY Statistics in Medicine

Confidence interval estimation for the changepoint of treatment stratification in the presence of a qualitative covariate-treatment interaction

Bernhard Haller[1] | **Ulrich Mansmann**[2] | **Dennis Dobler**[3] | **Kurt Ulm**[1] | **Alexander Hapfelmeier**[1]

[1]School of Medicine, Institute for Medical Informatics, Statistics and Epidemiology, Technical University of Munich, Munich, Germany

[2]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence
Bernhard Haller, School of Medicine, Institute for Medical Informatics, Statistics and Epidemiology, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany.
Email: bernhard.haller@tum.de

The goal in stratified medicine is to administer the "best" treatment to a patient. Not all patients might benefit from the same treatment; the choice of best treatment can depend on certain patient characteristics. In this article, it is assumed that a time-to-event outcome is considered as a patient-relevant outcome and a qualitative interaction between a continuous covariate and treatment exists, ie, that patients with different values of one specific covariate should be treated differently. We suggest and investigate different methods for confidence interval estimation for the covariate value, where the treatment recommendation should be changed based on data collected in a randomized clinical trial. An adaptation of Fieller's theorem, the delta method, and different bootstrap approaches (normal, percentile-based, wild bootstrap) are investigated and compared in a simulation study. Extensions to multivariable problems are presented and evaluated. We observed appropriate confidence interval coverage following Fieller's theorem irrespective of sample size but at the cost of very wide or even infinite confidence intervals. The delta method and the wild bootstrap approach provided the smallest intervals but inadequate coverage for small to moderate event numbers, also depending on the location of the true changepoint. For the percentile-based bootstrap, wide intervals were observed, and it was slightly conservative regarding coverage, whereas the normal bootstrap did not provide acceptable results for many scenarios. The described methods were also applied to data from a randomized clinical trial comparing two treatments for patients with symptomatic, severe carotid artery stenosis, considering patient's age as predictive marker.

KEYWORDS

changepoint, confidence intervals, covariate-treatment interaction, Cox regression, stratified medicine

1 | INTRODUCTION

For many medical indications, it is observed that treatments work differently in different patients, or that different patients with the same medical indication might benefit from different treatments or treatment strategies.[1,2] Selection or identification of optimal treatment regimes for individual patients in an evidence-based manner is currently discussed. The literature provides proposals for study designs to investigate differential treatment effects[3-5] and statistical methods for identification of relevant predictive markers[6,7] or subgroups benefiting from different treatments.[8]

It is possible to investigate differential treatment effects in data from a two-group randomized clinical trial by estimating a mean difference, an odds ratio, or a hazard ratio in dependence of a continuous variable with a regression model by including the main effect of treatment, the main effect of the covariate, and their interaction term.[9,10] A test for homogeneity of the treatment effect over the range of the covariate is obtained under common model assumptions. When a qualitative covariate-treatment interaction is observed,[11] the covariate value that is associated with no difference between the two treatment groups can be estimated from the regression coefficients for treatment and the covariate-treatment interaction. This is the value of the covariate, where the difference between the mean outcomes of the treatments is zero or the odds ratio or the hazard ratio between the treatments is one and the direction of the treatment effect changes.[12] We call the covariate value, where the superior treatment changes, the *changepoint of treatment stratification* in this article. As this estimate relies on a finite sample and is prone to sampling variability, a confidence interval for that value should be estimated and provided to indicate uncertainty, as commonly recommended for treatment effect estimates.[13,14]

In this article, we present and investigate different approaches for calculation of a confidence interval for this *changepoint of treatment stratification* based on data collected in a randomized clinical trial with a time-to-event outcome. An adaptation of Fieller's theorem for confidence interval estimation for the ratio of two means,[15,16] the delta method for transformation of maximum-likelihood estimates,[17,18] normal and percentile-based bootstrap approaches,[19,20] as well as the wild bootstrap[21,22] are presented and investigated in a simulation study. Relevant characteristics as the number of investigated patients, the censoring distribution, the location of the changepoint in the covariate distribution, and the distribution of the covariate of interest are varied in the simulation study, and their influence on the performance of the different confidence interval estimators is assessed. Extensions to situations with multiple covariates are presented and investigated. Performance of the methods is evaluated by the proportion of simulation runs, where the true changepoint is covered by the confidence interval (coverage)[23] and the distribution of confidence interval width.

Moreover, the different methods were applied to data from the stent-protected angioplasty versus carotid endarterectomy (SPACE) study, a randomized clinical trial comparing two interventions for treatment of patients with severe, symptomatic carotid stenosis.[24,25] In our analysis, the *changepoint of treatment stratification* regarding patient's age was investigated, as a qualitative age-treatment interaction was observed in the trial.

The article is organized as follows. In Section 2, estimation of the interaction term using a standard Cox regression model and calculation of the estimate for the *changepoint of treatment stratification* from the estimated regression coefficients is presented. In Section 3, the different methods for confidence interval estimation for that changepoint are introduced. The simulation study, including the investigated scenarios, methods for simulation of the data, methods for analysis of the generated data, and the observed results, is presented in Section 4. Application of the different methods for confidence interval estimation to the data from the SPACE trial is presented in Section 5, and a discussion of the observed results is given in Section 6.

All analyses presented in the manuscript were performed using the statistical software R.[26]

2 | ESTIMATION OF THE CHANGEPOINT OF TREATMENT STRATIFICATION FROM A COX REGRESSION MODEL

The Cox regression model,[27] is widely used for analysis of time-to-event data in clinical research as well as in epidemiology. Consequently, we decided to focus on the Cox model for estimation of the *changepoint of treatment stratification* in this article. Following the previously described scenario, the interaction between a continuous covariate of interest X and treatment $G \in \{0; 1\}$ is to be estimated from data observed in a randomized clinical trial with a potentially censored event-time outcome T. Moreover, scenarios with further prognostic or predictive variables, which will be called Z_1 to Z_k in our manuscript, were considered.

In order to estimate the *changepoint of treatment stratification* for the covariate of interest X, a Cox regression model, including the main effects of treatment G and the covariate X as well as their interaction term $G \times X$ and, where applicable,

main effects of further covariates $Z_1, \ldots, Z_k$ and interaction terms between those covariates and treatment group, is fitted to the data. In this article, we consider three different scenarios:

Model I – No further covariates:

$$\lambda(t \mid X, G) = \lambda_0(t) \exp(X\beta_X + G\beta_G + G \times X\beta_{G\times X}). \tag{1}$$

Model II – Considering the effect of k further prognostic variables $Z_1, \ldots, Z_k$ not interacting with treatment:

$$\lambda(t \mid X, G, Z_1, \ldots, Z_k) = \lambda_0(t) \exp\left(X\beta_X + G\beta_G + G \times X\beta_{G\times X} + \sum_{i=1}^{k} Z_i \beta_{Z_i}\right). \tag{2}$$

Model III – Moreover, considering the interaction with treatment for l covariates $Z_1, \ldots, Z_l$ (with $l \leq k$):

$$\lambda(t \mid X, G, Z_1, \ldots, Z_k) = \lambda_0(t) \exp\left(X\beta_X + G\beta_G + G \times X\beta_{G\times X} + \sum_{i=1}^{k} Z_i \beta_{Z_i} + \sum_{j=1}^{l} G \times Z_j \beta_{G\times Z_j}\right). \tag{3}$$

Here, t denotes time, $\lambda_0(t)$ is the unspecified baseline hazard rate, and β_G, β_X, β_{Z_i}, $\beta_{G\times X}$, and $\beta_{G\times Z_j}$ are the regression coefficients for treatment, the covariates, and the covariate-by-treatment interactions, respectively. Regression coefficients can be estimated by maximizing the partial (log-)likelihood, and the variance-covariance matrix can be derived as the inverse of the observed Fisher information matrix (more details can, eg, be found in the textbooks by Therneau and Grambsch[28] or Kalbfleisch and Prentice.[29])

When an interaction term between the covariate of interest X and treatment is incorporated in the regression model, it is investigated whether the hazard ratio between the two treatments $G = 1$ and $G = 0$ is the same for all values of the covariate X or whether the treatment effect depends on the covariate value. A statistical test for

$$\mathrm{H}_0 : \beta_{G\times X} = 0 \qquad \text{vs.} \qquad \mathrm{H}_1 : \beta_{G\times X} \neq 0,$$

ie, a test on homogeneity of the treatment effect over the range of X, can be performed using the estimated regression coefficients and the standard error derived from the variance-covariance matrix (Wald test). An estimate for the hazard ratio between the two treatment groups depending on the covariate of interest X at a given covariate value can be derived for the three different models. While the estimated hazard ratio does only depend on the value of X for Model I and Model II, the estimate of the hazard ratio is also a function of Z_1 to Z_l for Model III.

Model I:

$$\begin{aligned}\widehat{\mathrm{HR}}(x) &= \frac{\hat\lambda(t \mid x, G = 1)}{\hat\lambda(t \mid x, G = 0)} \\ &= \frac{\hat\lambda_0(t) \exp(\hat\beta_G + x\hat\beta_X + x\hat\beta_{G\times X})}{\hat\lambda_0(t) \exp(x\hat\beta_X)} = \exp(\hat\beta_G + x\hat\beta_{G\times X}).\end{aligned}$$

Model II:

$$\begin{aligned}\widehat{\mathrm{HR}}(x) &= \frac{\hat\lambda(t \mid x, G = 1, z_1, \ldots, z_k)}{\hat\lambda(t \mid x, G = 0, z_1, \ldots, z_k)} \\ &= \frac{\hat\lambda_0(t) \exp\left(\hat\beta_G + x\hat\beta_X + x\hat\beta_{G\times X} + \sum_{i=1}^{k} z_i \hat\beta_{Z_i}\right)}{\hat\lambda_0(t) \exp\left(x\hat\beta_X + \sum_{i=1}^{k} z_i \hat\beta_{Z_i}\right)} = \exp(\hat\beta_G + x\hat\beta_{G\times X}).\end{aligned}$$

Model III – Hazard ratio depending on the values for the covariates Z_1 to Z_l:

$$\widehat{\mathrm{HR}}(x, z_1, \ldots, z_l) = \frac{\hat{\lambda}(t \mid x, G = 1, z_1, \ldots, z_k)}{\hat{\lambda}(t \mid X, G = 0, z_1, \ldots, z_k)} = \frac{\hat{\lambda}_0(t) \exp\left(\hat{\beta}_G + x\hat{\beta}_X + x\hat{\beta}_{G\times X} + \sum_{i=1}^{k} z_i \hat{\beta}_{Z_i} + \sum_{j=1}^{l} z_j \hat{\beta}_{G\times Z_j}\right)}{\hat{\lambda}_0(t) \exp\left(x\hat{\beta}_X + \sum_{i=1}^{k} z_i \hat{\beta}_{Z_i}\right)} = \exp\left(\hat{\beta}_G + x\hat{\beta}_{G\times X} + \sum_{j=1}^{l} z_j \hat{\beta}_{G\times Z_j}\right).$$

Consequently, the estimated covariate value $\hat{x}_{\mathrm{cp}}$, at which the estimated hazard ratio between the two treatments equals one or equivalently the log-hazard ratio equals zero, ie, the covariate value, for which a patient is considered to have the same expected outcome from both treatments, is as follows.

For Model I and II:

$$\hat{x}_{cp} = -\frac{\hat{\beta}_G}{\hat{\beta}_{G\times X}}. \tag{4}$$

For Model III:

$$\hat{x}_{\mathrm{cp}}(z_1, \ldots, z_l) = -\frac{\hat{\beta}_G + \sum_{j=1}^{l} z_j \hat{\beta}_{G\times Z_j}}{\hat{\beta}_{G\times X}}. \tag{5}$$

In the following, we call this covariate value the estimated *changepoint of treatment stratification*.

3 | METHODS FOR CONFIDENCE INTERVAL ESTIMATION

3.1 | Adaptation of Fieller's theorem

Fieller's theorem was originally proposed for estimation of a confidence interval for the ratio of two means from bivariate normal and possibly correlated data.[15] Later, the approach was adapted to allow estimation of ratios of regression coefficients in multivariable regression models.[30,31] Following Cox,[31] a confidence interval for the ratio of linear combinations of regression coefficients

$$\theta = \frac{\boldsymbol{K^T\beta}}{\boldsymbol{L^T\beta}} \tag{6}$$

can be estimated by calculating

$$A = (\boldsymbol{L}^T\hat{\boldsymbol{\beta}})^2 - z^2_{1-\alpha/2}\boldsymbol{L^T I^{-1} L}, \tag{7}$$

$$B = 2\left(z^2_{1-\alpha/2}\boldsymbol{K^T I^{-1} L} - (\boldsymbol{K^T}\hat{\boldsymbol{\beta}})(\boldsymbol{L^T}\hat{\boldsymbol{\beta}})\right), \tag{8}$$

and

$$C = (\boldsymbol{K^T}\hat{\boldsymbol{\beta}})^2 - z^2_{1-\alpha/2}\boldsymbol{K^T I^{-1} K}, \tag{9}$$

with $\boldsymbol{I^{-1}}$ representing the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

Confidence interval limits can be determined by solving the quadratic equation

$$A\theta^2 + B\theta + C = 0, \tag{10}$$

providing an interval of the form

$$100(1-\alpha)\%\ \text{ci} = \left[-B - \frac{\sqrt{B^2 - 4AC}}{2A}\ \text{ to }\ -B + \frac{\sqrt{B^2 - 4AC}}{2A}\right] \tag{11}$$

if $A > 0$ (which implies $B^2 - 4AC > 0$).[32]

If $A < 0$ and $B^2 - 4AC > 0$, the confidence interval will be the complement of a finite interval

$$100(1-\alpha)\%\ \text{ci} = \left(-\infty\ \text{ to }\ -B + \frac{\sqrt{B^2 - 4AC}}{2A}\right. \quad \& \quad \left.-B - \frac{\sqrt{B^2 - 4AC}}{2A}\ \text{ to }\ \infty\right). \tag{12}$$

For $A < 0$ and $B^2 - 4AC < 0$, the confidence interval is

$$100(1-\alpha)\%\ \text{ci} = (-\infty\ \text{ to }\ \infty). \tag{13}$$

For our question of interest, the vectors $\boldsymbol{K}$ and $\boldsymbol{L}$ have to be chosen appropriately to represent the relationship described in Equations (4) and (5).

Model I: The negative ratio of the estimated regression coefficient for treatment $\hat{\beta}_G$ and the coefficient for the interaction between treatment and the covariate of interest $\hat{\beta}_{G\times X}$ has to be considered. Consequently, for $\hat{\boldsymbol{\beta}} = (\hat{\beta}_G, \hat{\beta}_X, \hat{\beta}_{G\times X})^T$, up to constant factors, $\boldsymbol{K}$ and $\boldsymbol{L}$ have to be specified as

$$\boldsymbol{K} = -(1, 0, 0)^T \tag{14}$$

and

$$\boldsymbol{L} = (0, 0, 1)^T. \tag{15}$$

Model II: The same ratio as for Model I has to be calculated. The covariates $Z_1, \ldots, Z_k$ and their according regression coefficients are not directly used in the confidence interval estimation, but it has to be considered that $\hat{\beta}_G$, $\hat{\beta}_{G\times X}$ and the estimated variance-covariance matrix are from a model considering $Z_1, \ldots, Z_k$ as covariates and consequently other results will be obtained compared to the confidence interval for Model I described above. Thus, for $\hat{\boldsymbol{\beta}} = (\hat{\beta}_G, \hat{\beta}_X, \hat{\beta}_{G\times X}, \hat{\beta}_{Z_1}, \ldots, \hat{\beta}_{Z_k})^T$, the vectors $\boldsymbol{K}$ and $\boldsymbol{L}$ have to be chosen as

$$\boldsymbol{K} = -(1, 0, 0, 0, \ldots, 0)^T \tag{16}$$

and

$$\boldsymbol{L} = (0, 0, 1, 0, \ldots, 0)^T. \tag{17}$$

Model III: The estimated changepoint and consequently the estimated confidence interval will depend on the values of $Z_1, \ldots, Z_l$. With $\hat{\boldsymbol{\beta}} = (\hat{\beta}_G, \hat{\beta}_X, \hat{\beta}_{G\times X}, \hat{\beta}_{Z_1}, \ldots, \hat{\beta}_{Z_k}, \hat{\beta}_{G\times Z_1}, \ldots, \hat{\beta}_{G\times Z_l})^T$, up to constant factors, the vectors $\boldsymbol{K}$ and $\boldsymbol{L}$ have to be chosen as

$$\boldsymbol{K} = -(1, 0, 0, 0, \ldots, 0, z_1, \ldots, z_l)^T \tag{18}$$

and

$$\boldsymbol{L} = (0, 0, 1, 0, \ldots, 0, 0, \ldots, 0)^T, \tag{19}$$

where z_1 to z_l represent the values of interest for the l further predictive variables, at which the confidence interval for the *changepoint of treatment stratification* is to be estimated.

As a finite interval is obtained if $A > 0$, which only contains the vector $\boldsymbol{L}$ but not $\boldsymbol{K}$ and consequently does not depend on the chosen values of Z_1 to Z_l, a finite confidence interval is either obtained for all combinations of z_1 to z_l or for none.

An alternative representation for Model I based on the original proposal replacing means of normally distributed variables by regression coefficients can be found in the supplemental material (Section S1.1). The resulting confidence interval for the *changepoint of treatment stratification* corresponds to the cutpoints of a pointwise $100(1-\alpha)\%$ confidence interval around the log-hazard ratio in dependence of the covariate values (sometimes called the treatment-effect plot[6]) with the line of zero. This relationship is shown in the supplemental material (Section S1.2) and is illustrated for the three different cases using simulated data in Figure S1.

3.2 | Delta method

The delta method[17] can be used for calculation of a variance-covariance estimator for transformations of maximum likelihood estimates (see, eg, Davison[18]). Generally, the estimator for the variance-covariance matrix of a transformation $g(\hat{\boldsymbol{\theta}})$ of a parameter vector $\hat{\boldsymbol{\theta}}$ can be obtained as

$$\widehat{\text{cov}}\left(g(\hat{\boldsymbol{\theta}})\right) = D(\hat{\boldsymbol{\theta}})^T \widehat{\text{cov}}(\hat{\boldsymbol{\theta}}) D(\hat{\boldsymbol{\theta}}), \tag{20}$$

where $D(\hat{\boldsymbol{\theta}})$ contains the partial derivatives of $g(\hat{\boldsymbol{\theta}})$.

Models I and II: The relevant parameter vector is $\hat{\boldsymbol{\theta}} = (\hat{\beta}_G, \hat{\beta}_{G\times X})^T$ and the relevant transformation is $g(\hat{\boldsymbol{\theta}}) = -\hat{\beta}_G/\hat{\beta}_{G\times X}$ (see Equation (4)). Following the delta method, the partial derivatives of $g(\hat{\boldsymbol{\theta}})$ have to be calculated. Here, the vector of the partial derivatives is

$$D(\hat{\boldsymbol{\theta}}) = \left(-\frac{1}{\hat{\beta}_{G\times X}}, \frac{\hat{\beta}_G}{\hat{\beta}^2_{G\times X}}\right)^T. \tag{21}$$

It has to be considered that the estimated regression coefficients and the variance-covariance matrix will generally differ for Models I and II.

Model III: The vector $\hat{\boldsymbol{\theta}} = (\hat{\beta}_G, \hat{\beta}_{G\times X}, \hat{\beta}_{G\times Z_1}, \ldots, \hat{\beta}_{G\times Z_l})^T$ and the transformation $g(\hat{\boldsymbol{\theta}}) = -(\hat{\beta}_G + \sum_{j=1}^{l} z_j \beta_{G\times Z_j})/\hat{\beta}_{G\times X}$ have to be considered. The partial derivatives are

$$D(\hat{\boldsymbol{\theta}}) = \left(-\frac{1}{\hat{\beta}_{G\times X}}, \frac{\hat{\beta}_G + \sum_{j=1}^{l} z_j \hat{\beta}_{G\times Z_j}}{\hat{\beta}^2_{G\times X}}, -\frac{z_1}{\hat{\beta}_{G\times X}}, \quad \ldots, -\frac{z_l}{\hat{\beta}_{G\times X}}\right)^T. \tag{22}$$

The variance of $g(\hat{\boldsymbol{\theta}})$ can now be estimated following Equation (20).

For all models, an asymptotic $100(1-\alpha)\%$ confidence interval for $\hat{x}_{\text{cp}}$, assuming normality of $\hat{x}_{\text{cp}}$, is given by

$$100(1-\alpha)\%\ \text{ci} = \left[g(\hat{\boldsymbol{\theta}}) - z_{1-\alpha/2}\sqrt{D(\hat{\boldsymbol{\theta}})^T \widehat{\text{cov}}(\hat{\boldsymbol{\theta}}) D(\hat{\boldsymbol{\theta}})}\ \text{ to }\ g(\hat{\boldsymbol{\theta}}) + z_{1-\alpha/2}\sqrt{D(\hat{\boldsymbol{\theta}})^T \widehat{\text{cov}}(\hat{\boldsymbol{\theta}}) D(\hat{\boldsymbol{\theta}})}\right], \tag{23}$$

where z_q is the $(100\cdot q)$th percentile of the standard normal distribution. The variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is represented by the according components used in the calculation of $g(\hat{\boldsymbol{\theta}})$ of the inverse of the observed Fisher information matrix of the Cox model fitted to the data.

3.3 | Bootstrap

As an alternative to the analytical approaches described above, bootstrap methods[20] investigated in the simulation study and applied to the data collected in the clinical trial are described here.

When a dataset of n patients is considered, for each bootstrap sample, n individuals are drawn with replacement. This procedure is repeated k times, where k should be a large number. Recommendations on the minimum number of bootstrap replications vary depending on the parameters that are to be estimated.[20] It appears that a minimum number of 1000 replications should be considered when the percentile-based bootstrap, which is described in Section 3.3.1, is to be applied in order to obtain reliable results. For each of the k datasets, a Cox regression model is fitted to the data and the estimated *changepoint of treatment recommendation* $\hat{x}_{\text{cp}}$ is calculated as described in Equations (4) and (5). The estimate obtained for the jth bootstrap sample ($j \in \{1, \dots, k\}$) is denoted as $\hat{b}_j$.

3.3.1 | Percentile-based intervals

In the percentile-based bootstrap approach, limits of the $100(1-\alpha)\%$ confidence interval for $\hat{x}_{cp}$ are given by the $(100 \cdot \alpha/2)$th and the $(100 \cdot (1-\alpha/2))$th percentiles of the approximated changepoint distribution derived from the k bootstrap samples. With $\hat{b}_{(1)}, \hat{b}_{(2)}, \dots, \hat{b}_{(k)}$ denoting the k ordered estimates of the changepoints derived from the bootstrap samples, the $100(1-\alpha)\%$ confidence interval is

$$100(1-\alpha)\%\ \text{ci} = \left[\hat{b}_{(k\cdot\alpha/2)} \quad \text{to} \quad \hat{b}_{(k\cdot(1-\alpha/2))}\right]. \tag{24}$$

3.3.2 | Normal intervals

For the normal interval approach, the estimated changepoint of treatment stratification is derived as mean of the changepoints obtained in the bootstrap samples

$$\hat{x}_{\text{cp}}^{\text{boot}} = \frac{1}{k}\sum_{j=1}^{k}\hat{b}_j. \tag{25}$$

The standard error for the estimate is calculated as standard deviation of the changepoints obtained from the k samples

$$\hat{\sigma}^{\text{boot}} = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}\left(\hat{b}_j - \hat{x}_{\text{cp}}^{\text{boot}}\right)^2}. \tag{26}$$

Under the assumption of an asymptotic normal changepoint estimate, the asymptotic $100(1-\alpha)\%$ confidence interval can now be derived as

$$100(1-\alpha)\%\ \text{ci} = \left[\hat{x}_{\text{cp}}^{\text{boot}} - z_{1-\alpha/2}\,\hat{\sigma}^{\text{boot}} \quad \text{to} \quad \hat{x}_{\text{cp}}^{\text{boot}} + z_{1-\alpha/2}\,\hat{\sigma}^{\text{boot}}\right]. \tag{27}$$

3.3.3 | Wild bootstrap

The wild bootstrap is a resampling procedure which has been proposed by Wu[21] in the context of linear regression analyses. The idea was to introduce random centered weights to the residuals from which the bootstrapped response variables are derived. As Wu[21] wrote, the wild bootstrap benefits from a "bias-robustness against error variance heteroscedasticity." In this sense, the wild bootstrap seems to be a particularly reasonable choice for Cox regression models in survival analysis because different covariate measurements and unobserved patient heterogeneity involve heteroscedasticity among the individuals.

A variant of the wild bootstrap has been developed for resampling in Cox models with right-censored survival data.[33] In this framework, where the patient-related counting process $t \mapsto N(t)$ for the event of interest admits a martingale structure through a Doob-Meyer decomposition $M = N - \Lambda$, the martingale increments $\mathrm{d}M = \mathrm{d}N - \mathrm{d}\Lambda$ take the role of unobservable error terms. The, so to say, bootstrapped residuals are now given as $\xi \mathrm{d}N$, where the multiplier ξ has a standard normal distribution. It was later shown that the multipliers can have any distribution with zero mean and unit variance.[34,35] Another useful feature is that the $\xi \mathrm{d}N$ are again martingale increments.[35]

To apply the wild bootstrap for Model I, we first note that, for large sample sizes n, $(\hat{\beta}_X - \beta_X, \hat{\beta}_G - \beta_G, \hat{\beta}_{G\times X} - \beta_{G\times X})^T = [\frac{1}{n} I_n(\hat{\beta}_X, \hat{\beta}_G, \hat{\beta}_{G\times X})]^{-1} \frac{1}{n} U_n(\beta_X, \beta_G, \beta_{G\times X}) + o_p(n^{-1/2})$, where U_n is the score function and I_n is the negative of its gradient with respect to the parameter vector.[36] Now, to obtain a wild bootstrap version of this, we replace all martingale increments dM in the score function U_n by ξdN, with independent multipliers ξ for the counting processes of different individuals. Moreover, we replace in I_n the counting processes dN with $\xi^2 dN$; see Dobler and Pauly[37] for a similar approach in the context of nonparametric cumulative incidence functions. It was shown that using these squared multipliers corresponds to using the optional variation process of the bootstrapped processes when considered as martingales in time.[35] We denote the resulting wild bootstrap version of $(\hat{\beta}_X - \beta_X, \hat{\beta}_G - \beta_G, \hat{\beta}_{G\times X} - \beta_{G\times X})^T$ by $(\widehat{W}_X, \widehat{W}_G, \widehat{W}_{G\times X})^T$.

The confidence intervals for the changepoint, which are based on the wild bootstrap, are of a particularly simple form. Instead of the standard normal quantile $z_{1-\alpha/2}$ we use the $(1-\alpha)$-quantile $q^{\text{wild}}_{1-\alpha}$ of $\left| D(\hat{\theta})^T (\widehat{W}_G, \widehat{W}_{G\times X})^T \right|$, for which the original survival and covariate data are considered as fixed values. Here, the gradient $D(\hat{\theta})^T$ again results from the delta-method applied to the changepoint functional g (see Section 3.2). The wild bootstrap approach makes a separate estimation of variances and covariances superfluous, which is why the $1-\alpha$ confidence interval equals

$$100(1-\alpha)\%\text{ci} = \left[-\frac{\hat{\beta}_G}{\hat{\beta}_{G\times X}} - q^{\text{wild}}_{1-\alpha} \text{ to } -\frac{\hat{\beta}_G}{\hat{\beta}_{G\times X}} + q^{\text{wild}}_{1-\alpha} \right]. \tag{28}$$

Intervals for Model II and Model III can be estimated accordingly.

4 | SIMULATION STUDY

A simulation study covering different relevant aspects, which might have relevant influence on the performance of the presented methods, was performed in order to compare the quality of different approaches for confidence interval estimation. The scenarios differed in sample size, censoring distribution, distribution of the covariate of interest, location of the true *changepoint of treatment stratification*, strength of association between further covariates and outcome, strength of interaction between further predictive covariates and treatment, and the correlation structure between considered covariates. A detailed description of the scenarios investigated for Model I, Model II, and Model III is given below. An overview over all considered scenarios can be found in the supplemental material in Tables S5, S6, and S7. The quality of the estimated confidence intervals was measured by their resulting width and coverage. For each scenario, 2000 simulation runs were performed.

As the data are assumed to be collected in a randomized clinical trial with two treatment arms, group allocation was performed randomly with equal probability for both groups for each individual ($P(G = 0) = P(G = 1) = 0.5$). Event times were drawn following the specification of a Cox regression model as shown in Equations (1) to (3) with a constant baseline hazard and regression coefficients as defined below.

4.1 | Model I

Sample size: Different sample sizes were investigated in order to assess performance of the methods in small, moderate, and large samples. Sample sizes of 200, 500, 1000, 2000, and 5000 independent observations were considered.

Covariate distribution: In order to assess the influence of the distribution of the covariate X, covariates were simulated following

- A standard normal distribution: $X \sim N(0; 1)$;
- A uniform distribution between -0.5 and 0.5: $X \sim U(-0.5; 0.5)$.

Regression coefficients: The regression coefficient for the main effect of the covariate X was chosen to be $\beta_X = \ln(1.25) = 0.223$ for all scenarios, indicating a higher risk of death for individuals with larger covariate values. The regression coefficient for treatment β_G, indicating the group difference for an individual with $X = 0$ was varied in order to obtain scenarios with different true values for the *changepoint of treatment stratification* (see description below). The regression coefficient for the interaction between the covariate X and treatment G was chosen to be

- $\beta_{G\times X} = \ln(1.4) = 0.336$ for scenarios with a normally distributed covariate;
- $\beta_{G\times X} = \ln(3.2) = 1.163$ for scenarios with a uniformly distributed covariate.

in order to have a probability (power) of about 90% to observe a statistically significant covariate-treatment interaction for scenarios with a sample size of 500 individuals and a low amount of censored observations. Consequently, the statistical power for detection of a covariate-treatment interaction was smaller for scenarios with a lower number of observed events and higher for scenarios with larger event numbers.

Location of the changepoint: In order to assess the impact of the location of the changepoint in the covariate distribution, different scenarios with the true changepoint located at

- the median (50th percentile)
- the 70th percentile
- the 90th percentile

of the covariate distribution were considered. These different settings were achieved by choosing the regression coefficient for treatment accordingly. Regression coefficients for treatment G of $\beta_G = -z_q\beta_{G\times X}$ were used for scenarios with normally distributed X, where z_q is the $(100{\cdot}q)$th percentile of the standard normal distribution for the desired changepoint location. For scenarios with uniformly distributed X, β_G was chosen to be $\beta_G = -u_q\beta_{G\times X}$, where u_q is the $(100 \cdot q)$th percentile of a uniform distribution with minimum -0.5 and maximum 0.5.

Censoring distribution: Censoring time distributions were chosen to generate event time data with different proportions of censored observations.

- Low to moderate amount of censored observations of about 25% (leading to about 150, 375, 750, 1500, or 3750 expected events for the different sample sizes) were simulated by drawing censoring times from an exponential distribution with a hazard rate of $\lambda_{\text{cens.}} = 0.3$.
- High amount of censored observations of about 70% (translating to about 60, 150, 300, 600, or 1500 expected events) were simulated by drawing censoring times from an exponential distribution with a hazard rate of $\lambda_{\text{cens.}} = 2.2$.

If the generated censoring time for an individual was smaller than the generated event time, the individual was considered as a censored observation. The shorter time was allocated as observed time.

4.2 | Model II

Simulations were performed in order to derive whether inclusion of further prognostic variables can increase the performance of the estimation procedures. In these simulations, two additional prognostic variables Z_1 and Z_2 ($k = 2$) were considered. Properties of confidence interval estimators with and without consideration of Z_1 and Z_2 in the regression model used for estimation of the regression coefficients, which are considered for estimation of the *changepoint of treatment stratification*, were compared. Only analytical approaches (Fieller's and delta method) were considered in these simulations. Covariates were drawn from a multivariable normal distribution with means of zero and variances of one for X, Z_1, and Z_2. Covariances between the variables were set as described below. For all simulations, the regression coefficients for the covariate of interest X, treatment G, and their interaction were set to $\beta_X = \ln(1.25)$, $\beta_G = 0$, and $\beta_{G\times X} = \ln(1.4)$. The following aspects were varied in order to derive their influence on performance of the estimators.

Sample size: Datasets with 200, 1000, and 5000 observations were considered.

Censoring distribution: Exponentially distributed censoring times were generated leading to scenarios with

- Low censoring: about 25% censored observations ($\lambda_{\text{cens.}} = 0.3$);
- High censoring: about 70% censored observations ($\lambda_{\text{cens.}} = 2.5$).

Correlation between covariates: Covariates (X, Z_1 and Z_2) were drawn from multivariate normal distributions with two different variance-covariance matrices Σ_0 and Σ_1 giving either independent covariates or covariates with moderate pairwise correlations ($r = 0.5$)

$$\Sigma_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \Sigma_1 = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}. \tag{29}$$

Strength of association between the prognostic variables Z_1 and Z_2 and the outcome: To investigate the influence of the strength of association between the prognostic variables Z_1 and Z_2 and the event time of interest, two different settings for the according regression coefficients were chosen with

- $\beta_{Z_1} = \beta_{Z_2} = \ln(1.2)$;
- $\beta_{Z_1} = \beta_{Z_2} = \ln(1.5)$.

4.3 | Model III

Simulations were also performed for scenarios considering Model III, ie, in the presence of further predictive variables that also interact with treatment. As for the simulations for Model II, different scenarios in the presence of two further covariates Z_1 and Z_2 were investigated ($k = l = 2$). As described in Equation (5), the *changepoint of treatment stratification* depends on the values of Z_1 and Z_2 in this situation. For all simulations investigating properties of confidence interval estimators under Model III, the regression coefficients for X, G, and their interaction were set to $\beta_X = \ln(1.25)$, $\beta_G = 0$, and $\beta_{G\times X} = \ln(1.4)$. The coefficients for Z_1 and Z_2 were chosen to be $\beta_{Z_1} = \beta_{Z_2} = \ln(1.25)$. Censoring times were drawn from an exponential distribution with a hazard rate of $\lambda_{\text{cens.}} = 0.3$, leading to a mean proportion of censored observations of about 25%. The following aspects were varied in the simulation study.

Sample size: Datasets with 200, 1000, and 5000 observations were considered.

Correlation between covariates: Covariates were chosen to be either independent or moderately correlated (Σ_0, Σ_1) as described in Section 4.2.

Strength of interaction between Z_1 and Z_2 and treatment: Two situations with smaller and larger interaction effects between the two predictive variables Z_1 and Z_2 and treatment were considered with

- $\beta_{G\times Z_1} = \beta_{G\times Z_2} = \ln(1.2)$;
- $\beta_{G\times Z_1} = \beta_{G\times Z_2} = \ln(1.5)$.

4.4 | Results of the simulation study

4.4.1 | Model I

All methods described in Section 3 were applied to the generated data. For the bootstrap methods, 1000 bootstrap samples were drawn. In order to evaluate and compare the confidence interval approaches, different measures are provided. For each method, the coverage, ie, the proportion of simulation runs for which the true changepoint is covered by the confidence interval,[23] was calculated and is presented in Figures 1 and 2 and tabulated in the appendix (Tables A1 and A2). Moreover, the width of the interval, ie, the difference between the upper limit and the lower limit, was derived. Scatterplots showing combinations of confidence interval coverages and median confidence interval widths are presented in Figure 1 (for a standard normally distributed covariate) and Figure 2 (for a uniformly distributed covariate) for all methods stratified for different scenarios (true location of the changepoint, proportion of censored observations). Different sample sizes are indicated by different symbols within the according figures, and different methods are represented by different colors of symbols and lines. As a confidence interval with a width of more than four was considered to be uninformative, when the covariate of interest follows a standard normal distribution, median confidence interval widths larger than four are all presented as ">4" in Figure 1. For scenarios with a uniformly distributed X, median confidence interval widths larger than one are presented as ">1" in Figure 2. For a better visual comparability, median width is presented on a logarithmic scale. An overview over the results obtained for all scenarios stratified by method is presented in Figure 3.

Observed median confidence interval widths with 10th and 90th percentile are also tabulated for each method stratified by scenario. These tables are presented in the supplemental material (Tables S1 and S2). Additionally, the number of estimated confidence intervals using Fieller's approach with one or both limits being plus or minus infinity is given in Table 1.

Confidence intervals estimated based on Fieller's theorem (green symbols and lines in the Figures) provided a good confidence interval coverage close to 95% for all investigated scenarios irrespective of sample size and true location of the changepoint (observed coverage proportions between 93.6% and 96.0%, Tables A1 and A2). This comes at the cost of very wide or infinite confidence intervals, especially in the scenarios with small sample sizes. With a sample size of 200 observations and a high amount of censored observations (about 60 expected events), an infinite confidence interval (ie, at least one infinite confidence interval limit) was observed in about 75% of the simulation runs (Table 1). With expected numbers of 300 to about 375 events, an infinite confidence interval was still obtained in about 10 to 25% of the simulation runs.

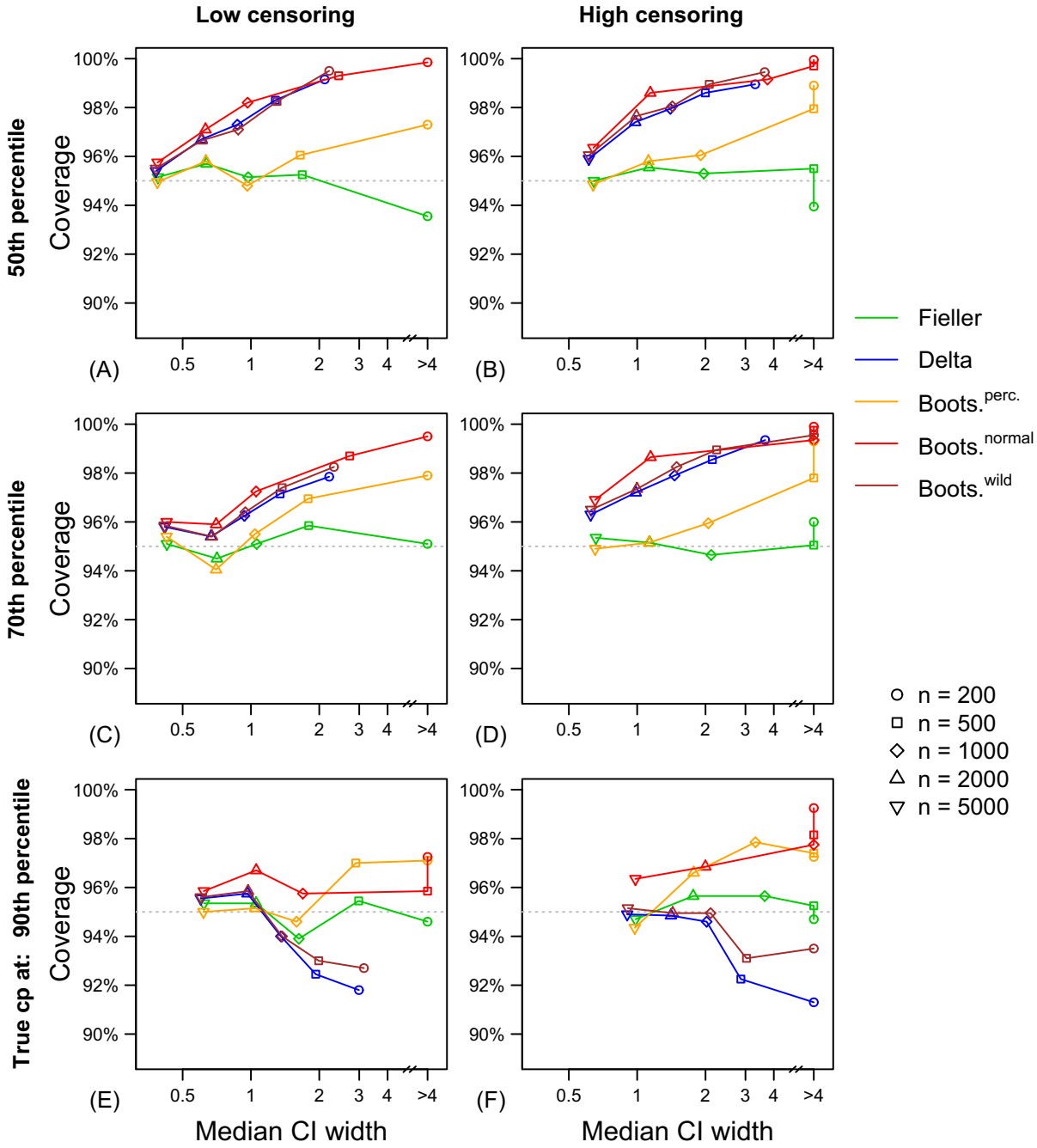

FIGURE 1 Scatter plots illustrating median confidence interval (CI) widths and observed coverage proportions for all methods under investigation for a normally distributed covariate. Scenarios with different true changepoints for treatment stratification are presented in different rows, and columns indicate different censoring distributions. The dotted grey line illustrates the desired level of 95%. Different symbols indicate different sample sizes used in the simulation study, with circles showing results for the smallest sample size ($n = 200$) and inverse triangles for the largest sample size ($n = 5000$). Median confidence interval widths exceeding a value of 4 are presented as ">4" [Colour figure can be viewed at wileyonlinelibrary.com]

With the standard delta method (blue lines and symbols in the Figures), obtained confidence interval widths were smaller than for Fieller's approach, but confidence interval coverage varied tremendously for small sample sizes depending on the true location of the changepoint. For a sample size of 200 observations, a confidence interval coverage of 99.2% was observed for the scenario with a low amount of censored observations and of 99.0% with a high amount of censored observation, when the true changepoint was located at the median of the covariate distribution (normally distributed covariate). When the true changepoint was located at the 90th percentile, observed coverage proportions were 91.8% (low censoring) and 91.3% (high censoring). This was caused by the facts that standard errors of the changepoint estimate were

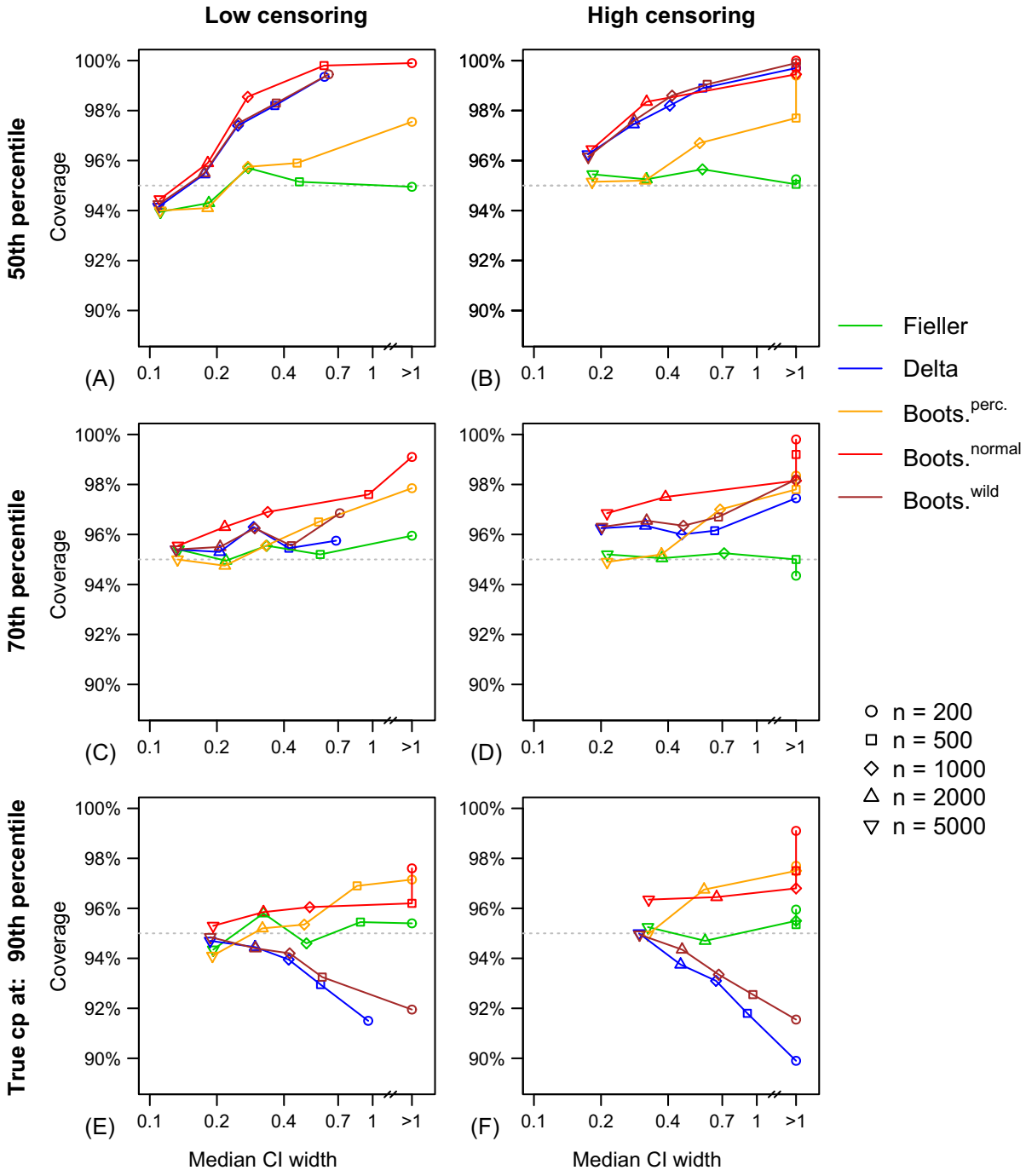

FIGURE 2 Scatter plots illustrating median confidence interval (CI) widths and observed coverage proportions for all methods under investigation for a uniformly distributed covariate. Scenarios with different true changepoints for treatment stratification are presented in different rows, and columns indicate different censoring distributions. The dotted grey line illustrates the desired level of 95%. Different symbols indicate different sample sizes used in the simulation study, with circles showing results for the smallest sample size ($n = 200$) and inverse triangles for the largest sample size ($n = 5000$). Median confidence interval widths exceeding a value of 1 are presented as ">1" [Colour figure can be viewed at wileyonlinelibrary.com]

overestimated by the delta method in scenarios with small to moderate numbers of observed events and that the delta method always provides symmetrical confidence intervals, but the distribution of estimated changepoints was skewed in scenarios with the true changepoint not located at a value of zero (ie, not at the mean and median of the covariate distribution).

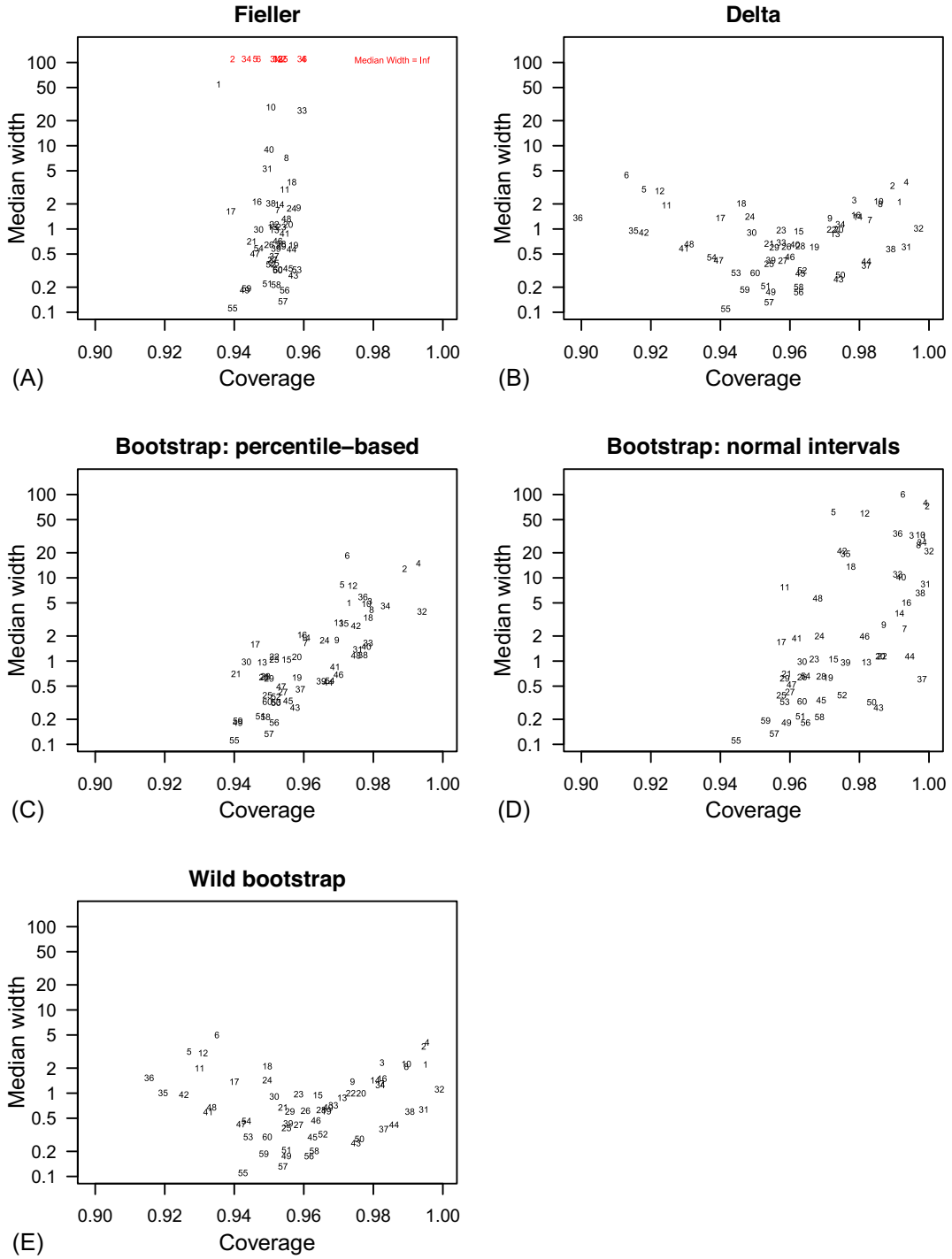

FIGURE 3 Summary of all observed confidence interval coverages and median widths stratified by estimation method. Scenarios are indicated by numbers as presented in the supplemental Table S5. For Fieller's approach (A), scenarios with an observed infinite median confidence interval width are illustrated by red numbers [Colour figure can be viewed at wileyonlinelibrary.com]

For the percentile-based bootstrap, observed confidence interval coverage was above 95% for all scenarios with 200 or 500 observations with a maximum observed coverage of 99.4%. For all scenarios with at least 500 expected events (at least 1000 observations with low amount of censoring or at least 2000 observations with high amount of censoring), observed coverage proportions were between 94.0% and 96.8%. Median confidence interval widths were smaller than those obtained following Fieller's theorem but larger than those based on the delta method.

TABLE 1 Proportion of infinite intervals obtained with Fieller's approach in the different scenarios

	Covariate distribution	True loc. of x_{cp}	β_Z	$\beta_{G\times Z}$	Σ	Cens.	Sample size 200	500	1000	2000	5000
Model I	normally	50th	—	—	—	low	49.4%	11.7%	0.5%	0.0%	0.0%
	normally	50th	—	—	—	high	74.0%	43.4%	15.6%	1.1%	0.0%
	normally	70th	—	—	—	low	50.5%	12.6%	0.4%	0.0%	0.0%
	normally	70th	—	—	—	high	76.7%	48.8%	20.6%	1.8%	0.0%
	normally	90th	—	—	—	low	51.4%	13.5%	0.8%	0.0%	0.0%
	normally	90th	—	—	—	high	77.9%	51.4%	22.6%	2.2%	0.0%
	uniformly	50th	—	—	—	low	47.8%	11.2%	0.6%	0.0%	0.0%
	uniformly	50th	—	—	—	high	74.9%	44.0%	18.0%	1.6%	0.0%
	uniformly	70th	—	—	—	low	49.4%	10.0%	0.4%	0.0%	0.0%
	uniformly	70th	—	—	—	high	77.7%	48.4%	20.8%	1.9%	0.0%
	uniformly	90th	—	—	—	low	52.8%	10.9%	0.6%	0.0%	0.0%
	uniformly	90th	—	—	—	high	78.4%	53.2%	25.0%	2.8%	0.0%
Model II (not considering Z_1 and Z_2)	normally	—	ln(1.2)	—	Σ_0	low	50.9%	—	1.2%	—	0.0%
	normally	—	ln(1.2)	—	Σ_0	high	76.9%	—	20.3%	—	0.0%
	normally	—	ln(1.2)	—	Σ_1	low	52.2%	—	1.0%	—	0.0%
	normally	—	ln(1.2)	—	Σ_1	high	76.8%	—	21.4%	—	0.0%
	normally	—	ln(1.5)	—	Σ_0	low	59.4%	—	4.2%	—	0.0%
	normally	—	ln(1.5)	—	Σ_0	high	78.0%	—	25.2%	—	0.0%
	normally	—	ln(1.5)	—	Σ_1	low	61.6%	—	3.9%	—	0.0%
	normally	—	ln(1.5)	—	Σ_1	high	78.6%	—	27.0%	—	0.0%
Model II (considering Z_1 and Z_2)	normally	—	ln(1.2)	—	Σ_0	low	48.4%	—	0.6%	—	0.0%
	normally	—	ln(1.2)	—	Σ_0	high	76.3%	—	19.2%	—	0.0%
	normally	—	ln(1.2)	—	Σ_1	low	49.4%	—	1.0%	—	0.0%
	normally	—	ln(1.2)	—	Σ_1	high	76.3%	—	19.6%	—	0.0%
	normally	—	ln(1.5)	—	Σ_0	low	48.4%	—	1.0%	—	0.0%
	normally	—	ln(1.5)	—	Σ_0	high	74.5%	—	19.2%	—	0.0%
	normally	—	ln(1.5)	—	Σ_1	low	53.0%	—	0.9%	—	0.0%
	normally	—	ln(1.5)	—	Σ_1	high	76.4%	—	20.4%	—	0.0%
Model III	normally	—	ln(1.2)	ln(1.2)	Σ_0	low	49.6%	—	0.6%	—	0.0%
	normally	—	ln(1.2)	ln(1.2)	Σ_1	low	61.6%	—	5.2%	—	0.0%
	normally	—	ln(1.2)	ln(1.5)	Σ_0	low	50.6%	—	0.8%	—	0.0%
	normally	—	ln(1.2)	ln(1.5)	Σ_1	low	63.5%	—	5.6%	—	0.0%

The normal bootstrap approach described in Section 3.3.2 provided wide confidence intervals and coverage proportions exceeding the desired level, as standard errors of the changepoint estimate were overestimated by the standard deviation of the bootstrap estimates, especially in scenarios with small to moderate event numbers. The observed coverage proportions were between 97% and 100% for most scenarios with 200, 500, or 1000 observations. The normal bootstrap also performed worst regarding coverage proportion and confidence interval width for scenarios with 5000 observations.

The wild bootstrap approach performed similarly to the delta method for most of the scenarios, with coverage proportions being larger than the desired level for scenarios with the true changepoint located at the 50th or the 70th percentile, when sample sizes were small to moderate. For scenarios with the true changepoint located at the 90th percentile, the wild bootstrap also provided confidence interval coverages smaller than 95%, when sample sizes of 200 or 500 observations were used, but coverage proportions were closer to 95% than those obtained using the delta method (91.6% to 93.5% as compared to 89.9% to 93.0%). Distributions of confidence interval widths were very similar to those obtained from the delta method, with slightly larger median widths observed for confidence intervals derived using the wild bootstrap approach for most of the scenarios.

For all methods and all scenarios, confidence interval coverage approached the desired value of 95% with increasing sample size. With a sample size of 5000 observations and a low amount of censored observations (3750 expected events), observed proportions of confidence interval coverage were between 94.0% and 96.0% for all methods, changepoint locations, and covariate distributions.

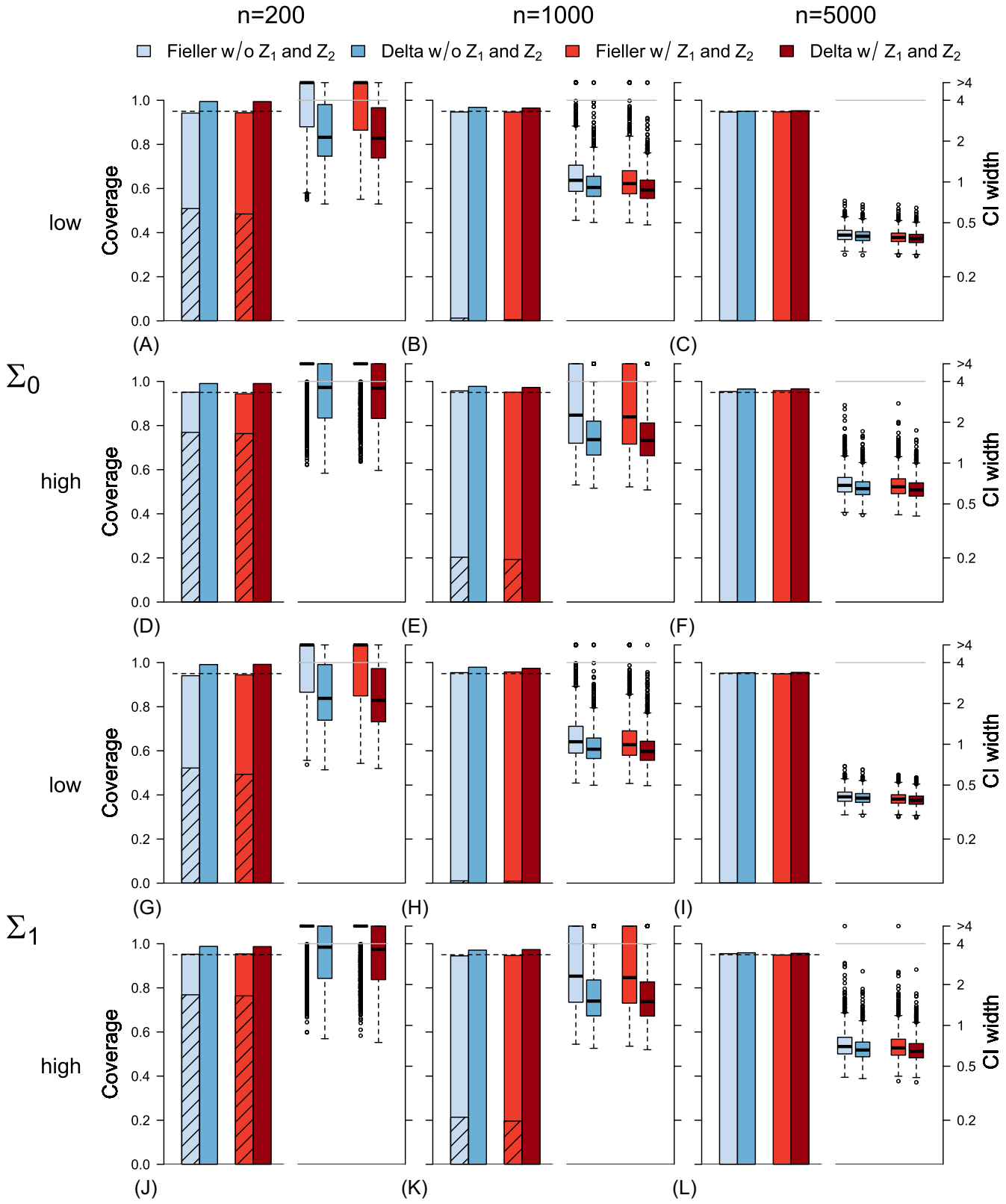

FIGURE 4 Observed confidence interval (CI) coverage (bars on the left side of the figures) and distributions of confidence interval widths (boxplots on the right side) for scenarios with a small effect of Z_1 and Z_2 on the outcome ($\beta_{Z_1} = \beta_{Z_2} = \ln(1.2)$) obtained following Fieller's approach (light colors) or the delta method (dark colors) from models not considering Z_1 and Z_2 as covariates (blue) or from models also considering Z_1 and Z_2 (red). Proportions of confidence intervals of infinite width using Fieller's approach are presented as hatched boxes. Confidence interval widths larger than 4 are displayed as ">4" [Colour figure can be viewed at wileyonlinelibrary.com]

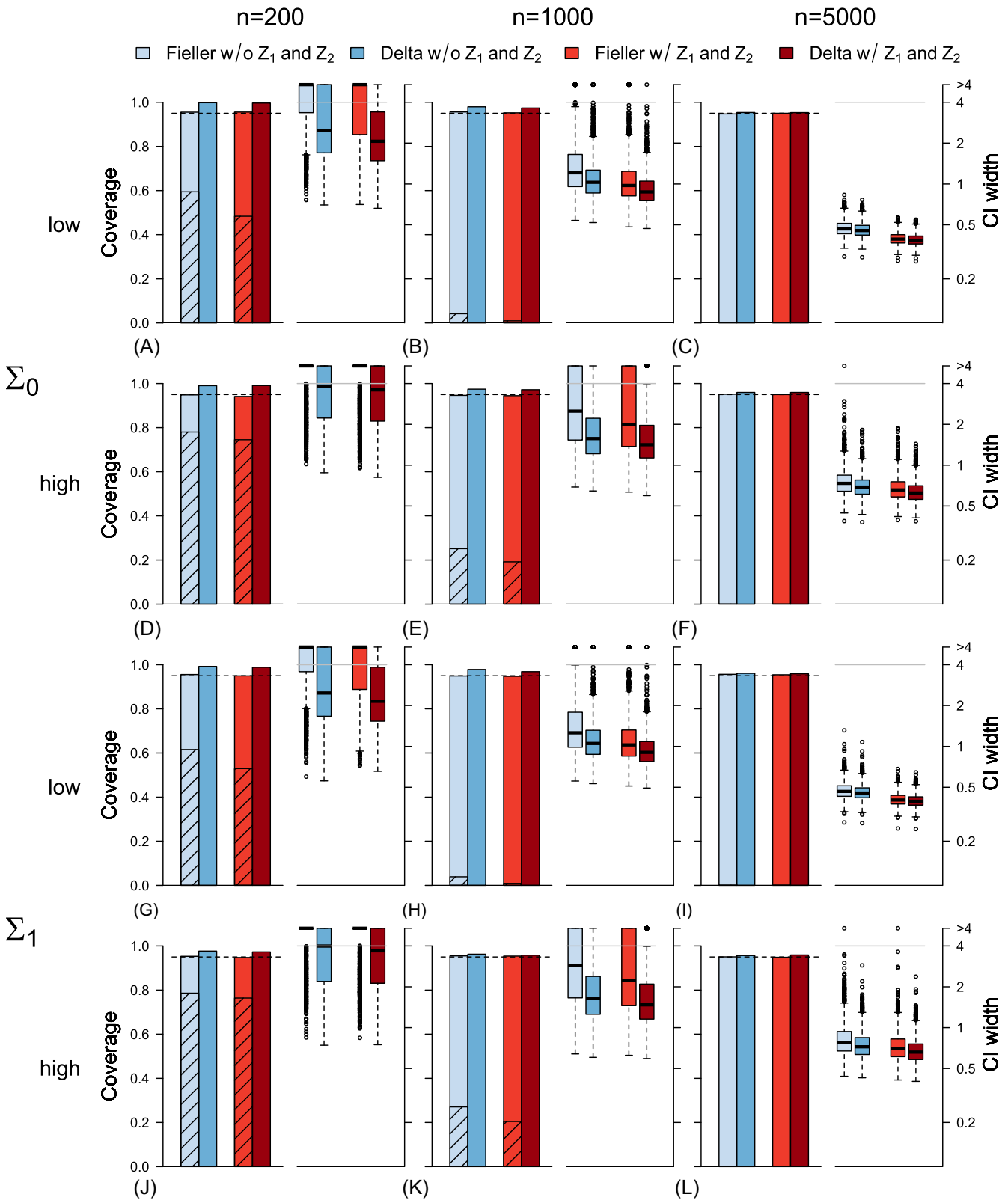

FIGURE 5 Observed confidence interval (CI) coverage (bars on the left side of the figures) and distributions of confidence interval widths (boxplots on the right side) for scenarios with a large effect of Z_1 and Z_2 on the outcome ($\beta_{Z_1} = \beta_{Z_2} = \ln(1.5)$) obtained following Fieller's approach (light colors) or the delta method (dark colors) from models not considering Z_1 and Z_2 as covariates (blue) or from models also considering Z_1 and Z_2 (red). Proportions of confidence intervals of infinite width using Fieller's approach are presented as hatched boxes. Confidence interval widths larger than 4 are displayed as ">4" [Colour figure can be viewed at wileyonlinelibrary.com]

4.4.2 | Model II

The results of the simulations for Model II are illustrated in Figures 4 and 5. Observed coverage proportions for confidence intervals based on regression coefficients not considering the prognostic variables Z_1 and Z_2 (blue bars) and based on regression models obtained from models also considering the effects of Z_1 and Z_2 (red bars) are illustrated for Fieller's approach (light colors) and the delta method (dark colors). The desired coverage level of 95% is illustrated by a dashed line. Obtained proportions of infinite confidence intervals when Fieller's approach was used are visualized by the hatched boxes within the bars indicating the coverage proportions and are tabulated in Table 1. Moreover, the distribution of confidence interval widths is presented by boxplots (on a logarithmic scale given on the right side of each Figure). As described for Model I, obtained widths larger than four are summarized as ">4." Results are presented stratified for strength of association between Z_1 and Z_2 and outcome (weak effect of Z_1 and Z_2 on the outcome in Figure 4, strong effect in Figure 5), sample size (columns), censoring distribution (rows), and correlation structure of the covariates (top and bottom half of the Figures).

Under the considered settings, coverage proportions of confidence intervals based on Fieller's approach were very close to the desired level of 95% irrespective of inclusion of Z_1 and Z_2 into the regression model, with all portions ranging between 94.0% and 95.7% when Z_1 and Z_2 were not considered and 94.2% and 95.8% when Z_1 and Z_2 were included as predictors in the regression model as given in Equation (2). The proportion of infinitely wide confidence intervals obtained by the use of Fieller's approach was up to 78.6% for the scenarios with a sample size of 200 when Z_1 and Z_2 were not considered and up to 76.4%, else. For large sample sizes, infinitely wide confidence intervals were observed less often (up to 27.0% and 20.4% for $n = 1000$ with and without inclusion of Z_1 and Z_2 and none for $n = 5000$). For the delta method, coverage proportions were slightly too large for the sample sizes of $n = 200$ and $n = 1000$, irrespective of consideration of Z_1 and Z_2 (between 97.6% and 99.8% for models without and between 97.3% and 99.6% for models with inclusion of Z_1 and Z_2 in the regression model for $n = 200$ and between 96.2% and 98.0% or 95.8% and 97.4% for $n = 1000$, respectively). As also observed for Model I, confidence interval widths were generally smaller for the confidence intervals estimated using the delta method than for those obtained by Fieller's approach. Confidence interval width could be reduced by inclusion of the prognostic variables in the regression model, especially for the scenarios with the stronger effect of Z_1 and Z_2 on the outcome variable, which are shown in Figure 5. Median widths were, eg, 1.03 and 0.972 using Fieller's approach without and with consideration of Z_1 and Z_2 as compared to 0.910 and 0.871 using the delta method when the sample size was $n = 1000$, the amount of censored observations was low, $\beta_{Z_1} = \beta_{Z_2} = \ln(1.2)$, and X, Z_1, and Z_2 were independent (Figure 4B). With a stronger association between Z_1 and Z_2 and the outcome ($\beta_{Z_1} = \beta_{Z_2} = \ln(1.5)$), a sample size of 1000 observations, low amount of censoring and independent covariates, median confidence interval widths were 1.21 (not considering Z_1 and Z_2 in the regression model) and 0.975 (including Z_1 and Z_2 as covariates in the regression model) following Fieller's approach and 1.03 and 0.874 when the delta method was applied. Exact numbers of observed confidence interval coverage and confidence interval widths (medians with 10th and 90th percentile) are presented in the appendix (Table A3) and in the supplemental material (Table S3).

4.4.3 | Model III

In Figures 6 and 7, results of the simulations with additional predictive variables Z_1 and Z_2 that also interact with treatment are presented. The scenarios are described in detail in Section 4.3, and an overview is given in the supplemental material (Table S7). As discussed in Section 2, the estimated changepoint of treatment stratification depends on Z_1 and Z_2. Consequently, the estimated coverage proportions are presented for given values of the covariates Z_1 and Z_2, namely at the 5th, 25th, 50th = median = mean, 75th, and 95th percentile of the theoretical covariate distribution, which was the standard normal distribution in our simulations. In the Figures, corresponding percentiles for Z_1 are given at the x-axis, and percentiles of Z_2 are illustrated in different grey scales, so results for all possible combinations of the given percentiles for Z_1 and Z_2 are shown. Results are presented stratified for sample size (columns, $n = 200$, $n = 1000$, $n = 5000$), method under investigation (rows, estimation following Fieller's approach, and the delta method), correlation between covariates X, Z_1 and Z_2 (Σ_0: top of the Figures, Σ_1: bottom, as defined in Section 4.2), and strength of interaction between Z_1 and Z_2 and treatment ($\beta_{G\times Z_1} = \beta_{G\times Z_2} = \ln(1.2)$: Figure 6, $\beta_{G\times Z_1} = \beta_{G\times Z_2} = \ln(1.5)$: Figure 7). Observed coverage proportions are illustrated by barplots and median confidence interval widths by red dots. Again, median widths larger than four are indicated as ">4." Proportions of confidence intervals of infinite width obtained following Fieller's approach, which do not depend on the chosen values for Z_1 and Z_2 as described in Section 3.1, are presented by hatched boxes and are given in Table 1. Observed confidence interval coverages are also tabulated in the appendix (Table A4). Median confidence interval widths with observed 10th and 90th percentiles are shown in the supplemental material (Table S4).

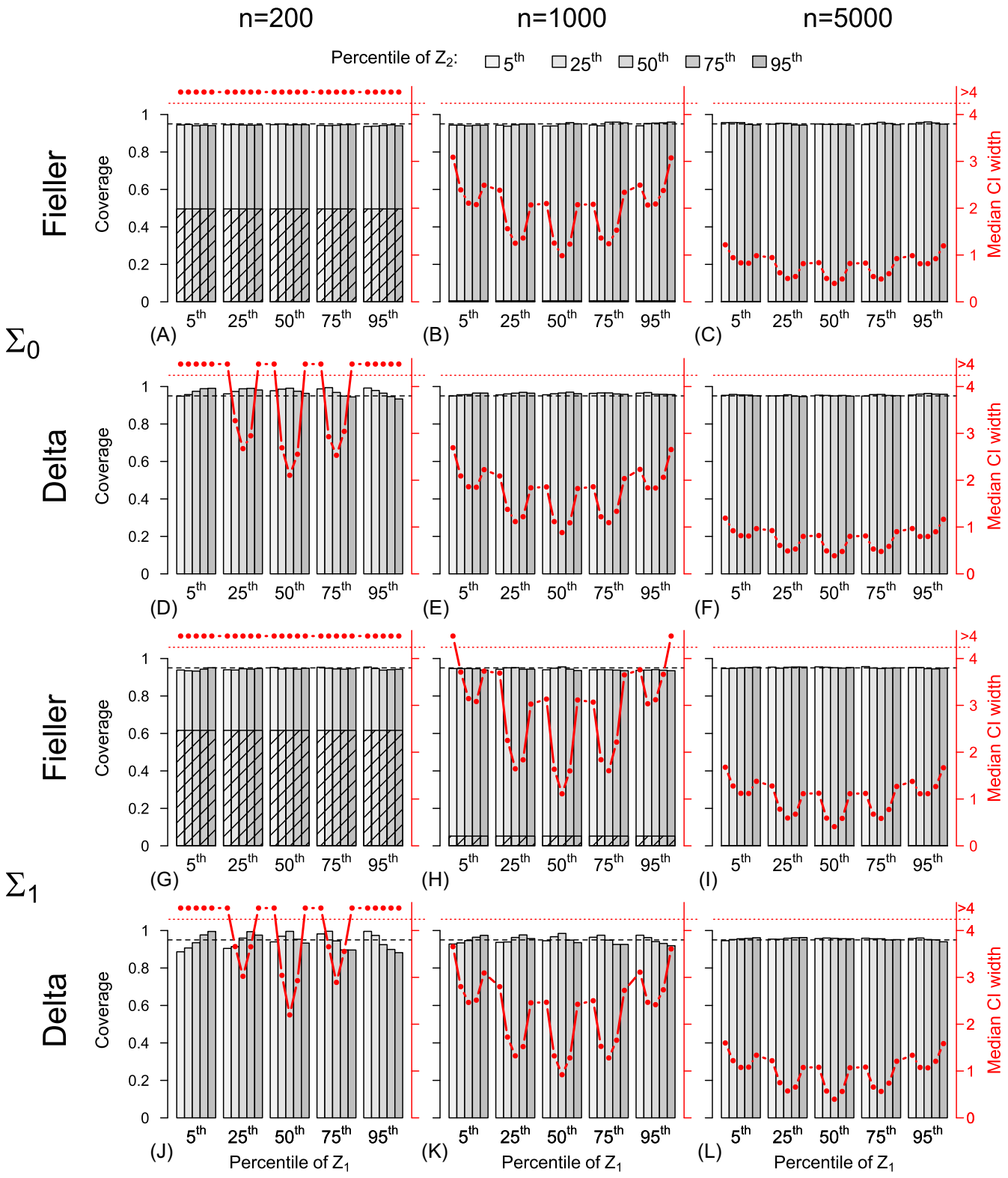

FIGURE 6 Observed confidence interval (CI) coverage proportions (grey bars) and median confidence interval widths (red dots) for scenarios with small interaction effects between covariates Z_1 and Z_2 and treatment ($\beta_{G\times Z_1} = \beta_{G\times Z_2} = \ln(1.2)$). Coverage proportions and median confidence interval widths are displayed for given combinations of covariate values of Z_1 (x-axis) and Z_2 (grey scale), namely for the 5th, 25th, 50th, 75th, and 95th percentile of the theoretical distributions. Proportions of infinite confidence intervals obtained following Fieller's approach are illustrated by hatched boxes. Confidence interval widths larger than 4 are displayed as ">4" [Colour figure can be viewed at wileyonlinelibrary.com]

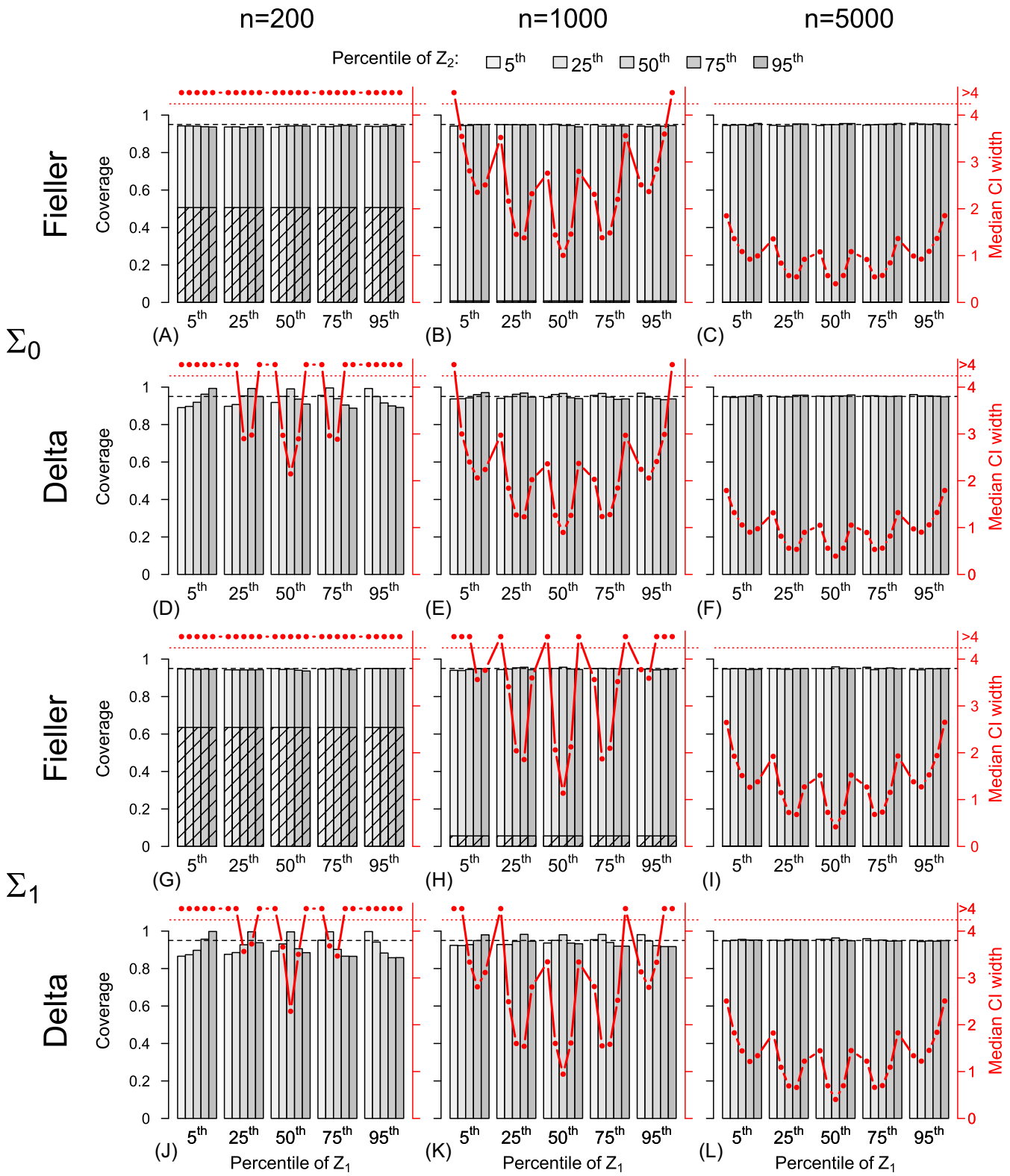

FIGURE 7 Observed confidence interval (CI) coverage proportions (grey bars) and median confidence interval widths (red dots) for scenarios with large interaction effects between covariates Z_1 and Z_2 and treatment ($\beta_{G\times Z_1} = \beta_{G\times Z_2} = \ln(1.5)$). Coverage proportions and median confidence interval widths are displayed for given combinations of covariate values of Z_1 (x-axis) and Z_2 (grey scale), namely for the 5th, 25th, 50th, 75th, and 95th percentile of the theoretical distributions. Proportions of infinite confidence intervals obtained following Fieller's approach are illustrated by hatched boxes. Confidence interval widths larger than 4 are displayed as ">4" [Colour figure can be viewed at wileyonlinelibrary.com]

As observed for the other investigated scenarios, the observed coverage proportions for Fieller's approach were close to the desired level under all scenarios and for all investigated covariate values of Z_1 and Z_2 (observed proportions between 93.2% and 96.0%). For the scenario with $n = 200$, strong interaction between Z_1 and Z_2 and treatment and the correlation structure Σ_1, 63.5% of the estimated confidence intervals were of infinite width. For $n = 1000$, the proportions ranged from 0.6% to 5.6% for the different settings. No confidence intervals of infinite width were observed in simulations with a sample size of $n = 5000$. When the delta method was applied to estimate confidence intervals for the *changepoint of treatment stratification* for X, the observed coverage proportions varied, strongly depending on the values of Z_1 and Z_2, especially for scenarios with the variance-covariance structure Σ_1. While coverage proportions ranged from 93.4% to 99.4% under Σ_0 for $n = 200$ and weak interaction between Z_1 and Z_2 and treatment (Figure 6D), the observed proportions were between 88.2% and 99.6% using Σ_1 (Figure 6J). Large deviations from the desired confidence level of 95% were especially observed for combinations of very small values of Z_1 and Z_2 (5th percentile and 5th percentile) or both very large values (95th percentile and 95th percentile).

As observed in the other simulated scenarios, median confidence interval widths were generally larger for Fieller's approach as compared to the delta method, which was more pronounced for sample sizes of $n = 200$ or $n = 1000$. For both methods and under all settings, median confidence interval widths were smaller when a confidence interval for the changepoint of treatment stratification was to be estimated at the center of the distribution of the predictive covariates Z_1 and Z_2 than at the tails of the distribution, eg, median widths of the observed confidence intervals for a sample size of $n = 1000$, $\beta_{G\times Z_1} = \beta_{G\times Z_2} = \ln(1.5)$, and the variance-covariance structure Σ_0 were 1.00 following Fieller's approach and 0.897 using the delta method, when the interval was estimated at the 50th percentile of Z_1 and Z_2, each. For the same scenario, median widths were 4.88 for Fieller's approach and 4.05 for the delta method, when a confidence interval for the changepoint was estimated at the 95th percentile of Z_1 and Z_2.

5 | APPLICATION

The methods presented in Section 3 and investigated in Section 4 were applied to data from the randomized clinical SPACE trial.[24,25] In the SPACE trial, patients with symptomatic, severe ($\geq$ 70% ECST) carotid artery stenosis in the previous six months were randomly assigned to either carotid artery endarterectomy (CEA) or carotid artery angioplasty with stenting (CAS). The study was originally designed as noninferiority trial and was intended to show noninferiority of CAS as compared to CEA for ipsilateral stroke or death within 30 days, which could not be established in the primary analysis.[24] In the analysis of the two-year follow-up data, a relevant qualitative interaction between treatment and age (dichotomized at an age of 68 years) was found, when time to any stroke or death was analyzed, indicating a higher risk from CEA as compared to CAS for younger patients and a higher risk from CAS as compared to CEA for older patients.[25] This interaction was also observed in other randomized trials comparing these procedures.[38,39]

We applied the different methods for confidence interval estimation for the *changepoint of treatment stratification* to the two-year per protocol data of the SPACE trial considering any stroke or death as events of interest. Overall, 146 events (77 in 573 patients treated with CAS and 69 in 563 patients treated with CEA) were observed. We fitted a Cox regression model, including main effects of treatment (CEA: $G = 0$, CAS: $G = 1$) and age (as continuous variable) as well as their interaction to the data as described in Section 2. The results are shown in Table 2. In Figure 8, the estimated hazard ratio between the treatment groups in dependence of age is illustrated and a pointwise 95% confidence interval, as described in Shen et al,[12] is given (see supplemental material, Section S1.2).

Following Equation (4), the estimate for the changepoint is

$$\hat{x}_{\text{cp}} = -\frac{\hat{\beta}_G}{\hat{\beta}_{G\times\text{Age}}} = -\frac{-3.302}{0.049} = 67.95 \text{ years}, \tag{30}$$

which is at the 49th percentile of the age distribution of the included patients. Estimated confidence boundaries derived from the different approaches are given in Table 3 and are illustrated as colored lines in Figure 8.

TABLE 2 Results of the Cox regression model fitted to the stent-protected angioplasty versus carotid endarterectomy (SPACE) data

	$\hat{\beta}$	$\exp(\hat{\beta})$	$se(\hat{\beta})$	z	p
Treatment (G)	−3.302	0.037	1.476	−2.236	0.025
Age	0.018	1.018	0.014	1.257	0.209
Treatment(G) × Age	0.049	1.050	0.021	2.347	0.019

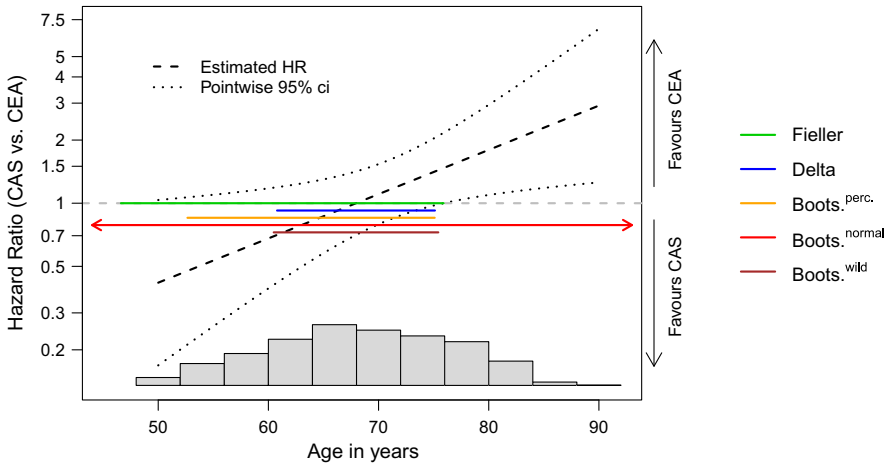

FIGURE 8 Illustration of the estimated treatment effect in dependence of patient's age using a Cox regression model with main effects of treatment and age and their interaction (dashed line). Dotted lines indicate a pointwise 95% confidence interval for the hazard ratio. Colored lines show estimated 95% confidence intervals for the *changepoint of treatment stratification* obtained from the different approaches. The histogram at the bottom illustrates the age distribution in the data. CAS, carotid artery angioplasty with stenting; CEA, carotid artery endarterectomy; HR, hazard ratio [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Confidence intervals for the changepoint derived from the stent-protected angioplasty versus carotid endarterectomy (SPACE) study estimated using Fieller's approach, the delta method, and the bootstrap approaches

	95% conf. int.		
Fieller	46.6	to	75.9
Delta	60.8	to	75.1
Bootstrap (percentile)	52.7	to	75.1
Bootstrap (normal)	-2.8	to	134.6
Bootstrap (wild)	60.5	to	75.4

It can be seen that the normal bootstrap (red line) provides a confidence interval that exceeds the range of the covariate of interest for both limits. The confidence interval based on Fieller's approach (green line) is wider than the interval for the delta method (blue line) and the percentile-based (orange line) and wild bootstrap (brown line). Again, the wild bootstrap provides confidence intervals similar to those obtained by the delta method. The delta method provides the smallest confidence interval, but based on the result of the simulation study, coverage of the delta method may differ relevantly from the desired level of 95%. The percentile-based bootstrap gives a confidence interval with a width between the intervals obtained from Fieller's approach and that derived using the delta method.

6 | DISCUSSION

Stratified or personalized medicine aims to find the best available treatment for each individual patient based on his or her characteristics.[2] Various examples for suggested or established treatment stratification based on, eg, molecular biomarkers exist in the literature, mainly for cancer therapies.[40-42] In order to establish stratified treatment decisions, relevant covariates have to be identified and validated. While some strategies are motivated based on biological plausibility and preclinical data, many predictive biomarkers appear to be identified in retrospective analyses of clinical trial data.[43] In our article, we consider the setting of a two-armed randomized clinical trial, which compares two treatment groups or an experimental treatment versus placebo. Often, a post hoc analysis is performed to assess treatment effect heterogeneity in dependence on one continuous covariate. When a qualitative interaction between the covariate and treatment exists, ie, when not all patients benefit most from the same treatment, treatment allocation should depend on the patient's covariate value, and a cut-off value for treatment stratification has to be determined. We call this cut-off value the *changepoint of treatment stratification*.

In practice, treatment effect heterogeneity is often investigated by estimation of treatment effects in predefined or post hoc–defined subgroups. While this approach is intuitive for categorical covariates of interest, categorization of continuous variables was criticized due to loss of information leading to decreased power for detection of interaction effects and due to biological implausibility.[44,45] While different approaches for classification of patients that respond differently to the therapies were proposed in recent years,[8] we focus on estimation of the cut-off value based on a common regression model, including a covariate-treatment interaction term.[9] As most predictive biomarkers were identified for cancer therapies, we considered a time-to-event outcome and used a Cox regression model for analysis. As the estimated cut-off value is a very relevant quantity derived from clinical trial data, presentation of the estimated value should be accompanied by

an adequate confidence interval indicating uncertainty of the estimate, as it is recommended for estimated treatment effects.[13,14]

We performed a simulation study to investigate behavior of various confidence interval estimators under different scenarios. An adaptation of Fieller's theorem,[15,46] originally proposed for a confidence interval for the ratio of two means from a bivariate normal distribution, the delta method for transformation of maximum likelihood estimates,[16] and various bootstrapping procedures (percentile-based interval, normal interval, wild bootstrap) were investigated regarding coverage probabilities and confidence interval widths considering one predictive covariate of interest as well as in the presence of further prognostic or predictive covariates. Different aspects as the number of observed individuals, the proportion of censored observations, the distribution of the covariate of interest, the location of the true changepoint, the strength of association between further prognostic covariates and the outcome, the strength of interaction between further predictive covariates and treatment, and the correlation between the considered covariates were varied in different scenarios. The simulation study for comparison of confidence interval methods was intended to be performed as a *neutral* study.[47,48] On that note, the main objective of the study was the comparison of methods and not the introduction or promotion of a new method. We also established a team of researchers that are equally experienced with each of the considered methods and do not have preferences for any particular method. Finally, the chosen evaluation criteria are objective, and the compared methods were selected based on a literature research on available approaches.

Results of the simulation study showed that the confidence intervals based on Fieller's theorem provided coverage probabilities close to the desired level of 95% for all scenarios irrespective of the number of observed events and the presence of further covariates, but the approach led to intervals of infinite width for a large number of generated datasets, when the number of observed events was small. The coverage proportions of the often applied delta method were observed to depend heavily on the sample size and the true location of the *changepoint of treatment stratification*. This dependence was caused by the two facts that the delta method provides symmetric confidence intervals, but the changepoint estimate followed a skewed distribution if the true changepoint was not located at the center of the covariate distribution, and that the delta method overestimated the variability of the changepoint estimate in scenarios with a low number of observed events. For large samples, the confidence intervals obtained following Fieller's theorem and the delta method were very similar, and the coverage proportions approximated the desired level of 95%. This was also shown before by Cox.[31] While calculation of a normal confidence interval from bootstrap samples performed inferior to other estimators with regard to confidence interval coverage and width, percentile-based bootstrap confidence intervals were generally conservative for small to moderate numbers of events, ie, confidence interval coverage exceeded the nominal level of 95% for most scenarios, and consequently, intervals were wider than those obtained by other methods. In our investigated scenarios, the wild bootstrap performed very similar to the delta method. When the methods were applied to data from the SPACE study, where only a number of 146 events was observed in 1136 patients, large differences between the confidence intervals obtained by application of the different estimators were observed.

There are several limitations in our simulation study that might relevantly influence the results and consequently our recommendations. Firstly, as it is often the case for simulation studies, only a moderate number of scenarios could be investigated due to limited time and space. We varied the total number of included subjects, the censoring distribution, the distribution of the covariate, and the true location of the changepoint but did not consider, eg, different strengths of interaction or asymmetric covariate distributions in the simulations where only treatment and the covariate of interest were considered. When further covariates, either prognostic or predictive, were included, we only focused on the changepoint of treatment stratification of the predefined variable of interest but did not estimate the changepoint for the other covariates. Moreover, we only investigated the approaches based on Fieller's theorem and the delta method when further covariates were considered. Furthermore, for estimation of the *changepoint of treatment stratification*, we only used the information of one time-to-event endpoint and did not involve further outcome variables (as, eg, quality of life) or costs and risks of the treatment options. The code used for performance of our simulations is provided as online supplemental material, so readers will be able to investigate further scenarios of interest.

We only used a standard Cox regression model with main effects and interaction term, and data were generated fulfilling the common model assumptions. Methods that relax the linearity assumptions[6,7,49] or that rely on classification methods[8,50] were not considered. When the established approach using multivariable fractional polynomials for estimation of nonlinear interactions (MFPI) with one polynomial transformation (FP1) is used, some of the proposed methods, as, eg, the delta method or the bootstrapping approaches, could be adapted easily for estimation of a confidence interval. Adaptation of the delta method and results of some simulations are shown in the online supplementary material to this manuscript (Section S4). This approach appears to work very well, when the true association is covered by the MFPI transformations. As the estimate for the changepoint of treatment stratification will be biased for functional relationships

not covered by the MFPI approach, the confidence interval coverage will decrease when the sample size is increased in that situation (see Figure S2 in the supplemental file). Consequently, in the presence of nonlinear associations and interactions, identification of the correct functional form and unbiased estimation of the changepoint are prerequisites for estimation of an adequate confidence interval. The same holds true for application of spline-based methods, where bootstrapping approaches can be used for confidence interval estimation. Moreover, incorporation of spline functions in the regression models might lead to multiple changepoint estimates. Further research is needed with regard to changepoint and confidence interval estimation in these more complex situations.

We believe that the *changepoint of treatment stratification* is an important quantity that should be estimated and reported, when a qualitative biomarker-treatment interaction was detected, and that a corresponding confidence interval should be presented. Based on the results of our simulation study, confidence interval estimation for the *changepoint of treatment stratification* following Fieller's theorem provided the most reliable results regarding confidence interval coverage but led to infinite intervals in a relevant number of simulation runs when the number of observed events was small to moderate. While all other methods always provide finite intervals, in small sample scenarios, a wide range of the covariate distribution was covered by the confidence intervals, which was also observed in our example using data from a randomized clinical trial, indicating high uncertainty regarding the estimated changepoint. Thus, we recommend application of the approach based on Fieller's theorem for data with small to moderate event numbers. For a large number of observed events, the delta method and the wild bootstrap will also provide confidence intervals with the desired properties but smaller confidence interval widths.

Generally, a large number of observations are necessary in order to precisely estimate the changepoint of treatment stratification from the data collected in randomized clinical trials. If this is one major goal in a certain study, this should be considered adequately in the sample size determination. Sharing of clinical research data will also be important in order to determine the changepoint with an adequate precision and consequently identify the correct patients for administration of a certain treatment. It has to be considered that the changepoint is estimated from a regression model that underlies certain assumptions and consequently is prone to model misspecification.

DATA AVAILABILITY STATEMENT

The R code used for the simulations is available as online supplemental material to this article.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

ORCID

Bernhard Haller https://orcid.org/0000-0002-9723-393X
Ulrich Mansmann https://orcid.org/0000-0002-9955-8906
Dennis Dobler https://orcid.org/0000-0002-9040-0854
Kurt Ulm https://orcid.org/0000-0001-8540-9849
Alexander Hapfelmeier https://orcid.org/0000-0001-6765-6352

REFERENCES

1. Jain K. *Textbook of Personalized Medicine*. New York, NY: Springer; 2016.
2. Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med*. 2010;363(4):301-304.
3. Le Tourneau C, Kamal M, Trédan O, et al. Designs and challenges for personalized medicine studies in oncology: focus on the SHIVA trial. *Target Oncol*. 2012;7(4):253-265.
4. Jürgensmeier JM, Eder JP, Herbst RS. New strategies in personalized medicine for solid tumors: molecular markers and clinical trial designs. *Clin Cancer Res*. 2014;20(17):4425-4435.
5. Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J Biopharm Stat*. 2016;26(1):99-119.

6. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statist Med*. 2004;23(16):2509-2525.
7. Tian L, Alizadeh AA, Gentles AJ, Tibshirani RJ. A simple method for estimating interactions between a treatment and a large number of covariates. *J Am Stat Assoc*. 2014;109(508):1517-1532.
8. Lipkovich I, Dmitrienko A, D'Agostino RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statist Med*. 2017;36(1):136-196.
9. Chen JJ, Lu T-P, Chen Y-C, Lin W-J. Predictive biomarkers for treatment selection: statistical considerations. *Biomark Med*. 2015;9(11):1121-1135.
10. Royston P, Sauerbrei W. *Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Chichester, UK: John Wiley & Sons; 2008.
11. Polley M-Y, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. *J Natl Cancer Inst*. 2013;105(22):1677-1683.
12. Shen Y-M, Le LD, Wilson R, Mansmann U. Graphical presentation of patient-treatment interaction elucidated by continuous biomarkers. *Methods Inf Med*. 2017;56(01):13-27.
13. Lewis JA. Statistical principles for clinical trials (ICH e9): an introductory note on an international guideline. *Statist Med*. 1999;18(15):1903-1942.
14. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Med*. 2010;8(1):18.
15. Fieller EC. The distribution of the index in a bivariate normal distribution. *Biometrika*. 1932;24(3-4):428-440.
16. Armitage P, Colton T. *Encyclopedia of Biostatistics*. Chichester, UK: John Wiley & Sons; 1998.
17. Ver Hoef JM. Who invented the delta method? *Am Stat*. 2012;66(2):124-127.
18. Davison AC. *Statistical Models*. Cambridge, UK: Cambridge University Press; 2003.
19. Efron B. Censored data and the bootstrap. *J Am Stat Assoc*. 1981;76(374):312-319.
20. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press; 1994.
21. Wu C-F. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat*. 1986;14(4):1261-1295.
22. Liu RY. Bootstrap procedures under some non-I.I.D. models. *Ann Stat*. 1988;16(4):1696-1708.
23. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statist Med*. 2006;25(24):4279-4292.
24. Ringleb PA, Allenberg J, Berger J, et al. 30 day results from the SPACE trial of stent-protected angioplasty versus carotid endarterectomy in symptomatic patients: a randomised non-inferiority trial. *Lancet*. 2006;368(9543):1239-1247.
25. Eckstein H-H, Ringleb P, Allenberg J-R, et al. Results of the stent-protected angioplasty versus carotid endarterectomy (SPACE) study to treat symptomatic stenoses at 2 years: a multinational, prospective, randomised trial. *Lancet Neurol*. 2008;7(10):893-902.
26. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
27. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc Series B Stat Methodol*. 1972;34:187-220.
28. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer Science & Business Media; 2000.
29. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: John Wiley & Sons; 2011.
30. Zerbe GO. On Fieller's theorem and the general linear model. *Am Stat*. 1978;32(3):103-105.
31. Cox C. Fieller's theorem, the likelihood and the delta method. *Biometrics*. 1990;46(3):709-718.
32. Buonaccorsi JP. On Fieller's theorem and the general linear model. *Am Stat*. 1979;33(3):162-162.
33. Lin D, Fleming TR, Wei L-J. Confidence bands for survival curves under the proportional hazards model. *Biometrika*. 1994;81(1):73-81.
34. Beyersmann J, Di Termini S, Pauly M. Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scand Stat Theory Appl*. 2013;40(3):387-402.
35. Dobler D, Pauly M, Scheike TH. Confidence bands for multiplicative hazards models: flexible resampling approaches. *Biometrics*. 2019.
36. Aalen O, Borgan O, Gjessing H. *Survival and Event History Analysis: A Process Point of View*. New York, NY: Springer Science & Business Media; 2008.
37. Dobler D, Pauly M. Bootstrapping Aalen-Johansen processes for competing risks: handicaps, solutions, and limitations. *Electron J Stat*. 2014;8(2):2779-2803.
38. Mas J-L, Chatellier G, Beyssen B, et al. Endarterectomy versus stenting in patients with symptomatic severe carotid stenosis. *N Engl J Med*. 2006;355(16):1660-1671.
39. Howard G, Roubin GS, Jansen O, et al. Association between age and risk of stroke or death from carotid endarterectomy and carotid stenting: a meta-analysis of pooled patient data from four randomised trials. *Lancet*. 2016;387(10025):1305-1311.
40. Alymani NA, Smith MD, Williams DJ, Petty RD. Predictive biomarkers for personalised anti-cancer drug use: discovery to clinical implementation. *Eur J Cancer*. 2010;46(5):869-879.
41. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. *Transl Cancer Res*. 2015;4(3):256.
42. Koch C, Trojan J. Established and potential predictive biomarkers in gastrointestinal cancer-c-Kit, Her2, Ras and Beyond. *Digestion*. 2015;91(4):294-302.
43. Perez-Gracia JL, Sanmamed MF, Bosch A, et al. Strategies to design clinical studies to identify predictive biomarkers in cancer research. *Cancer Treat Rev*. 2017;53:79-97.
44. Royston P, Sauerbrei W. Interactions between a treatment and continuous covariates: a step toward individualizing therapy. *J Clin Oncol*. 2008;26(9):1397-1399.
45. Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *Am J Neuroradiol*. 2011.

46. Fieller EC. The biological standardization of insulin. *J R Stat Soc Series B Stat Methodol*. 1940;7(1):1-64.
47. Boulesteix A-L, Lauer S, Eugster MJ. A plea for neutral comparison studies in computational sciences. *PLoS One*. 2013;8(4):1-11.
48. Boulesteix A-L, Wilson R, Hapfelmeier A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Med Res Methodol*. 2017;17(1):138.
49. Liu Y, Jiang W, Chen BE. Testing for treatment-biomarker interaction based on local partial-likelihood. *Statist Med*. 2015;34(27):3516-3530.
50. Zhang Z, Seibold H, Vettore MV, Song W-J, François V. Subgroup identification in clinical trials: an overview of available methods and their implementations with R. *Ann Transl Med*. 2018;6(7).

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Haller B, Mansmann U, Dobler D, Ulm K, Hapfelmeier A. Confidence interval estimation for the changepoint of treatment stratification in the presence of a qualitative covariate-treatment interaction. *Statistics in Medicine*. 2020;39:70–96. https://doi.org/10.1002/sim.8404

APPENDIX

COVERAGE TABLES

TABLE A1 Coverage for Model I for scenarios with normally distributed covariate for the different methods for confidence interval estimation. (2000 runs, 1000 bootstrap samples)

						Bootstrap	
n	**Loc.**	**Cens.**	**Fieller**	**Delta**	**Perc.**	**Normal**	**Wild**
200	50th perc.	low	93.6%	99.2%	97.3%	99.8%	99.5%
200	50th perc.	high	94.0%	99.0%	98.9%	100.0%	99.4%
200	70th perc.	low	95.1%	97.8%	97.9%	99.5%	98.2%
200	70th perc.	high	96.0%	99.4%	99.3%	99.9%	99.6%
200	90th perc.	low	94.6%	91.8%	97.1%	97.2%	92.7%
200	90th perc.	high	94.7%	91.3%	97.2%	99.2%	93.5%
500	50th perc.	low	95.2%	98.3%	96.0%	99.3%	98.2%
500	50th perc.	high	95.5%	98.6%	98.0%	99.7%	99.0%
500	70th perc.	low	95.8%	97.2%	97.0%	98.7%	97.4%
500	70th perc.	high	95.0%	98.6%	97.8%	99.8%	99.0%
500	90th perc.	low	95.4%	92.4%	97.0%	95.8%	93.0%
500	90th perc.	high	95.2%	92.2%	97.4%	98.2%	93.1%
1000	50th perc.	low	95.2%	97.3%	94.8%	98.2%	97.1%
1000	50th perc.	high	95.3%	98.0%	96.0%	99.2%	98.0%
1000	70th perc.	low	95.1%	96.2%	95.5%	97.2%	96.4%
1000	70th perc.	high	94.6%	97.9%	96.0%	99.4%	98.2%
1000	90th perc.	low	93.9%	94.0%	94.6%	95.8%	94.0%
1000	90th perc.	high	95.6%	94.6%	97.8%	97.8%	95.0%
2000	50th perc.	low	95.7%	96.7%	95.8%	97.1%	96.6%
2000	50th perc.	high	95.6%	97.4%	95.8%	98.6%	97.6%
2000	70th perc.	low	94.5%	95.4%	94.0%	95.9%	95.4%
2000	70th perc.	high	95.2%	97.2%	95.2%	98.6%	97.4%
2000	90th perc.	low	95.4%	95.8%	95.2%	96.7%	95.8%
2000	90th perc.	high	95.6%	94.8%	96.6%	96.8%	95.0%
5000	50th perc.	low	95.2%	95.4%	95.0%	95.8%	95.5%
5000	50th perc.	high	95.0%	95.9%	94.8%	96.4%	96.0%
5000	70th perc.	low	95.1%	95.8%	95.4%	96.0%	95.8%
5000	70th perc.	high	95.4%	96.3%	94.9%	96.9%	96.5%
5000	90th perc.	low	95.4%	95.6%	95.0%	95.8%	95.6%
5000	90th perc.	high	94.7%	94.9%	94.4%	96.4%	95.2%

Loc., True location of x_{cp}; Cens., Censoring; Perc., percentile-based.

TABLE A2 Coverage for Model I for scenarios with uniformly distributed covariate for the different methods for confidence interval estimation. (2000 runs, 1000 bootstrap samples)

						Bootstrap	
n	**Loc.**	**Cens.**	**Fieller**	**Delta**	**Perc.**	**Normal**	**Wild**
200	50th perc.	low	95.0%	99.4%	97.6%	99.9%	99.4%
200	50th perc.	high	95.2%	99.7%	99.4%	100.0%	99.9%
200	70th perc.	low	96.0%	95.8%	97.8%	99.1%	96.8%
200	70th perc.	high	94.4%	97.4%	98.4%	99.8%	98.2%
200	90th perc.	low	95.4%	91.5%	97.2%	97.6%	92.0%
200	90th perc.	high	96.0%	89.9%	97.7%	99.1%	91.6%
500	50th perc.	low	95.2%	98.2%	95.9%	99.8%	98.3%
500	50th perc.	high	95.0%	98.9%	97.7%	99.8%	99.0%
500	70th perc.	low	95.2%	95.4%	96.5%	97.6%	95.6%
500	70th perc.	high	95.0%	96.2%	97.8%	99.2%	96.7%
500	90th perc.	low	95.4%	93.0%	96.9%	96.2%	93.2%
500	90th perc.	high	95.4%	91.8%	97.5%	97.5%	92.6%
1000	50th perc.	low	95.7%	97.4%	95.8%	98.6%	97.5%
1000	50th perc.	high	95.6%	98.2%	96.7%	99.4%	98.6%
1000	70th perc.	low	95.6%	96.3%	95.6%	96.9%	96.2%
1000	70th perc.	high	95.2%	96.0%	97.0%	98.2%	96.4%
1000	90th perc.	low	94.6%	94.0%	95.4%	96.0%	94.2%
1000	90th perc.	high	95.5%	93.1%	97.5%	96.8%	93.4%
2000	50th perc.	low	94.3%	95.4%	94.1%	95.9%	95.5%
2000	50th perc.	high	95.2%	97.4%	95.2%	98.4%	97.6%
2000	70th perc.	low	95.0%	95.3%	94.8%	96.3%	95.5%
2000	70th perc.	high	95.0%	96.4%	95.2%	97.5%	96.6%
2000	90th perc.	low	95.8%	94.4%	95.2%	95.8%	94.4%
2000	90th perc.	high	94.7%	93.8%	96.8%	96.4%	94.4%
5000	50th perc.	low	94.0%	94.2%	94.0%	94.4%	94.2%
5000	50th perc.	high	95.4%	96.2%	95.2%	96.4%	96.2%
5000	70th perc.	low	95.4%	95.4%	95.0%	95.6%	95.4%
5000	70th perc.	high	95.2%	96.2%	94.9%	96.8%	96.3%
5000	90th perc.	low	94.4%	94.7%	94.1%	95.3%	94.8%
5000	90th perc.	high	95.2%	95.0%	95.0%	96.4%	95.0%

Loc., True location of x_{cp}; Cens., Censoring; Perc., percentile-based.

TABLE A3 Coverage for simulations performed for Model II (2000 runs for each scenario)

				Not considering Z_1, Z_2		Considering Z_1, Z_2	
$\beta_{Z_{1,2}}$	Σ	**n**	**Cens.**	**Fieller**	**Delta**	**Fieller**	**Delta**
ln(1.2)	Σ_0	200	low	94.2%	99.4%	94.4%	99.4%
ln(1.2)	Σ_0	200	high	95.1%	99.0%	94.4%	99.0%
ln(1.2)	Σ_1	200	low	94.0%	99.1%	94.4%	99.2%
ln(1.2)	Σ_1	200	high	95.2%	98.8%	95.4%	98.7%
ln(1.5)	Σ_0	200	low	95.5%	99.8%	95.6%	99.6%
ln(1.5)	Σ_0	200	high	94.8%	99.0%	94.2%	99.1%
ln(1.5)	Σ_1	200	low	95.6%	99.2%	95.0%	98.8%
ln(1.5)	Σ_1	200	high	95.3%	97.6%	94.8%	97.3%
ln(1.2)	Σ_0	1000	low	94.7%	96.8%	94.6%	96.5%
ln(1.2)	Σ_0	1000	high	95.7%	97.8%	95.1%	97.2%
ln(1.2)	Σ_1	1000	low	95.4%	97.9%	95.8%	97.4%
ln(1.2)	Σ_1	1000	high	94.5%	97.1%	94.6%	97.4%
ln(1.5)	Σ_0	1000	low	95.6%	98.0%	95.2%	97.4%
ln(1.5)	Σ_0	1000	high	94.6%	97.4%	94.5%	97.2%
ln(1.5)	Σ_1	1000	low	94.9%	97.8%	94.8%	96.8%
ln(1.5)	Σ_1	1000	high	95.5%	96.2%	95.4%	95.8%
ln(1.2)	Σ_0	5000	low	94.6%	95.0%	94.8%	95.2%
ln(1.2)	Σ_0	5000	high	95.4%	96.6%	95.8%	96.6%
ln(1.2)	Σ_1	5000	low	95.2%	95.4%	94.9%	95.6%
ln(1.2)	Σ_1	5000	high	95.5%	95.9%	94.9%	95.6%
ln(1.5)	Σ_0	5000	low	94.7%	95.4%	95.0%	95.3%
ln(1.5)	Σ_0	5000	high	95.2%	96.0%	95.0%	96.0%
ln(1.5)	Σ_1	5000	low	95.6%	96.1%	95.4%	95.9%
ln(1.5)	Σ_1	5000	high	95.0%	95.6%	94.8%	95.9%

TABLE A4 Coverage for simulations performed for Model III (2000 runs for each scenario). Due to the symmetry of the results, only results for 50th, 75th, and 95th percentile of Z_1 and Z_2 are shown

		Percentile of			Fieller			Delta	
$\beta_{G\times Z_{1,2}}$	Σ	Z_1	Z_2	$n = 200$	$n = 1000$	$n = 5000$	$n = 200$	$n = 1000$	$n = 5000$
ln(1.2)	Σ_0	50th	50th	94.4%	95.0%	94.7%	99.1%	96.6%	95.0%
ln(1.2)	Σ_0	50th	75th	94.6%	95.7%	94.8%	97.5%	97.0%	95.2%
ln(1.2)	Σ_0	50th	95th	94.6%	95.0%	94.4%	96.4%	96.2%	94.8%
ln(1.2)	Σ_0	75th	50th	94.2%	95.9%	95.8%	96.8%	96.6%	95.8%
ln(1.2)	Σ_0	75th	75th	94.6%	96.0%	95.2%	94.9%	96.0%	95.2%
ln(1.2)	Σ_0	75th	95th	94.6%	95.6%	94.6%	94.4%	95.8%	95.2%
ln(1.2)	Σ_0	95th	50th	94.2%	95.4%	96.0%	96.4%	95.9%	96.3%
ln(1.2)	Σ_0	95th	75th	94.5%	95.6%	95.5%	94.6%	95.8%	96.0%
ln(1.2)	Σ_0	95th	95th	94.0%	96.0%	94.9%	93.4%	95.7%	95.9%
ln(1.2)	Σ_1	50th	50th	94.8%	95.6%	95.1%	99.6%	98.4%	95.8%
ln(1.2)	Σ_1	50th	75th	94.3%	94.3%	94.9%	95.4%	94.9%	95.6%
ln(1.2)	Σ_1	50th	95th	94.8%	93.6%	95.2%	93.3%	93.6%	95.6%
ln(1.2)	Σ_1	75th	50th	94.4%	94.0%	94.6%	94.4%	94.9%	95.6%
ln(1.2)	Σ_1	75th	75th	94.3%	93.8%	94.9%	89.6%	92.6%	95.0%
ln(1.2)	Σ_1	75th	95th	94.6%	93.5%	95.0%	89.6%	92.6%	95.2%
ln(1.2)	Σ_1	95th	50th	93.8%	94.1%	94.6%	92.4%	94.1%	95.2%
ln(1.2)	Σ_1	95th	75th	94.0%	93.8%	94.4%	89.9%	93.1%	95.0%
ln(1.2)	Σ_1	95th	95th	94.2%	93.5%	94.8%	88.2%	91.9%	94.0%
ln(1.5)	Σ_0	50th	50th	94.2%	94.5%	94.9%	99.0%	96.6%	95.3%
ln(1.5)	Σ_0	50th	75th	94.4%	94.4%	95.5%	93.6%	94.3%	95.3%
ln(1.5)	Σ_0	50th	95th	94.2%	93.8%	95.5%	91.0%	93.8%	95.8%
ln(1.5)	Σ_0	75th	50th	94.2%	94.2%	95.1%	93.8%	94.7%	95.1%
ln(1.5)	Σ_0	75th	75th	94.6%	94.4%	95.2%	90.4%	93.4%	95.0%
ln(1.5)	Σ_0	75th	95th	94.2%	94.2%	95.6%	88.7%	93.7%	95.2%
ln(1.5)	Σ_0	95th	50th	94.1%	94.3%	95.1%	91.6%	93.8%	95.4%
ln(1.5)	Σ_0	95th	75th	94.4%	94.1%	95.4%	90.0%	93.3%	95.2%
ln(1.5)	Σ_0	95th	95th	94.1%	94.4%	95.1%	89.1%	93.7%	94.8%
ln(1.5)	Σ_1	50th	50th	94.7%	95.6%	95.9%	99.6%	98.0%	96.4%
ln(1.5)	Σ_1	50th	75th	94.0%	94.7%	95.2%	90.6%	93.6%	95.4%
ln(1.5)	Σ_1	50th	95th	93.6%	94.4%	95.0%	88.5%	93.2%	94.8%
ln(1.5)	Σ_1	75th	50th	95.1%	94.8%	95.0%	90.2%	93.9%	95.2%
ln(1.5)	Σ_1	75th	75th	94.4%	94.8%	95.3%	86.6%	92.0%	94.6%
ln(1.5)	Σ_1	75th	95th	94.4%	94.8%	94.8%	86.6%	92.0%	94.8%
ln(1.5)	Σ_1	95th	50th	94.9%	94.2%	94.9%	88.3%	92.2%	94.8%
ln(1.5)	Σ_1	95th	75th	95.0%	94.5%	94.8%	85.8%	91.8%	94.6%
ln(1.5)	Σ_1	95th	95th	95.0%	94.6%	95.0%	85.8%	91.8%	95.0%