WILEY

Performance of methods to conduct mediation analysis with time-to-event outcomes

Lizbeth Burgos Ochoa[1] | **Judith J.M. Rijnhart[2]** |
Brenda W. Penninx[3] | **Klaas J. Wardenaar[4]** |
Jos W.R. Twisk[2] | **Martijn W. Heymans[2]**

[1]Department of Obstetrics and Gynaecology, Erasmus MC, University Medical Centre Rotterdam, Rotterdam, The Netherlands

[2]Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, Amsterdam UMC, VU University Medical Center, Amsterdam, The Netherlands

[3]Department of Psychiatry, Amsterdam UMC, VU University Medical Center, Amsterdam, The Netherlands

[4]Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), University of Groningen, University Medical Center Groningen (UMCG), Groningen, The Netherlands

Correspondence
Lizbeth Burgos Ochoa, Department of Obstetrics and Gynaecology, Erasmus MC, University Medical Centre Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands.
Email: l.burgosochoa@erasmusmc.nl

Previous studies have discouraged the use of the Cox proportional hazards (PH) model for traditional mediation analysis as it might provide biased results. Accelerated failure time (AFT) models have been proposed as an alternative for Cox PH models. In addition, the use of the potential outcomes framework has been proposed for mediation models with time-to-event outcomes. The aim of this paper is to investigate the performance of traditional mediation analysis and potential outcomes mediation analysis based on both the Cox PH and the AFT model. This is done by means of a Monte Carlo simulation study and the illustration of the methods using an empirical data set. Both the product-of-coefficients method of the traditional mediation analysis and the potential outcomes framework yield unbiased estimates with respect to their own underlying indirect effect value for simple mediation models with a time-to-event outcome and estimated based on Cox PH or AFT.

KEYWORDS

accelerated failure time, causal inference, Cox proportional hazards, mediation analysis, potential outcomes, survival analysis, time-to-event

---

University Medical Center; GGZ inGeest;
Leiden University Medical Center; Leiden
University; GGZ Rivierduinen; University
Medical Center Groningen; University of
Groningen, Lentis; GGZ Friesland; GGZ
Drenthe; Rob Giel Onderzoekscentrum
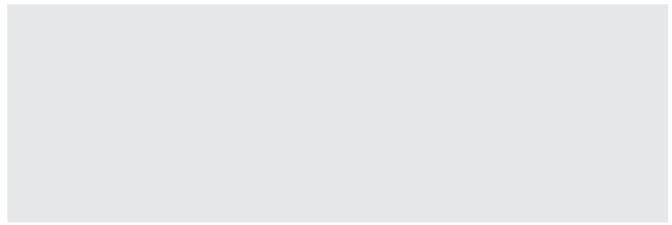
1 | INTRODUCTION

Statistical mediation analysis has recently raised interest in the field of epidemiology (e.g., Gerrits et al., 2014; Richiardi, Bellocco, & Zugna, 2013). Its main goal is to identify underlying mechanisms of exposure–outcome associations by investigating to what extent this association can be explained by a mediator (Gelfand, MacKinnon, DeRubeis, & Baraldi, 2016). Figure 1 is a graphical representation of a single mediator model. Following Figure 1, Equations (1), (2), and (3) represent the regression models necessary for traditional mediation analysis (Baron & Kenny, 1986; MacKinnon, 2012):

$$Y = i_1 + cX + e_1 \qquad (1)$$

$$Y = i_2 + c'X + bM + e_2 \qquad (2)$$

$$M = i_3 + aX + e_3. \qquad (3)$$

Equation (1) represents the total effect c of exposure X on outcome Y. Equation (2) represents both the direct effect c' and the b path. Lastly, Equation (3) represents the a path. In all the equations, the i terms represent the intercepts and the e terms represent error variability. In traditional mediation analysis, the indirect effect is calculated as either the product of the a and b coefficients or as the difference between the c and c' coefficients. These two methods are algebraically equivalent for continuous outcomes (MacKinnon, Warsi, & Dwyer, 1995). In the remainder of this paper, we will refer to these respective methods as the product (ab) and difference (c-c') methods, respectively.
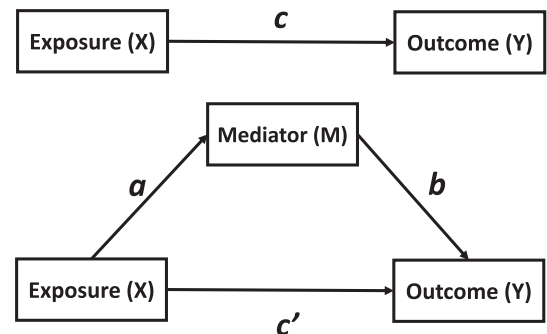

FIGURE 1 Single mediator model

In epidemiological research, time-to-event outcome variables are common. Time-to-event variables provide information about the time until an event occurs, for example, treatment responses or adverse events, and are common in fields such as oncology and cardiology (e.g., Fizazi et al., 2012; Sedlis et al., 2015). Two commonly used methods for the analysis of time-to-event outcomes are the Cox proportional hazards (PH) model and the accelerated failure time (AFT) model. Both of these models can be used for traditional mediation analysis with a time-to-event outcome by analyzing Equations (1) and (2) with either the Cox PH model or the AFT model.

Several authors have discouraged the use of the Cox PH model for traditional mediation analysis (Lange & Hansen, 2011; Lapointe-Shaw et al., 2018; Martinussen, Vansteelandt, Gerster, & Hjelmborg, 2011; VanderWeele, 2011) because the product and difference methods yield different indirect effect estimates when based on the Cox PH model (Gelfand et al., 2016; Tein & MacKinnon, 2003). These two methods are assumed to approximate each other only in case of rare outcomes (VanderWeele, 2011). However, in practice, rare outcomes are rather uncommon (Gelfand et al., 2016).

In contrast with the Cox PH model, the AFT model has shown to yield similar indirect effect estimates for the product and difference method when applied to uncensored time-to-event outcomes (Tein & MacKinnon, 2003). However, simulation studies showed that, in the presence of right censoring, the estimates for these two methods are not equivalent (Fulcher, Tchetgen Tchetgen, & Williams, 2017; Gelfand et al., 2016). While the product-of-coefficients method remains relatively unaffected by right censoring, the difference-between-coefficients method tends to underestimate the indirect effect (Gelfand et al., 2016). The results from these previous studies raise the question of whether either of these traditional methods for calculating the indirect effect provide an accurate and unbiased approximation of the underlying indirect effect (VanderWeele, 2011).

Recently, the application of the potential outcomes framework to mediation analysis has received more attention (Pearl, 2001; VanderWeele, 2011). This framework defines the indirect effect in terms of potential outcomes and, therefore, only provides a single estimate of the indirect effect. These potential outcomes can be estimated using various estimation approaches, for example, based on simulations, numerical integration, and regression models (Imai, Keele, & Tingley, 2010; Muthén, Muthén, & Asparouhov, 2017; VanderWeele, 2011; Vansteelandt, Bekaert, & Lange, 2012). In this paper, we will focus on the simulation-based approach described in Lange, Vansteelandt, and Bekaert (2012) and Vansteelandt et al. (2012). The potential outcomes framework can be used for many types of outcome variables, including time-to-event outcomes (Lange et al., 2012; VanderWeele, 2011). In case of time-to-event outcomes, both the Cox PH and AFT models can be used to estimate the potential outcomes. However, there is still little evidence of its performance, as it has been mostly illustrated through empirical data sets (e.g., Lange, Hansen, Sørensen, & Galatius, 2017; Tchetgen Tchetgen, 2013).

The aim of this paper is to investigate the performance of traditional mediation analysis and potential outcomes mediation analysis based on both the Cox PH model and the AFT model. This paper is organized as follows. First, we describe how to apply traditional mediation analysis and the potential outcomes framework to mediation models with time-to-event outcomes. Second, we outline the design of our Monte Carlo simulation study to investigate the performance of these methods. Third, we describe the empirical data set that we used for the illustration of the methods. Fourth, we report the obtained results from the simulation study and the empirical data example. Fifth, we discuss our findings and provide advice for substantive researchers who want to perform mediation analysis with time-to-event outcomes.

2 | METHODS

2.1 | Methods for performing mediation analysis with time-to-event outcomes

2.1.1 | Traditional mediation analysis

Within traditional mediation analysis, Equations (1) and (2) are fitted as either Cox PH models or AFT models to analyze the time-to-event outcome. The Cox PH model, which is most commonly applied in epidemiological studies (George, Seals, & Aban, 2014), is a semiparametric model that can be used to model the relationship between an exposure and a time-to-event outcome through the hazard function (George et al., 2014). The key assumption of the Cox PH model is proportional hazards, which means that the ratio of the hazard under two different values of the exposure variable remains constant over time. The regression coefficients in a Cox PH model are interpretable as the natural logarithms of hazard ratios. In the AFT model, the logarithm of the survival time is used as the outcome. AFT models require that the survival times follow a parametric error distribution and that this distribution is a priori specified by the researcher (e.g., exponential, Weibull, log- normal, log-logistic, etc.; Swindell, 2009). The regression coefficients of the AFT model are interpretable as the difference in the speed of developing the outcome under two different values of the exposure (Kleinbaum & Klein, 2010; Zhang, 2016).

2.1.2 | Potential outcomes mediation analysis

The potential outcomes framework defines causal effects in terms of potential outcomes. Ideally, we would like to observe the causal effect of the exposure on the outcome for an individual participant. This effect would be defined as the difference in the outcome if the individual participant was assigned to two different values of the exposure variable, for example, both the treatment and control conditions when the exposure reflects a treatment (Gelfand et al., 2016; Richiardi et al., 2013). In this case, $T(x)$ is the potential time-to-event outcome under the exposure $X = x$ (e.g., the treatment condition), and $T(x^*)$ is the potential outcome under the other exposure level $X = x^*$ (e.g., the control condition; Rubin, 1974). The difference between $T(x)$ and $T(x^*)$ would represent in this situation the causal exposure–outcome effect.

In the context of mediation analysis, the potential outcomes do not only depend on exposure values but also on mediator values (Gelfand et al., 2016). Therefore, the potential outcomes notation is extended to $T(x, m)$, where x reflects the value of the exposure and m reflects the value of the mediator (Lange et al., 2012). A unique feature of the potential outcomes framework is that the notation is even further extended by including nested potential outcomes. Instead of fixing m to a random value of the mediator, for example, its mean, the potential outcomes framework estimates the mediator under values that would be observed in the population, where $M(x)$ is the value of the mediator under the exposure $X = x$ (e.g., the treatment condition) and $M(x^*)$ is the value of the mediator under the exposure $X = x^*$ (e.g., the control condition). These potential values of the mediator are subsequently used to estimate the potential time-to-event outcomes. Based on this notation, four potential time-to-event outcomes can be observed: $T(x, M(x))$, $T(x, M(x^*))$, $T(x^*, M(x))$, and $T(x^*, M(x^*))$.

In the potential outcomes framework, the direct and indirect effect are referred to as natural direct and indirect effects, as they are based on a comparison of potential time-to-event outcomes that could have naturally been observed. The natural indirect effect (NIE) is estimated as the difference between the potential outcomes $T(x, M(x))$ and $T(x, M(x^*))$ and, thus, reflects the

difference in the time-to-event outcome when changing the value of the mediator. Likewise, the natural direct effect (NDE) is calculated as the difference between the potential outcomes T(x, M(x^*)) and T(x^*, M(x^*)) and, thus, reflects the difference in the time-to-event outcomes when changing the value of the exposure (Robins & Greenland, 1992). The total effect (TE) can subsequently be estimated as the summation of the natural direct and indirect effect (Pearl, 2001; Robins & Greenland, 1992). Note that, in practice, only two of the four potential outcomes can be observed for individual participants, as the two potential outcomes T(x, M(x^*)) and T(x^*, M(x)) assume that the individual participant is exposed to two different values of the exposure variable.

Vansteelandt et al. (2012) and Lange et al. (2012) proposed a simulation-based approach for the estimation of the four potential outcomes needed for the estimation of the natural direct and indirect effect. This simulation-based approach is based on the concept of natural effect models in which the potential outcomes are directly estimated. To our knowledge, the simulation-based approach has only been described for the Cox PH model (Lange et al., 2017). However, given the nature of the natural effect models used in the simulation-based approach, this approach can easily be extended to the AFT model. The simulation-based approach is based on the following five steps (Lange et al., 2017).

1. Use the original data set to fit the imputation model, that is, a Cox PH or AFT model in which the outcome is the dependent variable and both the exposure and mediator are independent variables.
2. Create a new data set based on the original data set in which all observations are present twice. Furthermore, two new variables are added to the new data set: one variable representing x, which represents the value of the exposure as observed in the original data set, and one variable representing x^*, which represents the opposite of the observed exposure value.
3. Impute the potential outcome Y(x, M(x^*)) as the observed value of the outcome for the observations of the database where x equals x^*, and as the predicted outcome based on the imputation model fitted at Step 1 for the other lines.
4. Fit the natural effect model to the data. In the natural effect model, the variable created at Step 3 is the dependent variable, and the x and x^* variables created at Step 2 are both included in the model as independent variables. In this model, the coefficient corresponding to the x variable represents the NDE, and the coefficient corresponding to the x^* variable represents the NIE.
5. Repeat the previous process 10 times and combine the results from these 10 imputations into one NDE and one NIE using standard multiple imputation formulas.

2.2 | Design of the simulation study

The performance of traditional mediation analysis and potential outcomes mediation analysis based on both the Cox PH and AFT models were assessed with a Monte Carlo simulation study. Both continuous and binary exposure and mediator variables were considered. The continuous exposure variable was generated from a normal distribution with a mean of 0 and a variance of 0.5. The continuous mediator variable was generated from a normal distribution with a mean of 0 and a variance of 1. Both the binary exposure and mediator were generated from a Bernoulli distribution with a probability of 0.5. The a coefficient, relating the exposure to the mediator, was set to 0.6 and represents a medium-to-large effect size (Cohen, 2013). The survival times (T) for the time-to-event outcome variable were simulated following a Weibull error distribution. The Weibull error distribution is flexible enough to represent a wide range of distributions often found

in epidemiologic research (Gelfand et al., 2016). There are several ways to parameterize a Weibull distribution, and we used the most common parameterization:

$$T = exp\left(i_2 + c'X + bM + \frac{1}{p}\epsilon\right), \tag{4}$$

where the parameter i_2 was set to 4, and c' and b were both set to 0.6. The term p is a shape parameter for the survival times, and $\frac{1}{p}$ scales the parameter of ϵ so that the survival times follow a Weibull error distribution. To reflect different possible scenarios encountered in clinical research, the shape parameter p for the survival times was set to represent decreasing, constant, and increasing hazards, respectively. The term ϵ is a random variable following an extreme value distribution, which is a limiting model for the maximums and minimums of the data set.

Typically in clinical studies, a proportion of participants reach the end of the study before developing the outcome of interest. This condition is referred to as study-duration censoring (Gelfand et al., 2016). In the generated data sets, we introduced study-duration censoring. The study-duration censoring was assumed to be noninformative, that is, the survival time and the censoring time variables are statistically independent of each other. To introduce different censoring rates, two steps were followed. First, no censoring was assumed and the survival time (t) for each individual was simulated. Second, censoring was introduced by simulating a censoring time from a uniform (0, δ) distribution (Crowther & Lambert, 2013). The observed survival time (T) was defined as the minimum value of δ and t for each individual. The observation was considered censored if δ was smaller than t. The value of δ was varied to achieve the desired censoring rate. For each data set, sampling was continued until the prespecified criterion for the censoring rate was met. For the nonrare outcome, a censoring rate of 40% was selected, as it has been considered enough to show the effects of censoring without being unreasonable in the context of clinical research (Gelfand et al., 2016; Swift & Greenberg, 2012). To generate rare outcomes we generated study-duration censoring rates of 90% based on the approach described by Ambler, Seaman, and Omar (2012) and Vanderweele et al. (2016b).

A total of 108 simulation scenarios were generated. The simulation scenarios vary by sample size (250, 500, and 1,000), type of exposure (binary or continuous), type of mediator (binary or continuous), hazard rate (constant, increasing, and decreasing hazards), and censoring rate (no censoring, 40%, and 90%). For each scenario, 500 data sets were generated. All the analyses and simulations were conducted using R software version 3.4.3 (R Core Team, 2017). In the scenarios with a continuous mediator, the a path was estimated with linear regression. In the scenarios with a binary mediator, the a path was estimated with a logistic regression based on a generalized linear model with a logit link. The Cox PH models were fitted using the *coxph* function and the AFT models were fitted using the *survreg* function in the R *survival* package (Therneau, 2015). Both models were estimated assuming a Weibull error distribution of the time-to-event outcome.

2.2.1 | Performance measures

The performance of the mediation methods is presented in terms of bias (absolute and proportional), precision (empirical standard error), and the mean squared error (MSE) for the estimates of the indirect effect (Burton, Altman, Royston, & Holder, 2006). These performance measures are all calculated using the underlying simulated indirect effect. In some of our simulation conditions, the underlying indirect effect differs between the traditional mediation methods and potential outcomes framework. Each of these performance measures is therefore calculated using underlying traditional indirect effect for the estimates

based on traditional mediation analysis and using the underlying potential outcomes indirect effect for the potential outcomes framework estimates.

The traditional indirect effect is calculated as the product of the a and b coefficients for both AFT and Cox PH models and regardless of whether the mediator was continuous or binary. Because the a and b coefficients were set to 0.6 in all scenarios, the true underlying indirect effect was equal to 0.36 across all scenarios for traditional mediation analysis.

The potential outcomes indirect effect for models with a continuous mediator and based on both AFT and Cox PH models was calculated as the product of the a and b coefficients. Therefore, in scenarios with a continuous mediator, the underlying potential outcomes indirect effect was equal to 0.36. In scenarios with a binary mediator, the underlying indirect effect for the potential outcomes framework was 0.081, based on the NIE formulas presented in VanderWeele (2016a).

The absolute bias was calculated by subtracting the underlying simulated indirect effect from the average indirect effect estimates. The percentage bias was calculated by dividing the absolute bias by the underlying indirect effect times 100. To measure the precision of the estimates, the empirical standard error (SE) was calculated as the standard deviation (SD) of the indirect effect estimates from all simulations (Burton et al., 2006). To measure the variability in the simulated estimates, the MSE was calculated as the averaged sum of the squared bias and squared empirical SE (Burton et al., 2006).

2.2.2 | Application to a real-life data example

We also applied the methods for mediation analysis with survival data to a real-life data example to illustrate their performance in a practical application. We used data from the Netherlands Study of Depression and Anxiety (NESDA; Gerrits et al., 2014; Penninx et al., 2008). NESDA is a longitudinal cohort study designed to investigate the long-term course and consequences of depressive and anxiety disorders (Penninx et al., 2008). The full data set consists of 2,981 participants (aged 18 to 65 years) who were included from the general population (n = 564), general practices (n = 1,610), and mental health care organizations (n = 807). Baseline data were collected between 2004 and 2007, with follow-up assessments 2, 4, 6, and 9 years later. At baseline, all 2,981 participants were screened for depression and anxiety at baseline using the DSM-IV–based Composite International Diagnostic Interview (CIDI, Version 2.1; L. N. Robins et al., 1988).

Our example is based on a study published by Gerrits et al. (2014) who aimed at investigating whether subthreshold depressive symptoms mediate associations between chronic pain and the recurrence of depressive disorders based on the baseline data and the 2- and 4-year follow-up assessments. A subset of the data set was used, based on 1,236 participants who had reported to have had a depressive or anxiety disorder in the past, but were currently in remission according to the CIDI, which meant they did not have a diagnosis of a depressive or anxiety disorder (in the previous six months) either at baseline (n = 628) or at the 2-year follow-up assessment (n = 608). Of these, 114 participants were excluded from the analyses because they did not participate in the follow-up assessments. Therefore, the final data set consists of 1,122 participants. Note that, because drop outs were not included in the final data set, censored data are based on study-duration censoring.

Recurrence of a depressive disorder (yes/no) was defined by the DSM-IV based CIDI during follow-up. The time to recurrence of a depressive disorder was calculated in months from the time the participant was assessed as being in remission (did not have a current diagnosis) until they were diagnosed with a depressive or anxiety disorder at one of the follow-up assessments. When a participant was diagnosed with recurrence at the 2- or 4-year follow-up assessment, participants

were asked to indicate the recency of onset with one of the following categories: less than a month ago, between 1 and 6 months ago, between 6 and 12 months ago, 12 months ago, and between 12 and 24 months ago. Based on this information, the median of the interval to recurrence was calculated. For participants with no recurrence of depression, time was censored as the time from the assessment in which remission was defined until the end of the follow-up period.

In our example, we will only focus on the exposure variable self-reported pain, as measured by the chronic pain grade (CPG). The CPG was assessed at the point at which the participants were in remission, that is, either at baseline or at the two-year follow up assessment. The CPG reflects the intensity of and disability caused by pain (Von Korff, Ormel, Keefe, & Dworkin, 1992). The CPG is based on grades ranging from 0 to 4, where 0 corresponds to no pain and 4 corresponds to high disability–severely limiting. Even though the CPG can potentially be treated as an ordinal variable, we will follow the same procedure as Gerrits et al. (2014) and treat the CPG as an ordinal approximation of a continuous variable.

The mediator, subthreshold depressive symptoms, was measured by the Quick Inventory of Depressive Symptomatology (QIDS), which is a self-report questionnaire consisting of 16 items that are rated on a 4-point Likert scale (0–3; Rush et al., 2003). Item scores are added up to a total severity score with a range of 0–27. Following the original study, we adjusted all analyses for potential confounding by sex, age, year of education, and the recency of last episode of the depressive and/or anxiety disorder at baseline. All analyses were conducted using R software version 3.4.3 (R Core Team, 2017).

3 | RESULTS

3.1 | Results from the simulation study

3.1.1 | Traditional mediation analysis

Table 1 shows the estimates obtained from the simulations of the traditional mediation approach based on both the Cox PH and AFT models for a sample size 500 and a model with a continuous exposure and mediator. For traditional mediation analysis based on the AFT model, both ab and $c\text{-}c'$ approximated the underlying indirect effect of 0.36. The lowest absolute difference between ab and $c\text{-}c'$ across all hazard rate conditions were obtained in the no-censoring conditions. The direct effect based on the AFT model was generally close to the underlying direct effect of 0.6.

Traditional mediation analysis based on the Cox PH model only approximated the underlying indirect effect of 0.36 when the hazard was constant and the indirect effect was calculated as ab. When the hazard was increasing, ab overestimated the indirect effect. When the hazard was decreasing, ab underestimated the indirect effect. A similar pattern was observed for the direct effect estimates. Furthermore, $c\text{-}c'$ underestimated the underlying indirect effect in all scenarios. The lowest absolute difference between the ab and $c\text{-}c'$ estimates across all hazard rate conditions were obtained for the 90% censoring scenarios, that is, the scenarios with rare outcomes.

For both the Cox PH and AFT model, and for both ab and $c\text{-}c'$, the lowest SE values were obtained in the no-censoring conditions. The simulation results for sample sizes of 250 and 1,000 and for other exposure-mediator type combinations showed similar patterns to the results presented above. Detailed simulation results for the traditional mediation analysis can be found in the Supplementary Tables (S1–S8).

Figure 2 shows the percentage bias for ab and $c\text{-}c'$ based on the AFT model (Figures 2A and 2C) and the Cox PH model (Figures 2B and 2D) for scenarios with a continuous exposure and

TABLE 1 Direct and indirect effect estimates from traditional mediation analysis based on the AFT and Cox PH models for the scenarios with a continuous mediator and outcome (N = 500)

| Model | Conditions | | c' | ab | Indirect effect | | | |
	Hazard	Censoring (%)			SE ab	c-c'	SE c-c'	Difference ab – (c-c')
AFT	Increasing	NC	0.596	0.354	0.054	0.356	0.074	0.001
		40	0.617	0.369	0.117	0.322	0.146	0.047
		90	0.568	0.352	0.089	0.334	0.099	0.018
	Constant	NC	0.594	0.354	0.058	0.355	0.078	0.001
		40	0.607	0.367	0.109	0.335	0.118	0.032
		90	0.609	0.417	0.855	0.312	2.443	0.105
	Decreasing	NC	0.592	0.354	0.065	0.355	0.084	0.001
		40	0.605	0.360	0.114	0.339	0.108	0.021
		90	0.625	0.364	0.151	0.357	0.150	0.008
COX PH	Increasing	NC	0.899	0.534	0.085	0.078	0.082	0.457
		40	0.943	0.561	0.191	0.163	0.194	0.399
		90	0.873	0.540	0.124	0.453	0.130	0.087
	Constant	NC	0.597	0.356	0.060	0.166	0.061	0.190
		40	0.609	0.367	0.107	0.229	0.100	0.138
		90	0.622	0.362	0.102	0.335	0.106	0.027
	Decreasing	NC	0.396	0.237	0.045	0.166	0.045	0.072
		40	0.403	0.240	0.072	0.197	0.066	0.044
		90	0.419	0.245	0.095	0.235	0.091	0.010

Note. NC = no censoring; SE = standard error; AFT = accelerated failure time model; COX PH = Cox proportional hazards model; SE = empirical standard error of estimated indirect effect across simulations.

mediator. For the AFT model, *ab* showed percentages of bias close to 0% across all scenarios, except for a rare outcome and a constant hazard (Figure 2A). In this scenario, the underlying indirect effect is overestimated for sample sizes of 250 and 500. When the sample size was 1,000, the percentage of bias approached 0%. When estimating *c-c'* based on the AFT model (Figure 2C), the no-censoring scenarios showed a percentage bias close to 0%. However, in the presence of right censoring, *c-c'* was a consistent underestimation of the underlying indirect effect. For the scenarios with 90% censoring, the percentage of bias decreased as the sample size increased.

When calculating *ab* based on the Cox PH model, only in the scenarios with a constant hazard, a percentage of bias close to 0% was obtained (Figure 2B). In the scenarios with an increasing hazard, *ab* was an overestimation of the indirect effect, that is, high positive percentages of bias, whereas in scenarios with a decreasing hazard, *ab* was an underestimation of the indirect effect, that is, high negative percentage bias. There was no apparent effect of the censoring rate or sample size on the percentage bias for *ab*. The *c-c'* estimate based on the Cox PH model (Figure 2D) showed a high percentage of bias for all scenarios. The only exception was the scenario with an increasing hazard and 90% censoring. In all other scenarios, the indirect effect was consistently underestimated.

In terms of variability, except for situations with 90% censoring, the *ab* and *c-c'* estimates based on the AFT model showed MSE values close to zero, indicating low variability in the estimates (see Supplementary Table S1). In addition, the *c-c'* estimate based on the Cox model showed MSE values close to zero. However, high MSE values were observed for the *ab* estimate based on the
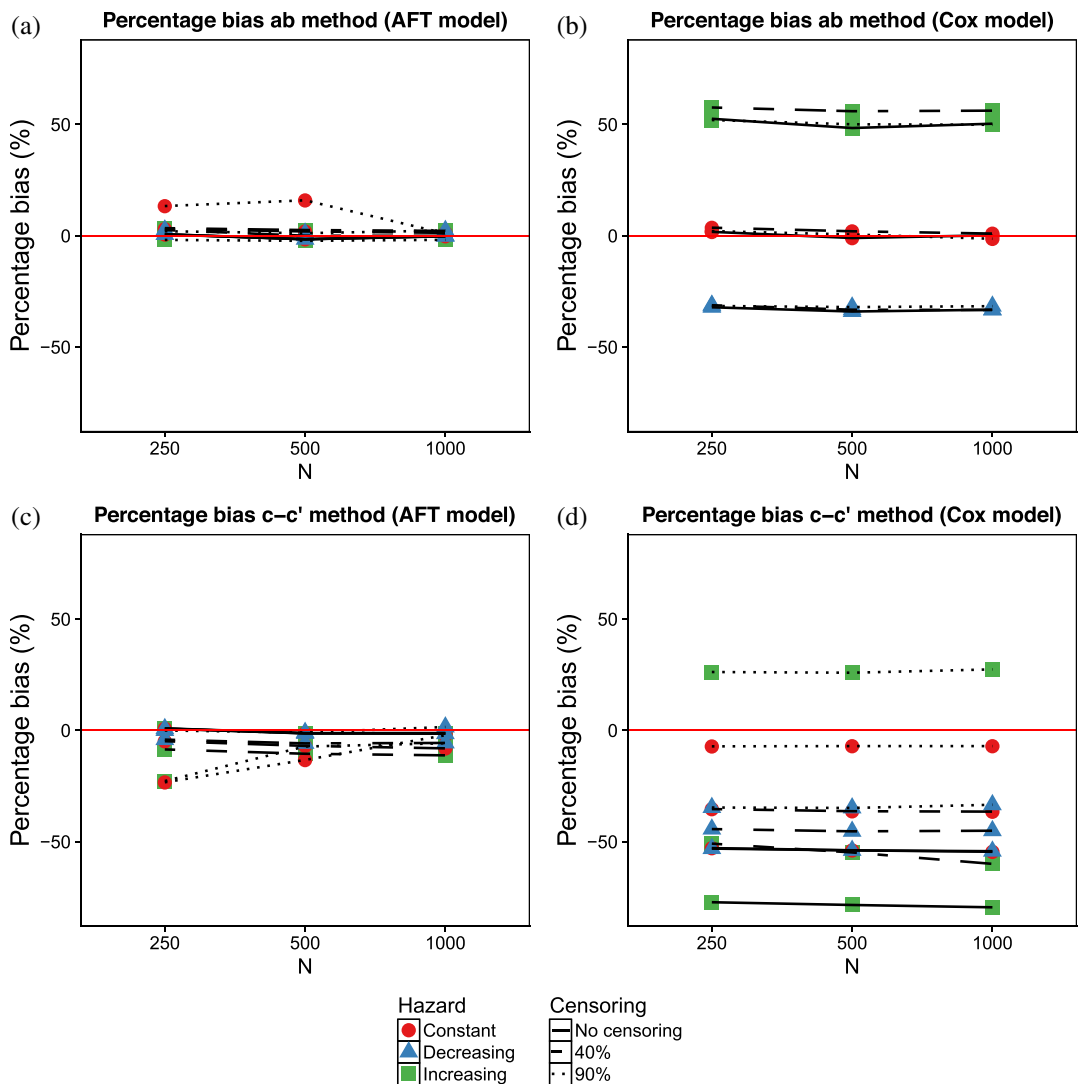
FIGURE 2 Percentage bias in the indirect effect for the traditional mediation approach based on both the accelerated failure time (AFT) model and the Cox proportional hazards model for scenarios with a continuous exposure and mediator

Cox model and for the estimates based on the AFT models when there was a high censoring rate, that is, 90%, indicating high variability in the estimates (see Supplementary Table S2).

Results for other exposure-mediator–type combinations generally followed similar patterns with the AFT model performing better than the Cox PH model (see Supplementary Tables S3–S8). However, in general the observed bias was higher for models with a binary exposure or mediator variable than for models with a continuous exposure and outcome.

3.1.2 | Potential outcomes framework

Table 2 shows the estimates obtained from the simulations of the potential outcomes frame-work based on both the Cox PH model and the AFT model for a sample size of 500 and a model

Model	Hazard	Censoring (%)	NDE	NIE	SE NIE
AFT	Increasing	NC	0.597	0.363	0.066
		40	0.594	0.363	0.120
		90	0.558	0.324	0.166
	Constant	NC	0.598	0.364	0.068
		40	0.588	0.360	0.097
		90	0.609	0.466	1.140
	Decreasing	NC	0.599	0.368	0.073
		40	0.591	0.367	0.089
		90	0.662	0.503	0.133
COX PH	Increasing	NC	0.632	0.392	0.065
		40	0.710	0.434	0.141
		90	0.841	0.510	0.128
	Constant	NC	0.493	0.300	0.053
		40	0.531	0.318	0.082
		90	0.630	0.351	0.109
	Decreasing	NC	0.365	0.217	0.042
		40	0.383	0.225	0.056
		90	0.475	0.259	0.085

TABLE 2 Direct and indirect effect estimates from potential outcomes framework based on the AFT and Cox PH models for the scenarios with a continuous mediator and outcome (N = 500)

Note. NC = no censoring; NDE = natural direct effect; NIE = natural indirect effect; SE = standard error; AFT = accelerated failure time model; COX PH = Cox proportional hazards model; SE = empirical standard error of estimated natural indirect effect across simulations.

with a continuous exposure and mediator variable. For the scenarios without censoring or moderate censoring levels (40%), the NIE based on the AFT model approximated the underlying simulated indirect effect of 0.36. For the scenarios with a high level of censoring (90%), the NIE estimates based on the AFT model showed high bias. The NDE estimates were generally close to the underlying direct effect of 0.6.

The NIE estimates based on the Cox PH model overestimated the indirect effect of 0.36 for scenarios with an increasing hazard and underestimated the indirect effect in scenarios with a constant or decreasing hazard. A similar pattern was observed for the NDE estimates.

In terms of precision, the scenarios without censoring obtained the lowest SE for the NIE for both the AFT and Cox PH models. The simulation results for sample sizes of 250 and 1,000 and for other exposure-mediator type combinations showed similar patterns to the results presented above. Detailed simulation results for the potential outcomes framework are provided in the Supplementary Tables S9–S16.

Figure 3 shows the percentage bias for the NIE estimates based on the AFT model (Figure 3A) and the Cox PH model (Figure 3B) for scenarios with a continuous exposure and mediator. When using the potential outcomes approach with the AFT model, the percentage bias in the NIE approximated 0% for the scenarios with no censoring or 40% censoring (Figure 3A). In contrast, with a censoring rate of 90% the percentage bias was higher, that is, overestimated the true indirect effect. This overestimation decreased with increasing sample size.

For the Cox PH model, the NIE was biased in almost all scenarios (Figure 3B). The NIE was overestimated in scenarios with an increasing hazard, and underestimated in scenarios with a constant or decreasing hazard. These results remained roughly stable across the three sample sizes.

In terms of variability, the NIE estimates based on the AFT model showed lower MSE values than the estimates of the Cox PH model (Supplementary Tables S9–S16 in the Supporting
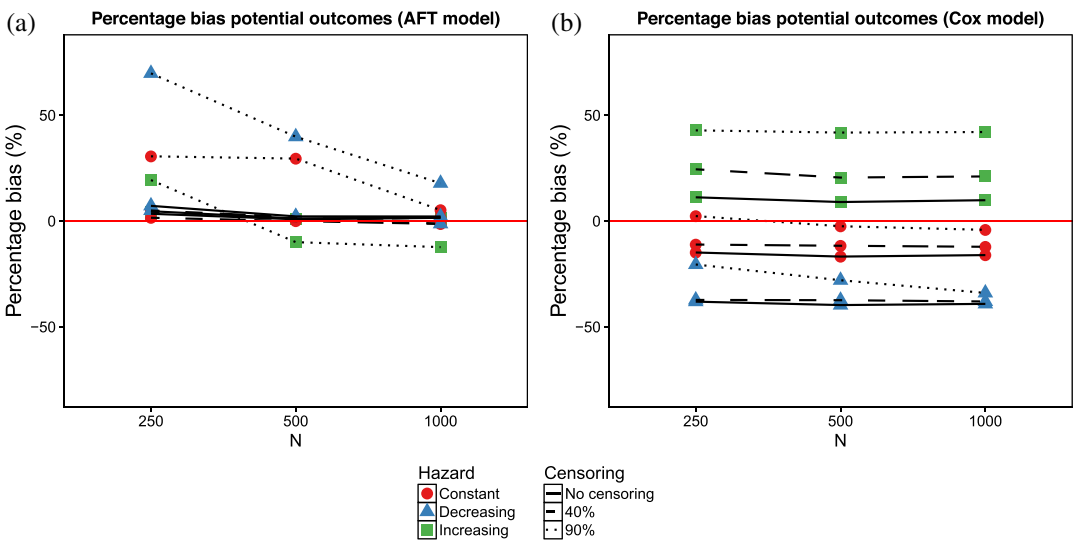
FIGURE 3 Percentage bias in the indirect effect for the potential outcomes framework based on both the accelerated failure time (AFT) model and the Cox proportional hazards model for scenarios with a continuous exposure and mediator

Information). However, high MSE values were observed for the estimates based on the AFT models when there was a high censoring rate, that is, 90%.

Results with similar patterns were found in scenarios with other exposure-mediator–type combinations (Supplementary Tables S11–S16 in the Supporting Information). Results for other exposure-mediator–type combinations generally followed similar patterns with the AFT model performing better than the Cox PH model (Supplementary Tables S3-S8 in the Supporting Information). However, differences were observed between the indirect effect estimates based on traditional and potential outcomes mediation methods when the mediator was binary.

3.2 | Application to the real-life data example

Out of the 1,122 participants, 292 (26%) experienced recurrence of a depressive disorder during the study duration and the remaining 830 (74%) were right-censored, corresponding to a moderate-to-high study-duration censoring rate. The mean follow-up duration was 30.0 months with an SD of 13.4. A graphical inspection of the hazard rate revealed an increasing hazard rate.

Before fitting the Cox PH and AFT models, we assessed whether the main assumptions of each model were met. For the Cox PH model, the PH assumption was checked using the Schoenfeld residuals test and visual inspection of the graphs based on the scaled Schoenfeld residuals. Based on the Schoenfeld Residuals Test, we did not find a significant relationship between residuals and time. Furthermore, the residual plots displayed a random pattern. Therefore, there were no signs of violation of the PH assumption in the data. The AFT model relies on the assumption that the survival times follow a specific parametric error distribution. In real-life data, the underlying error distribution of the data is unknown. To assess the goodness- of-fit of the Weibull AFT model, we compared the Akaike information criterion (AIC) of the model based on Equation (2) across different potential error distributions: Weibull, exponential, log- normal, and log-logistic. The Weibull model showed the best fit with the smallest AIC value.

Table 3 displays the summary of the direct, indirect, and TE estimates obtained through the four mediation analysis approaches. The 95% percentile bootstrap confidence intervals were

TABLE 3 Results of the application of the four methods for mediation analysis to the real-life data example

Approach Traditional	Model[a]	Direct effect		Indirect effect				Total effect	
		c'	95% CI	ab	95% CI	c-c'	95% CI	c	95% CI
	AFT	−0.039	−0.002, 0.175	−0.038	−0.054, −0.029	−0.041	−0.063, −0.028	−0.080	−0.092, −0.025
	COX PH	0.082	−0.017, 0.158	0.120	0.071, 0.176	0.103	0.050, 0.155	0.185	0.080, 0.261
Potential outcomes		NDE	95% CI	NIE	95% CI			TE	95% CI
	AFT	−0.021	−0.001, −0.094	−0.040	−0.090, −0.056			−0.061	−0.055, −0.010
	COX PH	0.060	−0.003, 0.32	0.080	0.046, 0.094			0.140	0.078, 0.356

Note. CI = confidence interval; AFT = accelerated failure time model; COX PH = Cox proportional hazards model; NDE = natural direct effect, NIE = natural indirect effect, TE = total effect.

[a]All models are adjusted for age, sex, years of education, and recency of last episode of depressive and/or anxiety disorder at baseline.

calculated for the indirect effect. It should be stressed that the analyses performed and results presented in this section are meant as an illustration tool. For a full discussion of the clinical implications, the reader is referred to the original paper by Gerrits et al. (2014).

For both traditional mediation analysis and the potential outcomes framework, the obtained AFT model estimates are on the natural logarithm scale. Exponentiation yields effect estimates that represent the ratio of the mean survival time between two different values of the exposure. For both traditional mediation analysis and the potential outcomes framework, the Cox PH model obtained estimates are on the natural logarithm of the hazard ratio scale.

For the traditional method, the exponentiated direct effect based on the AFT model equals 0.97, which means that, for one unit increase in the CPG score, the time to recurrence of depressive disorder is adjusted 0.97 times higher for the subthreshold depressive symptoms. In other words, when the CPG score increases, the time to recurrence of depressive disorder decreases. The *ab* and *c-c'* estimates are −0.03 and − 0.04, respectively. The exponentiated indirect effect *ab* for the traditional method based on the AFT model equals 0.97, which means that, for one unit increase in the CPG score, the time to recurrence of depressive disorder is 0.97 times higher through an increase in subthreshold depressive symptoms. The exponentiated direct effect based on the Cox PH model equals 1.08, which means that, for one unit increase in the CPG score, the risk of recurrence of the depressive disorder at any time point is adjusted 1.08 times higher for the subthreshold depressive symptoms. In other words, when the CPG score increases, the risk of recurrence of the depressive disorder increases. The results based on the AFT model and the Cox PH model do therefore point in the same direction. The *ab* and *c-c'* estimates are 0.12 and 0.10, respectively. The exponentiated indirect effect *ab* for the traditional method based on the Cox PH model equals 1.13, which means that, for one unit increase in the CPG score, the risk of recurrence of the depressive disorder at any time point is 1.13 times higher through an increase in subthreshold depressive symptoms.

For the potential outcomes framework, the effect estimates based on the AFT model can be interpreted in a similar way as the effect estimates based on traditional mediation analysis. The unexponentiated TE in the traditional method equals −0.08. Note that the summation of the direct effect and *ab* does not add up to −0.08 but to −0.06. In the potential outcomes framework, however, the NDE and NIE do add up to the TE. The effect estimates from the potential outcomes framework based on the Cox PH model can be interpreted in a similar way as the effect estimates based on traditional mediation analysis. Note that, also for the traditional method based on the Cox PH model, the summation of the direct effect and *ab* does not add up to 0.18 but to 0.20. Again, the NDE and NIE in the potential outcomes framework do add up to the TE.

4 | DISCUSSION

The aim of this paper was to investigate the performance of traditional mediation analysis and potential outcomes mediation analysis based on both the Cox PH and the AFT models, with respect to their own underlying indirect effect. In traditional mediation analysis based on the AFT model, the lowest percentage bias and lowest MSE were obtained for the *ab* estimate. The NDE and NIE based on the potential outcomes framework with the AFT model obtained a low percentage bias and low MSE except in the scenarios with a rare outcome. Both traditional mediation analysis and the potential outcomes framework based on the Cox PH model obtained estimates with a higher MSE than the AFT-based estimates. Only in two scenarios, the Cox PH model approximated the corresponding underlying indirect effect, when estimating *ab* based on

traditional mediation analysis in scenarios with constant hazards, and when estimating the NIE on the potential outcomes framework in scenarios with rare outcomes and constant hazards.

It is notable that both traditional and potential outcomes framework methods based on the AFT model with the Weibull distribution yielded more precise estimates (lower MSE and SE) than those based on the Cox model. However, this gain in precision comes at a price: the AFT being a parametric model poses more assumptions than the Cox model, and the violation of the Weibull assumption will result in biased estimates.

The *ab* estimate from traditional mediation analysis based on Cox PH model only had low bias in scenarios with a constant hazard (Figures 2B and 2D). A possible explanation for this result is that, in our study, the survival times were simulated following a Weibull distribution. When the shape parameter (i.e., hazard rate) is set to constant hazard, the Weibull distribution reduces to the exponential distribution (Lee & Wang, 2003), which may result in less biased estimates. Further research is needed to fully explain these results. It is, however, important to note that in practice it is uncommon to observe a constant hazard, as there are several factors that might affect the hazard rate during the study (Gürler, 2014), for example, the length of follow-up, intrinsic characteristics of a specific health condition, characteristics of treatment, and even selection bias (Hernan, 2010).

Within the traditional mediation approach, we also compared the *ab* and *c-c'* estimates. We observed that scenarios without censoring *ab* and *c-c'* based on the AFT model were approximately equivalent and both were minimally biased. This equality did not hold in situations that included censoring. In presence of right censoring, *c-c'* based on the AFT model underestimated the indirect effect, whereas *ab* provided accurate estimates. This is in line with the results from previous studies (Fulcher et al., 2017; Gelfand et al., 2016; Tein & MacKinnon, 2003). As suggested in previous research (VanderWeele, 2011), we observed a decreased difference between *ab* and *c-c'* for rare outcomes. However, for common outcomes, the equality between *ab* and *c-c'* did not hold for Cox PH models. Lange and Hansen (2011) described a method for mediation analysis with time-to-event outcomes, based on additive hazard models, which does not pose the rare outcome assumption. Unfortunately, this method has not yet been implemented in any commonly used software packages.

In our real-life data example, we also observed differences between the *ab* and *c-c'* estimates for both the Cox PH and AFT models. As expected, based on our simulation study and previous research (Fulcher et al., 2017; Gelfand et al., 2016; Tein & MacKinnon, 2003), the *c-c'* estimate was an underestimation of the *ab* estimate for both model types. In the real-life data example, we also observed that the TE in the traditional mediation analysis, that is, *c*, did not equal the sum of *c'* and *ab*. In contrast, the TE in the potential outcomes framework did equal the summation of the NDE and the NIE.

The nonequivalence of the *ab* and *c-c'* estimates, the nonequivalence of *c* and *ab + c'*, and the large bias in the Cox PH estimates can be explained by noncollapsibility. Noncollapsibility is related to the total variance in the models for the outcome, that is, Equations (1) and (2) (Mood, 2010). In these generalized linear models, the residual variance is fixed. As a consequence, the total variance in the model is dependent on the independent variables in the model. The more independent variables are included in the model, the higher the explained variance and, consequently, the higher the total variance. The magnitude of the coefficients in generalized linear models is related to the total variance. When the total variance increases, the magnitude of the coefficients will also increase. The total variance of the model based on Equation (2) will therefore be higher than the total variance of the model based on Equation (1). As a consequence, coefficients from these two models will have different magnitudes and, therefore, the use of *c-c'* is not

recommended for generalized linear models (MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007). Furthermore, as a consequence of noncollapsibility, also the c coefficient does not equal the summation of c' and ab. As shown in the potential outcomes literature, the summation of c' and ab is generally the preferred estimate of the TE (VanderWeele, 2011).

To assess the performance of the four reviewed mediation analysis methods, we simulated scenarios with various hazard shapes and censoring conditions. In our study, we only considered noninformative study-duration censoring. However, in practice, participants may drop out before the end of the study, which leads to drop-out censoring. Gelfand et al. (2016) included noninformative drop-out censoring in their simulations and concluded that drop-out censoring affects the estimates less than study-duration censoring. Future studies might consider scenarios with informative drop-out censoring, as Gelfand et al. (2016) only considered noninformative drop-out censoring. Informative censoring occurs when participants are lost to follow-up due to reasons related to the study (Ranganathan & Pramesh, 2012), for example, if more severely depressed patients are less likely to participate in the follow-up of the study. Furthermore, more research is needed to assess the robustness of the Weibull AFT model in scenarios where the Weibull assumption is violated, and on the performance of AFT models with other types of error distributions than the Weibull model. The flexibility of the potential outcomes mediation framework goes beyond the scenarios described in this paper as it can be applied to situations where there is an exposure-mediator interaction (Rijnhart, Twisk, Eekhout, & Heymans, 2019), and when the effects are in terms of differences instead of ratios.

Previous research has shown that, in the absence of exposure-mediator interaction and for continuous mediators and survival outcomes estimated with a Cox PH or AFT model, the NDE in the regression-based estimation reduces to the c' coefficient, and the NIE reduces to ab when there is no unobserved confounding and no exposure-mediator interaction. The NIE has been shown to reduce to the product of the a and b coefficients (Lange et al., 2012; VanderWeele, 2011). When the mediator is binary, the traditional and potential outcomes mediation analysis will yield different indirect effect estimates (Rijnhart et al., 2019). The difference can be explained by the difference in the formulas for the indirect effect between traditional and potential outcomes mediation analysis (VanderWeele, 2015). Further research is needed to investigate why and when the indirect effect estimates of these two methods differ.

This paper presented methods for mediation analysis with a time-to-event outcome that are based on either the AFT model or the Cox PH model. The choice for an effect measure, that is, the hazard ratio from a Cox PH model or the ratio of mean survival times from an AFT model, should primarily depend on the scientific context of the mediation analysis. In other words, the effect measure should match the research question at hand. When using either the Cox or AFT models with any of the two mediation frameworks, it is necessary to check the corresponding assumptions. The PH assumption can be checked using the Schoenfeld residuals test and visual inspection of the graphs based on the scaled Schoenfeld residuals. The Weibull assumption can be checked using either of the following two methods: (1) comparing the AIC from AFT models with different specified error distributions and select the error distribution with the lowest AIC, and (2) visual inspection of the Cox–Snell residuals. A further explanation on how to check the Cox and AFT model assumptions can be found in Kleinbaum and Klein (2010). Furthermore, in the Supporting Information, we provide an example R code for mediation analysis with time-to-event outcomes, which includes code to check the PH and Weibull assumptions.

If the Weibull assumption is violated, other distributions for the AFT model could be considered, such as the exponential, log-normal, and log-logistic error distribution. Fulcher et al. (2017) have shown that the product-of-coefficients method used with the AFT model might be minimally

biased for most choices of error distributions, such as exponential and log-logistic error distributions. For a situation with a normal mediator and a log-normal outcome, both the product and difference method provide minimally biased indirect effect estimates (Fulcher et al., 2017).

Various software packages can be used to estimate the models presented in this paper. In R, traditional mediation analysis based on the Cox PH and AFT models can be performed using the *coxph* and *survreg* functions, respectively, from the *survival* package (Therneau, 2015). In the Supporting Information, we provide example syntax for this. In SAS, traditional mediation analysis can be performed using the SAS *PHREG* procedure for the Cox model and the *LIFEREG* procedure for the AFT model. In Stata, traditional mediation analysis with Cox and AFT models can be performed using the *stcox* and the *streg* commands, respectively. In SPSS, traditional mediation analysis for the Cox model is performed using the *COXREG* command. Unfortunately, SPSS does not allow the use of the AFT model. Lange et al. (2017) provide R code for the implementation of the simulation-based potential outcomes framework for both Cox PH and AFT models. Valeri and VanderWeele (2015) developed a SAS macro for the regression-based approach for estimating the NDE and NIE with the potential outcomes framework based on both the AFT and Cox PH models. Unfortunately, we are not aware of automated procedures to implement the potential outcomes mediation approach with time-to-event data in SPSS or Stata.

In conclusion, both traditional mediation analysis and the potential outcomes framework yield unbiased effect estimates with respect to their own underlying true value for simple mediation models with a time-to-event outcome and estimated based on Cox PH or AFT.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Lizbeth Burgos Ochoa [ID] https://orcid.org/0000-0002-8379-2749

REFERENCES

Ambler, G., Seaman, S., & Omar, R. Z. (2012). An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in Medicine*, *31*(11–12), 1150–1161. https://doi.org/10.1002/sim.4371

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292. https://doi.org/10.1002/sim.2673

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Oxfordshire, UK: Taylor & Francis.

Crowther, M. J., & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, *32*(23), 4118–4134. https://doi.org/10.1002/sim.5823

Fizazi, K., Scher, H. I., Molina, A., Logothetis, C. J., Chi, K. N., Jones, R. J., & Saad, F. (2012). Abiraterone acetate for treatment of metastatic castration-resistant prostate cancer: Final overall survival analysis of the COU-AA-301 randomised, double-blind, placebo-controlled phase 3 study. *The Lancet Oncology*, *13*(10), 983–992.

Fulcher, I. R., Tchetgen Tchetgen, E. J., & Williams, P. L. (2017). Mediation analysis for censored survival data under an accelerated failure time model. *Epidemiology*, *28*(5), 660–666. https://doi.org/10.1097/ede.0000000000000687

Gelfand, L. A., MacKinnon, D. P., DeRubeis, R. J., & Baraldi, A. N. (2016). Mediation analysis with survival outcomes: Accelerated failure time vs. proportional hazards models. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00423

George, B., Seals, S., & Aban, I. (2014). Survival analysis and regression models. *Journal of Nuclear Cardiology: Official Publication of the American Society of Nuclear Cardiology*, *21*(4), 686–694. https://doi.org/10.1007/s12350-014-9908-2

Gerrits, M. M. J. G., van Oppen, P., Leone, S. S., van Marwijk, H. W. J., van der Horst, H. E., & Penninx, B. W. (2014). Pain, not chronic disease, is associated with the recurrence of depressive and anxiety disorders. *BMC Psychiatry*, *14*(1). https://doi.org/10.1186/1471-244x-14-187

Gürler, Ü. (2014). Hazard change point estimation. *Wiley StatsRef: Statistics Reference Online*.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, *21*(1), 13–15.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334.

Kleinbaum, D. G., & Klein, M. (2010). *Survival analysis: A self-learning text* (3rd ed.). New York, NY: Springer.

Lange, T., & Hansen, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology*, 575–581.

Lange, T., Hansen, K. W., Sørensen, R., & Galatius, S. (2017). Applied mediation analyses: A review and tutorial. *Epidemiology and Health*, *39*.

Lange, T., Vansteelandt, S., & Bekaert, M. (2012). A simple unified approach for estimating natural direct and indirect effects. *American Journal of Epidemiology*, *176*(3), 190–195. https://doi.org/10.1093/aje/kwr525

Lapointe-Shaw, L., Bouck, Z., Howell, N. A., Lange, T., Orchanian-Cheff, A., Austin, P. C., … Bell, C. M. (2018). Mediation analysis with a time-to-event outcome: A review of use and reporting in healthcare research. *BMC Medical Research Methodology*, *18*(1), 118. https://doi.org/10.1186/s12874-018-0578-7

Lee, E. T., & Wang, J. W. (2003). *Statistical methods for survival data analysis: Lee/survival data analysis*. Hoboken, NJ: John Wiley & Sons.

MacKinnon, D. P. (2012). *Introduction to statistical mediation analysis*. New York, NY: Routledge.

MacKinnon, D. P., Lockwood, C. M., Brown, C. H., Wang, W., & Hoffman, J. M. (2007). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, *4*(5), 499–513.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41–62.

Martinussen, T., Vansteelandt, S., Gerster, M., & Hjelmborg, J. v. B. (2011). Estimation of direct effects for survival data by using the Aalen additive hazards model: Estimation of direct effects for survival data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(5), 773–788. https://doi.org/10.1111/j.1467-9868.2011.00782.x

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82.

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2017). *Regression and mediation analysis using Mplus*. Los Angeles, CA: Muthén & Muthén.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420.

Penninx, B. W. J. H., Beekman, A. T. F., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., … Van Dyck, R. (2008). The Netherlands study of depression and anxiety (NESDA): Rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, *17*(3), 121–140. https://doi.org/10.1002/mpr.256

R Core Team. (2017). *R: A language and environment for statistical computing*.

Ranganathan, P., & Pramesh, C. S. (2012). Censoring in survival analysis: Potential for bias. *Perspectives in Clinical Research*, *3*(1), 40.

Richiardi, L., Bellocco, R., & Zugna, D. (2013). Mediation analysis in epidemiology: Methods, interpretation and bias. *International Journal of Epidemiology*, *42*(5), 1511–1519. https://doi.org/10.1093/ije/dyt127

Rijnhart, J. J. M., Twisk, J. W. R., Eekhout, I., & Heymans, M. W. (2019). Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. *BMC Medical Research Methodology*, *19*(1), 19.

Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155.

Robins L. N., Wing, J., Wittchen, H., Helzer, J. E., Babor, T. F., Burke, J., … Towle, L. H. (1988). The composite international diagnostic interview: An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry*, *45*(12), 1069–1077.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., & Manber, R. (2003). The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, *54*(5), 573–583.

Sedlis, S. P., Hartigan, P. M., Teo, K. K., Maron, D. J., Spertus, J. A., Mancini, G. B. J., & Lorin, J. D. (2015). Effect of PCI on long-term survival in patients with stable ischemic heart disease. *New England Journal of Medicine*, *373*(20), 1937–1946.

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(4), 547–559. https://doi.org/10.1037/a0028226

Swindell, W. R. (2009). Accelerated failure time models provide a useful statistical framework for aging research. *Experimental Gerontology*, *44*(3), 190–200.

Tchetgen Tchetgen, E. J. (2013). Inverse odds ratio-weighted estimation for causal mediation analysis. *Statistics in Medicine*, *32*(26), 4567–4580. https://doi.org/10.1002/sim.5864

Tein, J.-Y., & MacKinnon, D. P. (2003). Estimating mediated effects with survival data. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics: Proceedings of the International Meeting of the Psychometric Society IMPS2001 Osaka, Japan, July 15–19, 2001* (pp. 405–412). Tokyo, Japan: Springer.

Therneau, T. M. (2015). *A package for survival analysis in S (Version 2.38)*. Retrieved from https://CRAN.R-project.org/package=survival

Valeri, L., & VanderWeele, T. J. (2015). SAS macro for causal mediation analysis with survival data. *Epidemiology*, *26*(2), e23–e24. https://doi.org/10.1097/EDE.0000000000000253

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology*, *22*(4), 582–585. https://doi.org/10.1097/EDE.0b013e31821db37e

VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford, UK: Oxford University Press.

VanderWeele, T. J. (2016a). Explanation in causal inference: Developments in mediation and interaction. *International Journal of Epidemiology*, *45*, 1904–1908. https://doi.org/10.1093/ije/dyw277

VanderWeele, T. J. (2016b). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, *37*(1), 17–32. https://doi.org/10.1146/annurev-publhealth-032315-021402

Vansteelandt, S., Bekaert, M., & Lange, T. (2012). Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiologic Methods*, *1*(1). https://doi.org/10.1515/2161-962x.1014

Von Korff, M., Ormel, J., Keefe, F. J., & Dworkin, S. F. (1992). Grading the severity of chronic pain. *Pain*, *50*(2), 133–149.

Zhang, Z. (2016). Parametric regression model for survival data: Weibull regression model as an example. *Annals of Translational Medicine*, *4*(24), 484–484. https://doi.org/10.21037/atm.2016.08.45

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: Burgos Ochoa L, Rijnhart JJM, Penninx BW, Wardenaar KJ, Twisk JWR, Heymans MW. Performance of methods to conduct mediation analysis with time-to-event outcomes. *Statistica Neerlandica*. 2019;1–20. https://doi.org/10.1111/stan.12191