Published in final edited form as:

Stat Methods Med Res. 2018 October ; 27(10): 3010–3025. doi:10.1177/0962280217690414.

Estimation of Indirect Effect When the Mediator is A Censored Variable

Jian Wang¹ and Sanjay Shete^{1,2}

¹Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A

²Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A

Abstract

A mediation model explores the direct and indirect effects of an initial variable (X) on an outcome variable (Y) by including a mediator (M). In many realistic scenarios, investigators observe censored data instead of the complete data. Current research in mediation analysis for censored data focuses mainly on censored outcomes, but not censored mediators. In this study, we proposed a strategy based on the accelerated failure time model and a multiple imputation approach. We adapted a measure of the indirect effect for the mediation model with a censored mediator, which can assess the indirect effect at both the group and individual levels. Based on simulation, we established the bias in the estimations of different paths (i.e., the effects of X on $M[a]$, of M on $Y[b]$ and of X on Y given mediator $M[c]$) and indirect effects when analyzing the data using the existing approaches, including a naïve approach implemented in software such as Mplus, complete-case analysis, and the Tobit mediation model. We conducted simulation studies to investigate the performance of the proposed strategy compared to that of the existing approaches. The proposed strategy accurately estimates the coefficients of different paths, indirect effects and percentages of the total effects mediated. We applied these mediation approaches to the study of SNPs, age at menopause and fasting glucose levels. Our results indicate that there is no indirect effect of association between SNPs and fasting glucose level that is mediated through the age at menopause.

Keywords

Mediation analysis; censored mediator; accelerated failure time; multiple imputation; indirect effects

---

CORRESPONDING AUTHOR: Address for correspondence and reprints: Dr. Sanjay Shete, Department of Biostatistics, Unit 1411, The University of Texas MD Anderson Cancer Center, 1400 Pressler Dr, FCT4.6002, Houston, TX 77030, U.S.A. Phone: (713) 745-2483, Fax: (713) 563-4243, sshete@mdanderson.org.

Declaration of Conflicting Interests

The authors declare that there is no conflict of interest.

Introduction

A mediation model is a statistical approach that explores the direct and indirect effects of an initial variable on an outcome variable by including one or more mediating variables¹. In some scenarios, the mediation model can infer the causal effects from the initial variable to the mediator variable and then to the outcome variable^{1, 2}. Figure 1 (A) depicts a single mediator model that can be specified using the following three regressions^{1, 3–5}:

$$\begin{aligned} Y &= c_0 + cX + \varepsilon_1 \\ M &= a_0 + aX + \varepsilon_2 \\ Y &= b_0 + bM + c'X + \varepsilon_3 \end{aligned} \quad (1)$$

In this single mediator model, X denotes the initial variable, M denotes the mediating variable, Y denotes the outcome variable, and ε_1 , ε_2 and ε_3 are the residuals in the model. Mediation analysis decomposes the total effect of X on Y (path c) into a direct effect (path c') and an indirect effect (usually calculated as ab or $c-c'$); see Figure 1 (A).

Mediation models have been widely applied in many different fields, such as psychology, social science, health science, genetic epidemiology, prevention research, and political communication research^{4, 6–8}. In these research fields, investigators might observe censored data instead of the complete data^{9, 10}. For example, in epidemiologic cohort studies, an individual's age at the onset of disease is a right-censored variable¹⁰. Also for instance, some blood biomarkers (e.g., c-reactive protein concentration), are subject to left censoring because of the detection limitation for small values¹¹. Currently, most approaches to mediation analysis involving censored variables focus on the outcome variable Y being a censored variable^{10, 12–15}. However, the mediator in mediation analysis can also be a censored variable. For example, genes may impact the age at which an individual quits smoking (censored for current smokers), which in turn has been associated with lung cancer risk¹⁶. Therefore, the age at which an individual quits smoking may have a mediating effect on the association between genes and lung cancer risk. Similarly, the individual's age at smoking initiation might be a potential mediator for the association between genes and cardiovascular disease^{17, 18}. Previous studies have shown that several genes are associated with a woman's age at menopause^{19, 20}, and some studies have suggested that the age at menopause is associated with type-2 diabetes^{21–23}, which suggests the potential mediating effect of age at menopause on the association between genetic variants and type-2 diabetes. The conceptual mediation model of this example is shown in Figure 1 (B). Few approaches have been suggested for mediation analysis when the mediator is a censored variable. First, we review the existing approaches and point out their limitations.

One naïve approach, as implemented in the popular software program Mplus²⁴, is to estimate the paths b and c' using standard regression, which simply ignores the fact that M is a censored variable and uses all the data regardless of the censoring status. However, it is well established that simply using a censored covariate in regressions will lead to bias in the estimation of the corresponding coefficients for paths b and c' ^{41, 25–27}. Therefore, the results

obtained using the naïve approach are not valid. Furthermore, the naïve approach (e.g., Mplus) uses the Cox proportional model to assess the coefficient for path a . However, the estimated value of a obtained using the Cox proportional model does not have the same interpretation as the path a from the standard linear regression analyses such that ab is an indirect effect.

A second approach is complete-case analysis, in which data on subjects with censored covariates are discarded. This approach can provide a consistent but inefficient estimated value for path b , especially when the censoring percentage is high^{11, 25, 27}. Another shortcoming for the complete-case approach to mediation modeling is that it leads to a biased estimation of the coefficient of path a ²⁸.

A third existing approach is the Tobit mediation model proposed by Wang and Zhang⁹, which uses structural equation modeling with a truncated multivariate normal distribution to account for the censored data in the mediation analysis. In their model, any of the initial, mediator or outcome variable can be a censored variable, and is assumed to be continuous and normally distributed.

In this paper, we propose a strategy that estimates the coefficient of different paths (a , b and c) more accurately than the three existing approaches. We focus on the mediator that is a right-censored variable. We use simulation studies to establish the potential bias in the estimations of different paths (a , b and c) due to the censored mediator, when directly applying the three existing approaches, which in turn, will lead to bias in the estimation of the indirect effects. To accurately estimate the direct and indirect effects, we suggest using the accelerated failure time (AFT) model to evaluate the effect of the initial variable on the censored mediator variable (path a). To assess the coefficients of paths b and c , we use a multiple imputation approach designed for regression models with censored covariates. This approach imputes the censored covariates by their conditional quantiles given the observed data¹¹.

The measures commonly used for the indirect effects, such as the difference in regression coefficients $c-c'$ and the product of regression coefficients ab , are only applicable when both the mediator and outcome are continuous and normally distributed. The indirect effects have been investigated in previous studies when the outcome is a censored variable^{10, 12, 13}. However, when the mediator is a censored variable, the definition of the indirect effects is unclear. One may consider using the difference of $c-c'$ to measure the indirect effect based on two linear regressions if we assume a continuous outcome. Since it has been shown that the difference of $c-c'$ is equivalent to the product of ab for a continuous mediator²⁹, such a definition for the indirect effect ($c-c'$) implies the use of a linear probability model to model the censored mediator in order to assess the coefficient of a , which might not be valid in the mediation model with a censored mediator (e.g., the censored mediator has a Weibull distribution). Therefore, we adapt a measure of the indirect effect from the study of Li et al.³⁰, which was proposed for the binary mediator, for the mediation model with a censored mediator. The indirect effect measure proposed by Li et al.³⁰ is defined as the change in the mean value of Y associated with each unit change in X that can be attributed to the mediator M . This definition does not require the mediator to be continuous and normally distributed,

and therefore can be extended to the model with a censored mediator. This measure can assess the indirect effect at the study sample level as well as at the individual level for each subject³⁰.

We use simulated data to compare the estimated coefficients (a , b and c), the indirect effects and percentages of the total effects mediated that were obtained from our strategy to those obtained from the existing approaches, including the naïve approach, complete-case analysis, and the Tobit mediation model. We also apply the existing and proposed approaches to investigate the mediating effect of the age at menopause on the association between genetic variants and fasting glucose level, a quantitative measure of type-2 diabetes (Figure 1 (B)). Our findings show that there was no indirect effect of association between single nucleotide polymorphisms (SNPs) and fasting glucose level that is mediated through the age at menopause.

Methods

Proposed Strategy for Estimating Coefficients

We considered the mediation model shown in Figure 1 (A), in which X is the initial variable, M is the mediator and Y is the outcome variable. In this study, we assumed that both variables X and Y are continuous variables, while M is a time-to-event right-censored variable. That is, for each individual, we observe $m_i = \min(t_i, c_i)$, $i = 1, \dots, n$, where t_i is the time-to-event of interest and c_i is the individual's right censoring time; as well as an indicator $\delta_i = I(t_i \leq c_i)$ that indicates whether the individual's data are censored ($\delta_i = 0$) or observed ($\delta_i = 1$).

Coefficient for Path a: Accelerated Failure Time Model—We used the AFT model to analyze the effect of X on M (i.e., path a) in the mediation model. The major advantage of using the AFT model is that it provides the change in the length of survival as a function of the effect of the per unit change in the initial variable. The AFT model is a parametric model. Given an individual i , the AFT model can be specified using the following log-linear representation^{13, 31, 32},

$$\log(t_i) = a_0 + ax_i + \sigma\epsilon_i, \quad (2)$$

where a_0 and σ are the intercept and scale parameters, respectively, and ϵ_i represents the residual in the log-transformed survival time. For a parametric model, the distribution of the survival time must be specified. The distributions most commonly used are Weibull, log-logistic, log-normal, gamma, and Gompertz³¹. One can select the distribution using model selection approaches, such as the Akaike information criterion³³. In this study, the AFT model was implemented using the R package “survival”³⁴.

Coefficients for Paths b and c': Multiple Imputation—We used a multiple imputation approach to assess the effect of X and M on Y (i.e., the coefficients for paths b and c'). The multiple imputation approach was proposed by Wang and Feng¹¹ for linear regression models with censored covariates. The multiple imputation approach imputes the

censored covariates using the conditional quantiles given the observed data instead of specifying parametric likelihoods. Specifically, in our mediation model, given an individual i , the linear regression model was fitted as

$$y_i = b_0 + bt_i + c'x_i + \varepsilon_i, \quad (3)$$

where, as defined previously, x_i is the initial variable, and t_i is a time variable that is subject to right censoring, and we only observe $m_i = \min(t_i, c_i)$, where c_i is the right censoring time, and an indicator $\delta_i = I(t_i < c_i)$.

The imputation approaches commonly used generally impute the censored variable t_i by randomly sampling from the conditional distribution $F(t|y_i, x_i, t_i > c_i)$, which requires a parametric likelihood for t_i . Wang and Feng¹¹ noticed that this imputation is equivalent to filling in the censored variable t_i by the u th conditional quantile of t_i given the observed variables y_i and x_i , where u is randomly drawn from uniform $(0, \pi(y_i, x_i))$, and $\pi(y_i, x_i) = \Pr(t_i < c_i | y_i, x_i)$. Therefore, they proposed an imputation approach to impute t_i using its conditional quantiles instead of its conditional distribution. Using t_i as a dependent variable, its conditional quantiles were assessed by fitting censored quantile regression models $Q_{t_i}(u|y_i, x_i) = (y_i, x_i)\eta_0(u)$, where $\eta_0(u)$ is the quantile coefficient process that is assumed to be a smooth function of u . The quantile coefficient process can be estimated using established censored quantile regression methods, such as Powell's and Portnoy's estimators^{35, 36}. The imputation step was repeated multiple times generating multiple imputed datasets. The usual complete data approach (linear regression in this case) can be applied to the imputed datasets, and the average of the estimated coefficients from multiple imputed datasets is defined as the multiple imputation estimator. Wang and Feng showed that the resulting multiple imputation estimator is consistent, asymptotically normal, and has improved efficiency compared to that of the complete-case approach.

Definition of Indirect Effect with a Censored Mediator—As mentioned, the commonly used indirect effect measures, such as the product of ab and the difference of $c-c'$, are only applicable when the variables in the mediation model are continuous and normally distributed. The indirect effects have been investigated in previous studies when the outcome is a censored variable^{10, 12, 13}. However, when the mediator is a censored variable, the definition of the indirect effects is unclear. In this study, we adapted a measure from Li et al.³⁰ to provide a definition of the indirect effects in this scenario.

Specifically, Li et al.³⁰ proposed a definition of the indirect effect of the initial variable X on the outcome variable Y as the change in the mean value of Y associated with each unit change in X that could be attributed to the mediator M . Given a continuous initial variable X , the indirect effect can be specified as a partial differential equation

$$IE = \frac{\partial \{E(Y|X) - E(Y|X, M)\}}{\partial X},$$

where $E(Y|X)$ and $E(Y|X, M)$ are conditional expectations, which, based on equation (3), can be given as $E(Y|X) = b_0 + bE(M|X) + c'X$ and $E(Y|X, M) = b_0 + bM + c'X$. Therefore, the indirect effect is $IE = b \times \frac{\partial\{E(M|X)\}}{\partial X}$

. Note that if the mediator is a continuous and uncensored variable, the indirect effect is consistent with the product-of-coefficient approach (i.e., ab).

Using the Weibull distribution for the AFT model (equation (2)) with scale and shape parameters of λ and ν , respectively, the conditional expectation $E(M|X)$ is given as $\exp(a_0 + aX)\Gamma(1 + 1/\nu)$, where $\Gamma(\bullet)$ is the gamma function. Therefore, the indirect effect is $IE = ab\exp(a_0 + aX)\Gamma(1 + 1/\nu)$. We recommend using the AFT model and multiple imputation approach to estimate coefficients $\hat{a}_0$, $\hat{a}$, $\hat{b}$, and $\hat{\nu}$ for estimating the indirect effect (abbreviated as IE in the equation). For an individual i with a value of initial variable x_i , the indirect effect is estimated as $\hat{IE}_i = \hat{a}\hat{b}\exp(\hat{a}_0 + \hat{a}x_i)\Gamma(1 + 1/\hat{\nu})$. Even though we can estimate the individual level indirect effect for each study participant, we are interested in the indirect effect at the group level that can be estimated using the average of the individual indirect effects, $\hat{IE} = \sum_i \hat{IE}_i / n$, where n is the sample size.

Similarly, we can define the total effect of the initial variable X on the outcome variable Y as

$$TE = \frac{\partial\{E(Y|X)\}}{\partial X} = \frac{\partial\{b_0 + bE(M|X) + c'X\}}{\partial X} = ab\exp(a_0 + aX)\Gamma(1 + 1/\nu) + c'.$$

The total effect at the individual level ($\hat{TE}_i$) and the group level ($\hat{TE}$) can be assessed using the estimated coefficients $\hat{a}_0$, $\hat{a}$, $\hat{b}$, $\hat{c}'$, and $\hat{\nu}$. Furthermore, based on the estimated total effect and indirect effect, we can assess the percentage of the total effect of X on Y mediated by the mediator as $\hat{PM}_i = \hat{IE}_i / \hat{TE}_i$ (individual level), and $\hat{PM} = \hat{IE} / \hat{TE}$ (group level). We calculated PM using the absolute values of IE and c' , as suggested in the literature^{37, 38}, only when both a and b paths were statistically significant.

Simulation Approach—To simulate the time-to-event mediator M , we used a simulation approach proposed by Bender et al.³⁹, which can include multiple associated covariates. Bender et al. used the inverse transform sampling method⁴⁰ to randomly generate the survival times from the distribution function of the Cox proportional model given its cumulative distribution function. The baseline hazard function is assumed to follow an exponential, Weibull or Gompertz distribution. Without loss of generality, we used the Weibull distribution for the baseline hazard function in our study, since it is one of the most popular models in survival analysis⁴¹.

Specifically, for each individual i , the cumulative distribution function is given by

$$F(t_i|x_i) = 1 - \exp(-(1/\lambda^v)t_i^v \exp(-avx_i)),$$

where $x_i \sim N(0, 0.25^2)$ is the initial variable, a is the coefficient for path a (i.e., effect of X on M), and λ and v are the scale and shape parameters for the baseline hazard Weibull function, respectively. The time-to-event variable t_i was generated using

$$t_i = (-\lambda^v \log(u) / \exp(-avx_i))^{1/v},$$

where u is randomly generated from uniform(0, 1). We generated the right censoring time c_i using the similar formula $c_i = (-\lambda^v \log(w) / \exp(-s_c))^{1/v}$, where w is randomly generated from uniform(0, 1) and s_c is a constant chosen to reflect different censoring percentages in the simulated data. The observed censored variable m_i and censoring indicator δ_i were then obtained as $m_i = \min(t_i, c_i)$ and $\delta_i = I(t_i - c_i)$, $i = 1, \dots, n$. Finally, the outcome y_i was generated using the linear regression

$$y_i = b_0 + bt_i + c'X_i + \varepsilon_y,$$

for $i = 1, \dots, n$, where b and c' are the coefficients for paths b and c' , respectively, and ε_y is the residual term following $N(0, 1)$. Note that, in practice, investigators can only observe m_i and δ_i instead of t_i .

We considered 18 simulation scenarios for different values of coefficients a and b , as well as different censoring percentages (Table 1). The scale and shape parameters for the baseline hazard Weibull function were $\lambda = 1$ and $v = 2$. The first 10 simulation scenarios had no indirect effects for the mediator because either a or b was set to zero (null hypothesis); while simulation scenarios 11 to 18 had an indirect effect of the mediator (alternative hypothesis). The censoring percentage ranged from 20% to 40%. Note that we did not include the simulation scenarios in which all the subjects were observed (i.e., 0% censoring) because in such scenarios the existing approaches such as the naïve approach and the complete-case approach work well. We employed the bias-corrected and accelerated (BCa) bootstrap⁴² and standard bootstrap to determine the confidence intervals (CIs) for the indirect effects and the percentage of total effects mediated, respectively.

Simulation Results

Table 1 reports the means and standard errors of the estimated coefficients a , b and c' , obtained on the basis of 500 replicates, each replicate with 1,000 individuals, using different approaches: the naïve approach, complete-case analysis, the Tobit mediation model and the proposed approach. For the purpose of comparison, while using the complete-case approach, the estimated value of a was computed by the AFT model with a Weibull distribution. Because the Mplus program uses the Cox proportional model, we converted the coefficient to the AFT model for ease of comparison among the estimated parameters⁴³.

When the path b coefficient was non-zero, the naïve approach tended to underestimate the path b coefficient, especially when the censoring percentage was high (e.g., 40%; scenarios 4, 6, 12, 14, 16, and 18). In addition, when the path a coefficient was also non-zero (scenarios 11–18), the naïve approach also overestimated the path c' coefficient. For example, in scenario 18, when the true values of the b and c' coefficients were 0.5 and 0.5, the estimated values obtained using the naïve approach were 0.42 and 0.65, respectively. The complete-case analysis, in contrast, underestimated the path a coefficient when the path a coefficient was non-zero. This is expected since the complete-case analysis removed the subjects with censored covariates. For example, in scenario 18, when the true value of the path a coefficient was 0.5, the estimated value was 0.31. The pattern of bias using the Tobit mediation model was different from that of the other two approaches. The Tobit model gave accurate results only when both path a and b coefficients were zero. Otherwise, it underestimated the path a or b coefficient or overestimated the path c' coefficient. For example, in scenario 18, when the true values of the path a , b , and c' coefficients were respectively 0.5, 0.5, and 0.5, the Tobit mediation model estimated the respective values as 0.21, 0.41, and 0.64. For all these methods, the bias was higher when the censoring percentage was higher. On the other hand, for all the simulation scenarios presented, the proposed strategy provided accurate estimations of the coefficients of the different paths (3 rightmost columns of Table 1). For example, for scenario 18, the means of the estimated coefficients a , b and c' were 0.51, 0.48 and 0.51, respectively, which were close to the true simulation values of 0.5, 0.5 and 0.5.

Tables 2 and 3 report the means, standard errors and 95% CIs for the group level indirect effects and percentages of the total effects mediated respectively with different approaches. The results were based on 500 replicates, and each replicate had 1,000 subjects. The group level indirect effects obtained from the existing approaches were calculated using the product of ab . As expected, the estimated group level indirect effects and percentages of the total effects mediated were more accurately estimated with the proposed approach compared to the existing approaches. For example, under scenario 18, the mean of the estimated indirect effect was 0.22 (Table 2); and the mean of the percentage of the total effect mediated was 30.78% (Table 3), using the proposed strategy; these values were close to the expected values of 0.22 and 30.87%.

For the existing approaches, when there was no indirect effect of the mediator on the association between the initial and outcome variables (i.e., the null hypothesis; scenarios 1–10), all three approaches performed similarly and provided accurate estimations that were close to 0 for both IE and PM (Tables 2 and 3). However, when the indirect effect was non-zero (scenarios 11–18), as expected, all three existing approaches provided biased estimations of IE and PM for some simulation scenarios. Specifically, the Tobit mediation approach provided highly underestimated IE s and PM s for all the simulation scenarios. For example, using the Tobit mediation approach, for scenario 18, the mean of the estimated indirect effect was 0.09 (Table 2); and the mean of the percentage of the total effect mediated was 11.98% (Table 3), which highly underestimated the expected values of 0.22 and 30.87%. The complete-case approach provided relatively accurate estimations for IE and PM when the censoring percentage was low (i.e., 20%), but biased estimations when the censoring percentage was higher (i.e., 40%). For example, for scenario 18, the mean of the

estimated indirect effect was 0.15 (Table 2); and the mean of the percentage of the total effect mediated was 23.46% (Table 3), which were biased compared to the expected values of 0.22 and 30.87%, respectively. The naïve approach performed better than the complete-case approach, which provided relatively accurate estimations for the *IE* but underestimated the *PM* when the censoring percentage was higher (i.e., 40%). For example, for scenario 18, the mean of the estimated percentage of the total effect mediated was 24.93% (Table 3), which was biased compared to the expected value of 30.87%.

In Tables 2 and 3, we also report the 95% CIs for the estimations of *IEs* and *PMs*. The Tobit mediation model provided relatively narrow CIs for all the scenarios when assessing both *IEs* and *PMs*, compared to the other three approaches. The complete-case analysis provided narrow CIs when assessing the *IEs* (Table 2), which are similar to those from the Tobit model; this approach provided wider CIs when assessing *PMs* (Table 3), which are similar to those from the naïve and proposed approaches. When assessing *IEs* (Table 2), the naïve approach and proposed approach performed similarly, with the naïve approach providing slightly wider CIs. When assessing *PMs* (Table 3), the naïve approach and proposed approach performed similarly when their effect sizes were relatively small; while the proposed approach provided wider CIs when the effect sizes increased (e.g., scenario 18).

In Table 4, we report the 95% coverage probabilities of CIs for the group level *IEs* and *PMs*. For each simulation scenario, the coverage probability was calculated as the proportion of the replicates for which the 95% BCa CIs contained the true value of *IE* or *PM*. Under the null hypothesis of no indirect effect (scenarios 1–10), the naïve, Tobit model and the proposed approach each provided 95% coverage probabilities close to the nominal coverage probability of 0.95 for the group level *IEs*. On the other hand, the complete-case approach always led to coverage probabilities less than the nominal value of 0.95, except for the simulation scenarios when both paths *a* and *b* coefficients were zero (scenarios 1 and 2). Because the *PMs* cannot be negative, the lower bound of BCa CIs was set to be zero. Thus, under the null hypothesis (scenarios 1–10), 95% coverage probabilities of all four approaches were calculated as 1.00 for *PMs*. Under the alternative hypothesis (scenarios 11–18), while estimating *IE*, the complete-case approach and Tobit model both provided the coverage probabilities that are less than the nominal value of 0.95, whereas, while estimating *PM*, these approaches provided accurate coverage probabilities in some but not for all the simulations in scenarios 11–18. The naïve and the proposed approach performed similarly while estimating *IEs*, providing coverage probabilities close to the nominal value of 0.95. However, while estimating *PMs*, when the percentages mediated and the censoring percentages were higher (scenario 18), the coverage probability of the proposed approach was 0.95 whereas coverage probability for the naïve approach was 0.84.

In Table 5, we show the type I error rates (under the null hypothesis, scenarios 1–10) and powers (under the alternative hypothesis, scenarios 11–18) for the group level *IEs* using different approaches, based on 500 replicates, where each replicate had 1,000 subjects. The complete-case analysis provided inflated type I error rates, while the other three approaches well controlled the type I error rates for all scenarios. As the complete-case analysis provided inflated type I error rates, we don't recommend this approach and only compared the powers of the other three approaches. As expected, for all three approaches, the powers

decreased as the censoring percentage increased; while the powers increased as the indirect effect increased. For all scenarios, the Tobit model always had the lowest powers among the three approaches. Compared to the naïve approach, the proposed approach always had higher (i.e., scenarios 11, 12, 15, 16 and 18) or comparable (i.e., scenarios 13, 14 and 17) powers. The power gain of the proposed approach increased as the censoring percentage increased. For example, for scenario 11 (censoring percentage [CP] = 20.1%), the powers were 81.40 and 87.10%, respectively, for the naïve and the proposed approaches, where the power gain of the proposed approach was 5.7%. For scenario 12 (CP = 40.5%), the powers were 60.20 and 74.70%, respectively, where the power gain of the proposed approach was 14.5%.

Application to the Study of Diabetes, Age at Menopause and Genetic Variants

Type-2 diabetes is a metabolic disorder resulting from a complex interplay of genetic and environmental factors^{44–49}. Over 29 million people in United States have diabetes, with type-2 diabetes accounting for ~90% to 95% of all diagnosed cases of diabetes in adults⁵⁰. Some studies have suggested that a woman's age at menopause is associated with type-2 diabetes^{21–23}, while other studies have found that association to be either weak or not significant^{51–53}. Genetic epidemiologic data have indicated that several genes are associated with a woman's age at menopause^{19, 20}. Therefore, we assessed the dual pathways for the associations between genetic variants and type-2 diabetes: (1) independently via a direct effect on diabetes, and (2) through the age at menopause. We used the data from the Multi-Ethnic Study of Atherosclerosis (MESA) cohort provided by the National Center for Biotechnology Information and downloaded from dbGaP⁵⁴. For the mediation analysis, we considered SNPs as the initial variable (X), the age at menopause as the mediator (M), and the fasting glucose level, one of the quantitative measures of type-2 diabetes, as the outcome variable (Y). The conceptual mediation model is presented in Figure 1 (B). If a woman had gone through menopause, then the age at menopause was used; if she had not gone through menopause, then her age at menopause was considered to be censored.

This analysis included 2,956 women, and the censoring rate for the age at menopause was ~14.5%. Genotypes of 47,871 SNPs from candidate genes were available for the analysis. As per the standard quality control process, we excluded SNPs with minor allele frequency < 0.01. We conducted association analyses between all SNPs and the fasting glucose level, assuming an additive genetic model. We found that 733 SNPs had p-values less than the significance level of 0.01. We then analyzed these 733 SNPs for their association with the age at menopause, also assuming an additive genetic model. Two SNPs, rs10064631 and rs8178483, were found to have p-values of 1.51×10^{-4} and 1.29×10^{-6} , respectively, for their association with the age at menopause. The p-values of these two SNPs for their association with the fasting glucose level were 7.75×10^{-4} and 1.98×10^{-4} . We included these two SNPs in the mediation analysis since they showed association with both the fasting glucose level and age at menopause. When evaluating the coefficient of path a, we investigated various different functional forms for the AFT model, such as the Weibull, Gaussian, and lognormal, and selected the best fitted distribution (Weibull) using the Akaike information criterion³³.

Similar to the simulation studies, we employed the BCa bootstrap and standard bootstrap methods to determine the CIs for the indirect effects and the percentage of total effects mediated, respectively. For each SNP, we performed bootstrapping to calculate the 95% bootstrapping CIs. For the purpose of comparison, we also performed the mediation analyses using existing approaches, including the naïve approach, the complete-case approach and the Tobit mediation model. All the analyses were adjusted for age and/or ethnicity.

Table 6 reports the estimated coefficients for different paths, indirect effects and percentages of the total effects mediated for the SNP–fasting glucose association. Based on the MESA data analyses, the association between the age at menopause and fasting glucose level was not statistically significant, implying a non-significant indirect effect, and in turn a non-significant percentage of the total effect mediated. As expected, all approaches performed similarly because there was no indirect effect. For example, for SNP rs10064631, the indirect effects were 0.0007, 0.0002, 0.0113 and 0.0768 using the naïve approach, the complete-case approach, the Tobit mediation model and the proposed approach, respectively; and the percentages of total effect mediated were given as zero for all approaches. The 95% CIs of the indirect effects were $[-0.0035, 0.0054]$, $[-0.0039, 0.0043]$, $[-0.0786, 0.1242]$ and $[-0.1132, 0.6133]$ for the four approaches, respectively, which included zero. The 95% CIs of the percentages of the total effect mediated also included zero for all approaches. Similar results were obtained for the SNP rs8178483. These findings suggest that the two SNPs, rs10064631 and rs8178483, have no indirect effect on the fasting glucose level via their effects on the age at menopause.

Discussion

In this study, we focused on a single mediator model for which the mediator is a censored time-to-event variable. Using simulation, we established the bias of the existing methods in estimating coefficients for different paths (a , b and c') when the mediator is a censored time-to-event variable. To better assess the indirect effect and percentage of the total effect mediated, we proposed an approach that provides more accurate estimations for the coefficients a , b and c' . The simulation results showed that the proposed strategy, using the AFT model and multiple imputation, can better assess the indirect effect and percentage of the total effect mediated when the mediator is a censored time-to-event variable.

We also applied the proposed and existing approaches to assess the mediating effect of the age at menopause on the association between genetic variants and the fasting glucose level using the mediation model with a censored mediator. We included two SNPs, rs10064631 and rs8178483, in the mediation analyses because they were associated with both age at menopause and fasting glucose level. Based on the MESA data, the analyses showed that there were no mediating effects of the age at menopause on the SNP–fasting glucose association, suggesting that the two SNPs, rs10064631 and rs8178483, have no indirect effect on the fasting glucose level via their effects on the age at menopause.

The measure of the indirect effect used for our proposed approach, which was adapted from that of Li et al.³⁰, can assess the indirect effect at both the group level as well as at the

individual level for each subject. Compared to the group level measures (e.g., indirect effect and percentage of the total effect mediated through the mediator) that calculate the average mediation effects of a group, the individual level measures provide information about how different subjects could have different indirect effects and percentages of the total effects mediated through the mediator. Such individual level information can be used to identify the individuals who can benefit the most from interventions that can modify the mediator's effects.

This study focused on scenarios in which the mediator is a right-censored variable. However, the proposed approach is still valid when the mediator is a left-censored variable. Specifically, the multiple imputation approach works for both right- and left-censored covariates; therefore, the paths b and c' can still be accurately estimated. Also, the conditional expectation $E(M|X)$ will be the same given the left-censored variable M when assuming the Weibull distribution for the baseline hazard function in the AFT model, so the indirect effect can be estimated using the same formula. However, if other distributions are assumed (e.g., exponential) for the AFT model, one will need to derive the formula for the indirect effect accordingly. We conducted additional simulations to investigate the performance of the proposed approach when considering other assumed distributions for the AFT model, including exponential, Gaussian and log-normal distributions (online supplementary material, Appendix 1). Specifically, when assuming exponential distribution in the AFT model, we used the same simulation approach as that described above. When assuming Gaussian and log-normal distributions, we used the AFT model to generate the time-to-event variable. Given different distributions for the AFT model, the formulas to assess IE and PM were presented. From Table S1, we observe that the proposed approach can still provide relatively accurate estimations for coefficients a , b and c' , as well as the group level IE s and PM s.

Instead of investigating the baseline distribution for the AFT model, an alternative is to use the semiparametric AFT model. For instance, we can consider using the rank-based methods⁵⁵ or a sieve maximum likelihood⁵⁶ to estimate coefficient a , and using the censored residuals by the Kaplan-Meier estimator to assess the AFT model error distribution⁵⁷. However, assessing the indirect effects in such a scenario would require new methodology.

When the models are correctly specified, the estimated coefficients, including $\hat{a}_0$, $\hat{a}$, $\hat{b}$, $\hat{c}'$, and $\hat{v}$, are consistent estimators. Specifically, $\hat{a}_0$, $\hat{a}$, and $\hat{v}$ are maximum likelihood estimators assessed using the AFT parametric model, and $\hat{b}$ and $\hat{c}'$ obtained from the multiple imputation approach have been shown to be consistent¹¹. Therefore, we deduce that the estimated indirect effects (i.e., $\hat{IE}_i$ and $\hat{IE}$), total effects, (e.g., $\hat{TE}_i$ and $\hat{TE}$) and percentages of total effects mediated (i.e., $\hat{PM}_i$ and $\hat{PM}$), which are evaluated using the estimated coefficients, are consistent estimators. The estimators of the indirect effects (i.e., product of two regression coefficients) usually have non-normal skewed distributions^{30, 58}, therefore, bootstrap approaches are preferred^{59–61}. Bootstrapping is a nonparametric method that produces confidence intervals or hypothesis tests without making assumptions about the distribution of the statistic of interest. In particular, we used the BCa bootstrap method

proposed by Efron⁴² to evaluate the confidence intervals for the indirect effects. The BCa bootstrap method can adjust for both bias and skewness in the bootstrap distribution; it has been shown to have better asymptotic properties and is often recommended for assessing the indirect effects in the mediation analysis^{61, 62}.

We also investigated the performance of the three existing approaches, including the naïve approach implemented in popular software such as Mplus, complete-case approach, and Tobit mediation model. The Tobit mediation model led to biased estimation of *IE* and *PM* when the mediating effect of the mediator on the association between the initial variable and the outcome were not zero. Furthermore, in many scenarios, the complete-case approach and Tobit model provided anti-conservative coverage probabilities for *IEs* because of the underestimation of standard errors leading to narrow CIs. Of note, in all simulation scenarios, our approach provided coverage probabilities that were close to the nominal coverage probability.

When the censoring percentage was low (i.e., 20%), both the complete-case and naïve approach performed well and obtained results for *IE* and *PM* that were similar to those obtained from the proposed approach. This is not surprising as the majority of the subjects were observed in these scenarios, which can provide sufficient information for the analyses. However, when the censoring percentage was high (i.e., 40%), our proposed approach led to better estimation of *IE* and *PM*. Very high censoring percentages (e.g., 70%) have been identified in real clinical data^{63, 64}. In such cases, the proposed approach can perform noticeably better than the existing approaches. We conducted additional simulations to investigate the performance of the different approaches in the scenario of a high censoring percentage (i.e., 60%; online supplementary material, Appendix 2). From the results reported in Tables S2 and S3, we observe that, although the proposed approach provided a slightly inflated estimation for coefficient *b* in this scenario, compared to the existing approaches, the proposed approach still provided relatively more accurate estimations for coefficients *a*, *b* and *c'*, as well as the group level *IEs* and *PMs*.

In mediation analysis, there are important assumptions for testing mediation since the mediated effect is considered to be a causal effect^{38, 65}. In this study, we assumed that the mediation model is correctly specified, including correctly specified causal orders and causal directions. We also assumed that there are no unmeasured confounders that cause variables in the model, as well as there are no measurement errors for the variables in the model. For the data analysis to assess the mediating effect of the age at menopause on the association between genetic variants and the fasting glucose level, we conceptualized the causal orders and directions on the basis of previous studies^{19–23, 44–49}. We further assumed that the measurements for the variables of interest were perfect and there were no confounding factors mitigating associations among the genetic variants, age at menopause, and fasting glucose level.

Although the multiple imputation approach provided better estimations for coefficients *b* and *c'* in our simulation studies, we note certain limitations to this approach. First, it assumes a linear relationship between the mediator and the initial variable and might provide biased estimations when such an assumption is invalid. Therefore, we recommend performing

sensitivity analyses to check the model assumptions. Second, the multiple imputation approach was developed for outcome variables that are continuous. The analytics needs to be further developed for binary or censored outcomes. Finally, a single censored mediator is considered in this approach; further work is needed to extend this approach to include multiple censored mediators.

Several other approaches have been proposed for assessing the censored covariates in regression models^{25–27, 66–70}. These approaches might be potential candidates to address the problems that arise from censored mediator variables. However, the major limitation that is common to all the approaches is that they do not account for a causal relationship from the non-censored variable (e.g., initial variable) to the censored variable (e.g., mediator), which is usually implied in a mediation model^{71, 72}. Therefore, using these approaches in mediation model analysis with a censored mediator is not straightforward and needs further investigation.

In summary, we proposed an approach to accurately estimate the coefficients for the different paths in a mediation model when the mediator is a censored time-to-event variable. We also adapted a measure of the indirect effect, which is more appropriate when assessing the indirect effect and percentage of the total effect mediated at both the group level and the individual level for each subject when the mediator is a censored variable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at <http://www.mesa-nhlbi.org>.

Funding: Jian Wang was supported in part by the National Institutes of Health (NIH), grant R03 CA192197. Sanjay Shete was supported in part by the NIH grants 1R01CA131324, R01DE022891, and R25DA026120; the Cancer Prevention Research Institute of Texas grant RP130123, and the Barnhart Family Distinguished Professorship in Targeted Therapy. Jian Wang and Sanjay Shete were supported in part by the Cancer Center Support Grant P30CA016672. The MESA project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC- 95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169 and CTSA UL1-RR-024156. The MESA CARE data used for the analyses described in this manuscript were obtained through dbGaP (phs000209.v13.p3). Funding for CARE genotyping was provided by NHLBI Contract N01-HC-65226.

References

1. MacKinnon DP. Introduction to statistical mediation analysis. New York: Erlbaum; 2008. 488
2. MacKinnon DP, Fritz MS, Williams J, Lockwood CM. Distribution of the product confidence limits for the indirect effect: program PRODCLIN. Behav Res Methods. 2007; 39:384–389. [PubMed: 17958149]
3. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986; 51:1173–1182. [PubMed: 3806354]
4. Taylor AB, MacKinnon DP. Four applications of permutation methods to testing a single-mediator model. Behav Res Methods. 2012; 44:806–844. [PubMed: 22311738]

5. Selig JP, Preacher KJ, Little TD. Modeling Time-Dependent Association in Longitudinal Data: A Lag as Moderator Approach. *Multivar Behav Res.* 2012; 47:697–716.
6. Shrout PE, Bolger N. Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol Methods.* 2002; 7:422–445. [PubMed: 12530702]
7. Wang J, Spitz MR, Amos CI, et al. Method for evaluating multiple mediators: mediating effects of smoking and COPD on the association between the CHRNA5-A3 variant and lung cancer risk. *PLoS One.* 2012; 7:e47705. [PubMed: 23077662]
8. Wang J, Spitz MR, Amos CI, Wilkinson AV, Wu X, Shete S. Mediating effects of smoking and chronic obstructive pulmonary disease on the relation between the CHRNA5-A3 genetic locus and lung cancer risk. *Cancer.* 2010; 116:3458–3462. [PubMed: 20564069]
9. Wang LJ, Zhang ZY. Estimating and Testing Mediation Effects with Censored Data. *Struct Equ Modeling.* 2011; 18:18–34.
10. Lange T, Hansen JV. Direct and indirect effects in a survival context. *Epidemiology.* 2011; 22:575–581. [PubMed: 21552129]
11. Wang HJ, Feng XD. Multiple Imputation for M-Regression With Censored Covariates. *J Am Stat Assoc.* 2012; 107:194–204.
12. VanderWeele TJ. Causal mediation analysis with survival data. *Epidemiology.* 2011; 22:582–585. [PubMed: 21642779]
13. Tein JY, MacKinnon DP. Estimating mediated effects with survival data. *New Developments in Psychometrics.* 2003:405–412.
14. Tchetgen Tchetgen EJ. On causal mediation analysis with a survival outcome. *Int J Biostat.* 2011; 7 Article 33.
15. Luo P, Geng Z. Causal mediation analysis for survival outcome with unobserved mediator-outcome confounders. *Comput Stat Data An.* 2015
16. Wakai K, Seki N, Tamakoshi A, et al. Decrease in risk of lung cancer death in males after smoking cessation by age at quitting: findings from the JACC study. *JPN J Cancer Res.* 2001; 92:821–828. [PubMed: 11509112]
17. Huxley RR, Yatsuya H, Lutsey PL, Woodward M, Alonso A, Folsom AR. Impact of age at smoking initiation, dosage, and time since quitting on cardiovascular disease in african americans and whites: the atherosclerosis risk in communities study. *Am J of Epidemiol.* 2012; 175:816–826. [PubMed: 22396389]
18. Planas A, Clara A, Marrugat J, et al. Age at onset of smoking is an independent risk factor in peripheral artery disease development. *J Vasc Surg.* 2002; 35:506–509. [PubMed: 11877699]
19. Laven JS. Genetics of Early and Normal Menopause. *Semin Reprod Med.* 2015; 33:377–383. [PubMed: 26569518]
20. Day FR, Ruth KS, Thompson DJ, et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat Genet.* 2015; 47:1294–1303. [PubMed: 26414677]
21. Brand JS, van der Schouw YT, Onland-Moret NC, et al. Age at menopause, reproductive life span, and type 2 diabetes risk: results from the EPIC-InterAct study. *Diabetes Care.* 2013; 36:1012–1019. [PubMed: 23230098]
22. LeBlanc ES, Kapphahn K, Hedlin H, et al. Reproductive history and risk of type 2 diabetes mellitus in postmenopausal women: findings from the Women's Health Initiative. *Menopause.* 2016
23. Malacara JM, Huerta R, Rivera B, Esparza S, Fajardo ME. Menopause in normal and uncomplicated NIDDM women: physical and emotional symptoms and hormone profile. *Maturitas.* 1997; 28:35–45. [PubMed: 9391993]
24. Muthé LK, Muthén BO. *Mplus User's Guide.* 6th. Los Angeles, CA: Muthén & Muthén; 2011.
25. May RC, Ibrahim JG, Chu H. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Stat Med.* 2011; 30:2551–2561. [PubMed: 21710558]
26. Tsimikas JV, Bantis LE, Georgiou SD. Inference in generalized linear regression models with a censored covariate. *Comput Stat Data An.* 2012; 56:1854–1868.

27. Bernhardt PW, Wang HJ, Zhang DW. Statistical methods for generalized linear models with covariates subject to detection limits. *Stat Biosci.* 2013;1–22.
28. Demissie S, LaValley MP, Horton NJ, Glynn RJ, Cupples LA. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Stat Med.* 2003; 22:545–557. [PubMed: 12590413]
29. Mackinnon DP, Warsi G, Dwyer JH. A Simulation Study of Mediated Effect Measures. *Multivariate Behav Res.* 1995; 30:41. [PubMed: 20157641]
30. Li Y, Schneider JA, Bennett DA. Estimation of the mediation effect with a binary mediator. *Stat Med.* 2007; 26:3398–3414. [PubMed: 17066450]
31. Swindell WR. Accelerated failure time models provide a useful statistical framework for aging research. *Exp Gerontol.* 2009; 44:190–200. [PubMed: 19007875]
32. Collett D. *Modeling Survival Data in Medical Research.* Boca Raton: CRC Press; 2003.
33. Akaike A. A new look at the statistical model identification. *IEEE T Automat Contr.* 1974; 19:716–723.
34. Therneau TA. *Package for Survival Analysis in S.* 2014
35. Portnoy S. Censored regression quantiles. *J Am Stat Assoc.* 2003; 98:1001–1012.
36. Powell JL. Censored Regression Quantiles. *J Econometrics.* 1986; 32:143–155.
37. Alwin DF, Hauser RM. Decomposition of Effects in Path Analysis. *American Sociological Review.* 1975; 40:37–47.
38. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annual Review of Psychology.* 2007; 58:593–614.
39. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med.* 2005; 24:1713–1723. [PubMed: 15724232]
40. Devroye L. *Non-Uniform Random Variate Generation.* New York: Springer-Verlag; 1986.
41. Liu X. *Survival Analysis: Models and Applications.* United Kingdom: John Wiley & Sons; 2012.
42. Efron B. Better Bootstrap Confidence-Intervals. *Journal of the American Statistical Association.* 1987; 82:171–185.
43. Klein JP, Moeschberger M. *Survival Analysis: Techniques for Censored and Truncated Data.* 2nd. New York: Springer; 2003.
44. Fuchsberger C, Flannick J, Teslovich TM, et al. The genetic architecture of type 2 diabetes. *Nature.* 2016
45. Willemsen G, Ward KJ, Bell CG, et al. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Res Hum Genet.* 2015; 18:762–771. [PubMed: 26678054]
46. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet.* 2012; 44:981–990. [PubMed: 22885922]
47. Almind K, Doria A, Kahn CR. Putting the genes for type II diabetes on the map. *Nat Med.* 2001; 7:277–279. [PubMed: 11231616]
48. Cornelis MC, Hu FB. Gene-environment interactions in the development of type 2 diabetes: recent progress and continuing challenges. *Annu Rev Nutr.* 2012; 32:245–259. [PubMed: 22540253]
49. Lyssenko V, Laakso M. Genetic screening for the risk of type 2 diabetes: worthless or valuable? *Diabetes Care.* 2013; 36(Suppl 2):S120–126. [PubMed: 23882036]
50. Centers for Disease Control and Prevention. *National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States.* Atlanta, GA: U.S. Department of Health and Human Services; 2014.
51. Lopez-Lopez R, Huerta R, Malacara JM. Age at menopause in women with type 2 diabetes mellitus. *Menopause.* 1999; 6:174–178. [PubMed: 10374226]
52. Heianza Y, Arase Y, Kodama S, et al. Effect of postmenopausal status and age at menopause on type 2 diabetes and prediabetes in Japanese individuals: Toranomon Hospital Health Management Center Study 17 (TOPICS 17). *Diabetes Care.* 2013; 36:4007–4014. [PubMed: 24170752]
53. Szmulowicz ED, Stuenkel CA, Seely EW. Influence of menopause on diabetes and diabetes risk. *Nat Rev Endocrinol.* 2009; 5:553–558. [PubMed: 19687788]

54. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet.* 2007; 39:1181–1186. [PubMed: 17898773]
55. Wei LJ, Ying Z, Lin DY. Linear-Regression Analysis of Censored Survival-Data Based on Rank-Tests. *Biometrika.* 1990; 77:845–851.
56. Ding Y, Nan B. A Sieve M-Theorem for Bundled Parameters in Semiparametric Models, with Application to the Efficient Estimation in a Linear Model for Censored Data. *Annals of Statistics.* 2011; 39:3032–3061.
57. Kong SC, Nan B. Semiparametric approach to regression with a covariate subject to a detection limit. *Biometrika.* 2016; 103:161–174.
58. Mackinnon DP, Lockwood CM, Williams J. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behav Res.* 2004; 39:99. [PubMed: 20157642]
59. Preacher KJ, Kelley K. Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychol Methods.* 2011; 16:93–115. [PubMed: 21500915]
60. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav Res Methods Instrum Comput.* 2004; 36:717–731. [PubMed: 15641418]
61. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis.
62. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods.* 2008; 40:879–891. [PubMed: 18697684]
63. Wu T, Sun W, Yuan S, Chen CH, Li KC. A method for analyzing censored survival phenotype with gene expression data. *BMC Bioinformatics.* 2008; 9:417. [PubMed: 18837994]
64. Wan X, Peng L, Li Y. A review and comparison of methods for recreating individual patient data from published Kaplan-Meier survival curves for economic evaluations: a simulation study. *PLoS One.* 2015; 10:e0121353. [PubMed: 25803659]
65. Kenny DA. *Mediation.* 2016
66. Bernhardt PW, Zhang D, Wang HJ. A Fast EM Algorithm for Fitting Joint Models of a Binary Response and Multiple Longitudinal Covariates Subject to Detection Limits. *Comput Stat Data An.* 2015; 85:37–53.
67. Bernhardt PW, Wang HJ, Zhang D. Flexible Modeling of Survival Data with Covariates Subject to Detection Limits via Multiple Imputation. *Comput Stat Data An.* 2014; 69
68. Lee M, Kong L, Weissfeld L. Multiple imputation for left-censored biomarker data based on Gibbs sampling method. *Stat Med.* 2012; 31:1838–1848. [PubMed: 22359320]
69. Lynn HS. Maximum likelihood inference for left-censored HIV RNA data. *Stat Med.* 2001; 20:33–45. [PubMed: 11135346]
70. Lyles RH, Fan D, Chuachoowong R. Correlation coefficient estimation involving a left censored laboratory assay variable. *Stat Med.* 2001; 20:2921–2933. [PubMed: 11568949]
71. Imai K, Keele L, Tingley D. A General Approach to Causal Mediation Analysis. *Psychol Methods.* 2010; 15:309–334. [PubMed: 20954780]
72. Imai K, Keele L, Yamamoto T. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Stat Sci.* 2010; 25:51–71.

(A)

(B)

Figure 1.

(A) Diagram for a single mediator model. (B) The mediation model for the study of the mediating effect of age at menopause on the association between genetic variants and type-2 diabetes.

Table 1

Means and standard errors (se) of estimated coefficients for different paths, a , b and c' , obtained from 500 replicates, each with 1,000 individuals, using different approaches.

Scenario	a	b	c'	CP	Native			Complete-Case			Tobit			Proposed Approach		
					a (se)	b (se)	c' (se)	a (se)	b (se)	c' (se)	a (se)	b (se)	c' (se)	a (se)	b (se)	c' (se)
1	0	0	0.5	20.0%	0.00 (0.07)	-0.01 (0.08)	0.50 (0.13)	0.00 (0.07)	-0.01 (0.08)	0.49 (0.14)	0.00 (0.05)	-0.01 (0.08)	0.49 (0.13)	0.00 (0.07)	-0.01 (0.07)	0.50 (0.13)
2	0	0	0.5	40.2%	0.00 (0.08)	0.00 (0.09)	0.50 (0.13)	0.00 (0.08)	0.00 (0.11)	0.50 (0.17)	0.00 (0.05)	0.00 (0.09)	0.50 (0.13)	0.00 (0.08)	0.00 (0.09)	0.50 (0.13)
3	0	0.25	0.5	19.9%	0.00 (0.07)	0.23 (0.08)	0.51 (0.13)	0.00 (0.07)	0.25 (0.09)	0.51 (0.15)	0.00 (0.05)	0.23 (0.08)	0.51 (0.13)	0.00 (0.07)	0.23 (0.07)	0.51 (0.13)
4	0	0.25	0.5	40.2%	0.00 (0.08)	0.22 (0.09)	0.50 (0.12)	0.00 (0.08)	0.25 (0.12)	0.51 (0.17)	0.00 (0.04)	0.21 (0.09)	0.50 (0.12)	0.00 (0.08)	0.25 (0.09)	0.50 (0.12)
5	0	0.5	0.5	19.8%	0.00 (0.07)	0.47 (0.07)	0.49 (0.13)	0.00 (0.07)	0.50 (0.08)	0.49 (0.15)	0.00 (0.05)	0.47 (0.07)	0.49 (0.13)	0.00 (0.07)	0.46 (0.07)	0.49 (0.13)
6	0	0.5	0.5	40.3%	0.00 (0.08)	0.43 (0.09)	0.50 (0.14)	0.00 (0.08)	0.50 (0.11)	0.50 (0.17)	0.00 (0.05)	0.42 (0.09)	0.49 (0.14)	0.00 (0.08)	0.50 (0.09)	0.50 (0.14)
7	0.25	0	0.5	20.1%	0.25 (0.07)	0.00 (0.08)	0.50 (0.12)	0.20 (0.07)	0.00 (0.09)	0.50 (0.14)	0.16 (0.05)	0.00 (0.08)	0.50 (0.12)	0.25 (0.07)	0.00 (0.07)	0.50 (0.12)
8	0.25	0	0.5	40.3%	0.24 (0.08)	0.00 (0.09)	0.50 (0.13)	0.14 (0.08)	-0.01 (0.12)	0.50 (0.17)	0.10 (0.05)	-0.01 (0.09)	0.49 (0.13)	0.24 (0.08)	0.00 (0.08)	0.50 (0.13)
9	0.5	0	0.5	20.2%	0.50 (0.07)	-0.01 (0.08)	0.50 (0.13)	0.40 (0.07)	-0.01 (0.09)	0.49 (0.15)	0.32 (0.06)	-0.02 (0.08)	0.49 (0.13)	0.50 (0.08)	-0.01 (0.07)	0.50 (0.13)
10	0.5	0	0.5	40.6%	0.50 (0.09)	0.00 (0.09)	0.50 (0.12)	0.30 (0.08)	0.00 (0.12)	0.50 (0.16)	0.20 (0.05)	0.00 (0.09)	0.50 (0.12)	0.50 (0.09)	0.00 (0.09)	0.50 (0.13)
11	0.25	0.25	0.5	20.1%	0.25 (0.07)	0.23 (0.08)	0.52 (0.14)	0.20 (0.07)	0.25 (0.08)	0.50 (0.15)	0.16 (0.05)	0.22 (0.08)	0.51 (0.14)	0.25 (0.07)	0.22 (0.07)	0.50 (0.14)
12	0.25	0.25	0.5	40.5%	0.25 (0.08)	0.22 (0.09)	0.53 (0.13)	0.15 (0.08)	0.25 (0.11)	0.49 (0.17)	0.10 (0.04)	0.21 (0.09)	0.53 (0.13)	0.25 (0.08)	0.26 (0.08)	0.49 (0.13)
13	0.25	0.5	0.5	20.0%	0.25 (0.07)	0.47 (0.08)	0.54 (0.13)	0.20 (0.07)	0.51 (0.08)	0.50 (0.14)	0.16 (0.05)	0.46 (0.08)	0.54 (0.13)	0.25 (0.07)	0.46 (0.07)	0.50 (0.13)
14	0.25	0.5	0.5	40.5%	0.25 (0.08)	0.43 (0.09)	0.56 (0.13)	0.15 (0.08)	0.50 (0.12)	0.48 (0.16)	0.10 (0.04)	0.42 (0.09)	0.56 (0.13)	0.25 (0.08)	0.50 (0.08)	0.49 (0.14)
15	0.5	0.25	0.5	20.4%	0.50 (0.07)	0.24 (0.07)	0.54 (0.13)	0.40 (0.07)	0.25 (0.08)	0.50 (0.14)	0.32 (0.05)	0.23 (0.07)	0.54 (0.13)	0.50 (0.07)	0.23 (0.07)	0.50 (0.13)
16	0.5	0.25	0.5	40.5%	0.50 (0.08)	0.21 (0.09)	0.57 (0.12)	0.30 (0.09)	0.25 (0.12)	0.50 (0.16)	0.20 (0.05)	0.21 (0.09)	0.57 (0.12)	0.49 (0.08)	0.25 (0.09)	0.50 (0.13)
17	0.5	0.5	0.5	20.2%	0.50 (0.07)	0.47 (0.08)	0.58 (0.12)	0.39 (0.07)	0.50 (0.09)	0.51 (0.14)	0.31 (0.05)	0.46 (0.08)	0.58 (0.12)	0.50 (0.07)	0.45 (0.07)	0.51 (0.13)
18	0.5	0.5	0.5	40.6%	0.51 (0.08)	0.42 (0.09)	0.65 (0.13)	0.31 (0.08)	0.49 (0.11)	0.51 (0.17)	0.21 (0.05)	0.41 (0.09)	0.64 (0.13)	0.51 (0.08)	0.48 (0.09)	0.51 (0.14)

CP: censoring percentage

Table 2

Means and standard errors (se), as well as 95% confidence intervals (CI) of group level indirect effects, obtained from 500 replicates, each with 1,000 individuals, using different approaches.

Scenario	Expected	Naïve			Complete-Case			Tobit			Proposed Approach		
		Mean [CI]	se	Mean [CI]	Mean [CI]	se	Mean [CI]	Mean [CI]	se	Mean [CI]	Mean [CI]	se	Mean [CI]
1	0.00	0.00 [-0.02, 0.02]	0.005	0.00 [-0.01, 0.02]	0.006	0.004	0.00 [-0.01, 0.01]	0.00 [-0.02, 0.02]	0.004	0.00 [-0.02, 0.02]	0.00 [-0.02, 0.02]	0.004	0.00 [-0.02, 0.02]
2	0.00	0.00 [-0.02, 0.03]	0.007	0.00 [-0.02, 0.03]	0.009	0.004	0.00 [-0.01, 0.01]	0.00 [-0.02, 0.03]	0.004	0.00 [-0.02, 0.03]	0.00 [-0.02, 0.03]	0.007	0.00 [-0.02, 0.03]
3	0.00	0.00 [-0.04, 0.04]	0.020	0.00 [-0.03, 0.04]	0.019	0.013	0.00 [-0.03, 0.03]	0.00 [-0.03, 0.03]	0.013	0.00 [-0.03, 0.03]	0.00 [-0.03, 0.03]	0.016	0.00 [-0.03, 0.03]
4	0.00	0.00 [-0.04, 0.04]	0.023	0.00 [-0.02, 0.04]	0.018	0.010	0.00 [-0.02, 0.02]	0.00 [-0.04, 0.04]	0.010	0.00 [-0.04, 0.04]	0.00 [-0.04, 0.04]	0.019	0.00 [-0.04, 0.04]
5	0.00	0.00 [-0.07, 0.07]	0.040	0.00 [-0.05, 0.06]	0.035	0.025	0.00 [-0.05, 0.05]	0.00 [-0.06, 0.06]	0.025	0.00 [-0.06, 0.06]	0.00 [-0.06, 0.06]	0.032	0.00 [-0.06, 0.06]
6	0.00	0.00 [-0.07, 0.07]	0.044	0.00 [-0.05, 0.06]	0.037	0.018	0.00 [-0.04, 0.04]	0.00 [-0.08, 0.08]	0.018	0.00 [-0.08, 0.08]	0.00 [-0.08, 0.08]	0.040	0.00 [-0.08, 0.08]
7	0.00	0.00 [-0.04, 0.04]	0.034	0.00 [-0.02, 0.04]	0.035	0.024	0.00 [-0.03, 0.03]	0.00 [-0.03, 0.04]	0.024	0.00 [-0.03, 0.04]	0.00 [-0.03, 0.04]	0.029	0.00 [-0.03, 0.04]
8	0.00	0.00 [-0.05, 0.05]	0.036	0.00 [-0.02, 0.03]	0.042	0.019	0.00 [-0.02, 0.02]	0.00 [-0.05, 0.05]	0.019	0.00 [-0.05, 0.05]	0.00 [-0.05, 0.05]	0.037	0.00 [-0.05, 0.05]
9	0.00	0.00 [-0.08, 0.07]	0.017	0.00 [-0.05, 0.05]	0.018	0.013	0.00 [-0.05, 0.04]	0.00 [-0.07, 0.06]	0.013	0.00 [-0.07, 0.06]	0.00 [-0.07, 0.06]	0.015	0.00 [-0.07, 0.06]
10	0.00	0.00 [-0.09, 0.09]	0.020	0.00 [-0.03, 0.05]	0.023	0.011	0.00 [-0.04, 0.04]	0.00 [-0.08, 0.08]	0.011	0.00 [-0.08, 0.08]	0.00 [-0.08, 0.08]	0.020	0.00 [-0.08, 0.08]
11	0.06	0.06 [0.01, 0.12]	0.050	0.05 [0.02, 0.12]	0.050	0.034	0.04 [0.01, 0.08]	0.05 [0.01, 0.10]	0.034	0.05 [0.01, 0.10]	0.05 [0.01, 0.10]	0.041	0.05 [0.01, 0.10]
12	0.06	0.06 [0.00, 0.13]	0.058	0.04 [0.01, 0.11]	0.055	0.027	0.02 [0.00, 0.05]	0.06 [0.01, 0.13]	0.027	0.06 [0.01, 0.13]	0.06 [0.01, 0.13]	0.054	0.06 [0.01, 0.13]
13	0.11	0.12 [0.05, 0.20]	0.039	0.10 [0.06, 0.20]	0.040	0.028	0.07 [0.02, 0.13]	0.10 [0.04, 0.18]	0.028	0.10 [0.04, 0.18]	0.10 [0.04, 0.18]	0.033	0.10 [0.04, 0.18]
14	0.11	0.11 [0.03, 0.20]	0.042	0.08 [0.04, 0.17]	0.045	0.022	0.04 [0.01, 0.09]	0.11 [0.04, 0.21]	0.022	0.11 [0.04, 0.21]	0.11 [0.04, 0.21]	0.041	0.11 [0.04, 0.21]
15	0.11	0.12 [0.04, 0.21]	0.040	0.10 [0.07, 0.19]	0.038	0.026	0.07 [0.02, 0.13]	0.10 [0.04, 0.18]	0.026	0.10 [0.04, 0.18]	0.10 [0.04, 0.18]	0.033	0.10 [0.04, 0.18]
16	0.11	0.11 [0.02, 0.21]	0.048	0.07 [0.04, 0.16]	0.042	0.020	0.04 [0.00, 0.09]	0.11 [0.03, 0.21]	0.020	0.11 [0.03, 0.21]	0.11 [0.03, 0.21]	0.043	0.11 [0.03, 0.21]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Scenario	Expected	Naïve			Complete-Case			Tobit			Proposed Approach		
		Mean [CI]	se		Mean [CI]	se		Mean [CI]	se		Mean [CI]	se	
17	0.22	0.23 [0.14, 0.34]	0.026		0.20 [0.16, 0.30]	0.024		0.14 [0.08, 0.22]	0.018		0.20 [0.13, 0.29]	0.020	
18	0.22	0.21 [0.11, 0.34]	0.031		0.15 [0.12, 0.25]	0.028		0.09 [0.04, 0.14]	0.014		0.22 [0.12, 0.33]	0.029	

Table 3

Means and standard errors (se), as well as 95% confidence intervals (CI) of group level percentages of total effects mediated, obtained from 500 replicates, each with 1,000 individuals, using different approaches.

Scenario	Expected	Naïve		Complete-Case			Tobit			Proposed Approach		
		Mean [CI]	se	Mean [CI]	se	Mean [CI]	se	Mean [CI]	se	Mean [CI]	se	
1	0.00%	0.00% [0.00%, 2.20%]	0.000	0.01% [0.00%, 2.21%]	0.002	0.01% [0.00%, 1.50%]	0.003	0.00% [0.00%, 1.45%]	0.002			
2	0.00%	0.00% [0.00%, 2.70%]	0.000	0.00% [0.00%, 2.95%]	0.000	0.01% [0.00%, 1.38%]	0.001	0.42% [0.00%, 2.69%]	0.004			
3	0.00%	0.25% [0.00%, 10.35%]	0.022	0.36% [0.00%, 9.90%]	0.018	0.22% [0.00%, 7.66%]	0.013	1.19% [0.00%, 9.37%]	0.017			
4	0.00%	0.31% [0.00%, 11.82%]	0.020	0.41% [0.00%, 10.62%]	0.013	0.15% [0.00%, 6.29%]	0.007	0.31% [0.00%, 13.28%]	0.017			
5	0.00%	0.58% [0.02%, 20.19%]	0.036	0.66% [0.00%, 19.23%]	0.032	0.44% [0.00%, 15.62%]	0.025	2.47% [0.02%, 18.24%]	0.027			
6	0.00%	0.80% [0.00%, 20.40%]	0.038	0.87% [0.00%, 20.06%]	0.043	0.40% [0.00%, 12.21%]	0.017	0.88% [0.00%, 22.19%]	0.042			
7	0.00%	0.50% [0.00%, 12.89%]	0.030	0.35% [0.00%, 9.93%]	0.031	0.24% [0.00%, 8.39%]	0.025	0.32% [0.00%, 10.21%]	0.026			
8	0.00%	0.40% [0.00%, 13.89%]	0.038	0.15% [0.00%, 7.83%]	0.047	0.12% [0.00%, 5.67%]	0.018	1.57% [0.00%, 12.63%]	0.042			
9	0.00%	0.88% [0.00%, 21.91%]	0.015	0.76% [0.00%, 18.24%]	0.018	0.60% [0.00%, 15.06%]	0.012	0.47% [0.00%, 18.64%]	0.015			
10	0.00%	0.82% [0.00%, 24.07%]	0.018	0.79% [0.00%, 17.10%]	0.021	0.33% [0.00%, 11.56%]	0.010	0.73% [0.00%, 23.64%]	0.018			
11	9.97%	9.57% [0.51%, 26.12%]	0.069	8.00% [0.04%, 23.65%]	0.087	5.93% [0.13%, 19.17%]	0.059	9.26% [0.37%, 26.25%]	0.078			
12	9.97%	7.54% [0.18%, 24.31%]	0.068	3.57% [0.00%, 20.82%]	0.116	2.50% [0.03%, 12.57%]	0.043	8.97% [0.29%, 32.03%]	0.090			
13	18.17%	17.91% [3.80%, 35.57%]	0.072	16.18% [1.10%, 35.88%]	0.098	11.52% [1.49%, 27.45%]	0.066	17.18% [3.72%, 37.72%]	0.072			
14	18.17%	15.15% [1.52%, 32.83%]	0.082	9.08% [0.09%, 32.75%]	0.117	5.67% [0.16%, 18.92%]	0.055	18.10% [2.26%, 44.35%]	0.103			
15	18.14%	17.34% [2.69%, 37.82%]	0.092	16.20% [1.24%, 36.69%]	0.100	11.60% [1.37%, 28.71%]	0.069	17.26% [2.65%, 40.18%]	0.092			
16	18.14%	12.96% [0.95%, 34.02%]	0.102	10.56% [0.10%, 32.64%]	0.118	5.55% [0.18%, 18.39%]	0.052	17.95% [1.72%, 45.45%]	0.118			

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Scenario	Expected	Naïve			Complete-Case			Tobit			Proposed Approach		
		Mean [CI]	se		Mean [CI]	se		Mean [CI]	se		Mean [CI]	se	
17	30.87%	29.06% [17.09%, 46.96%]	0.066		28.78% [11.30%, 48.34%]	0.073		20.53% [10.77%, 36.92%]	0.052		29.10% [17.17%, 51.88%]	0.061	
18	30.87%	24.93% [11.08%, 40.92%]	0.072		23.46% [1.95%, 45.78%]	0.074		11.98% [3.22%, 23.50%]	0.038		30.78% [16.80%, 57.48%]	0.081	

95% coverage probabilities of confidence intervals (CIs) for group level indirect effects (*IE*) and percentages of total effects mediated (*PM*), based on 500 replicates, each with 1,000 individuals, using different approaches.

Table 4

Scenario	Naïve		Complete-Case		Tobit		Proposed Approach	
	<i>IE</i>	<i>PM</i> *	<i>IE</i>	<i>PM</i>	<i>IE</i>	<i>PM</i>	<i>IE</i>	<i>PM</i>
1	1.00	–	0.95	–	1.00	–	1.00	–
2	1.00	–	0.90	–	1.00	–	0.98	–
3	0.96	–	0.84	–	0.96	–	0.95	–
4	0.96	–	0.78	–	0.97	–	0.93	–
5	0.95	–	0.84	–	0.95	–	0.96	–
6	0.95	–	0.78	–	0.94	–	0.95	–
7	0.95	–	0.80	–	0.95	–	0.94	–
8	0.96	–	0.77	–	0.97	–	0.94	–
9	0.94	–	0.82	–	0.93	–	0.93	–
10	0.95	–	0.73	–	0.95	–	0.94	–
11	0.94	0.96	0.87	0.89	0.28	0.98	0.96	0.96
12	0.94	0.95	0.83	0.75	0.60	0.99	0.96	0.97
13	0.94	0.94	0.82	0.89	0.74	0.79	0.94	0.94
14	0.95	0.93	0.77	0.76	0.23	0.41	0.96	0.97
15	0.96	0.96	0.86	0.90	0.15	0.83	0.95	0.95
16	0.95	0.93	0.78	0.70	0.34	0.97	0.95	0.95
17	0.94	0.94	0.78	0.88	0.44	0.65	0.91	0.94
18	0.96	0.84	0.55	0.70	0.03	0.17	0.96	0.95

* The *PM* cannot be negative so the lower bound of CI was set to be 0. Therefore, the estimated coverage probabilities are 1.00 for the scenarios 1–10 (under null hypothesis).

Table 5

Type I error rates and power comparisons for group level indirect effects, based on 500 replicates, each with 1,000 individuals, using different approaches, at 0.05 significance level.

	Scenario	a	b	Expected	Naïve	Complete-Case*	Tobit	Proposed Approach
Null hypothesis								
Type I error	1	0	0	0.00	0.00	0.05	0.00	0.00
	2	0	0	0.00	0.00	0.10	0.00	0.02
	3	0	0.25	0.00	0.04	0.16	0.04	0.05
	4	0	0.25	0.00	0.04	0.22	0.03	0.07
	5	0	0.5	0.00	0.05	0.16	0.05	0.04
	6	0	0.5	0.00	0.05	0.23	0.06	0.05
	7	0.25	0	0.00	0.05	0.20	0.05	0.06
	8	0.25	0	0.00	0.04	0.23	0.03	0.06
	9	0.5	0	0.00	0.06	0.18	0.07	0.07
	10	0.5	0	0.00	0.05	0.27	0.05	0.06
Alternative hypothesis								
Power	11	0.25	0.25	0.06	81.40%	–	71.80%	87.40%
	12	0.25	0.25	0.06	60.20%	–	39.80%	74.70%
	13	0.25	0.5	0.11	95.60%	–	88.00%	94.80%
	14	0.25	0.5	0.11	85.40%	–	62.20%	84.04%
	15	0.5	0.25	0.11	87.00%	–	85.40%	91.38%
	16	0.5	0.25	0.11	68.40%	–	65.80%	78.46%
	17	0.5	0.5	0.22	100.00%	–	99.80%	100.00%
	18	0.5	0.5	0.22	99.60%	–	99.00%	100.00%

*. The power of the complete-case analysis is not reported since it has inflated type I error rate.

Table 6

Estimations of the coefficients for paths a , b and c' as well as the group level indirect effects (IE) and percentages of total effects mediated (PM) for the SNPs associated with both the fasting glucose level and age at menopause in the real data analysis. The 95% confidence intervals (CIs) were assessed using a bootstrap approach with 500 bootstraps.

SNP	CHR	Method	Coefficient			IE	IE 95% CI	PM	PM 95% CI
			a	b	c'				
rs10064631	5	Naïve	-0.0230	-0.0300	-6.5830	0.0007	[-0.0035, 0.0054]	0.00	[00.00%, 00.10%]
		Complete-Case	-0.0157	-0.0099	-5.7470	0.0002	[-0.0039, 0.0043]	0.00	[00.00%, 00.07%]
		Tobit	-0.2288	-0.0496	-6.1304	0.0113	[-0.0786, 0.1242]	0.00	[00.00%, 00.00%]
		Proposed Approach	-0.0345	-0.0462	-6.5269	0.0768	[-0.1132, 0.6133]	0.00	[00.00%, 05.98%]
rs8178483	21	Naïve	0.0098	-0.0260	10.6450	-0.0003	[-0.0046, 0.0030]	0.00	[00.00%, 00.00%]
		Complete-Case	0.0039	-0.0078	5.9765	0.0000	[-0.0042, 0.0035]	0.00	[00.00%, 00.00%]
		Tobit	0.0450	-0.0433	10.6379	-0.0019	[-0.2296, 0.1592]	0.00	[00.00%, 00.00%]
		Proposed Approach	0.0720	-0.0281	10.9353	-0.0980	[-1.6265, 0.2943]	0.00	[00.00%, 09.89%]

CHR: chromosome