Published in final edited form as:

Stat Med. 2019 March 30; 38(7): 1213–1229. doi:10.1002/sim.8028.

Mediation Analysis in a Case–control Study When the Mediator is a Censored Variable

Jian Wang¹, Jing Ning¹, 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

Mediation analysis is an approach for assessing the direct and indirect effects of an initial variable on an outcome through a mediator. In practice, mediation models can involve a censored mediator (e.g., a woman's age at menopause). The current research for mediation analysis with a censored mediator focuses on scenarios where outcomes are continuous. However, the outcomes can be binary (e.g., type-2 diabetes). Another challenge when analyzing such a mediation model is to use data from a case–control study, which results in biased estimations for the initial variable–mediator association if a standard approach is directly applied. In this study, we propose an approach (denoted as MAC-CC) to analyze the mediation model with a censored mediator given data from a case–control study, based on the semiparametric accelerated failure time model along with a pseudo-likelihood function. We adapted the measures for assessing the indirect and direct effects using counterfactual definitions. We conducted simulation studies to investigate the performance of MAC-CC and compared it to those of the naïve approach and the complete-case approach. MAC-CC accurately estimates the coefficients of different paths, the indirect effects and proportions of the total effects mediated. We applied the proposed and existing approaches to the mediation study of genetic variants, a woman's age at menopause and type-2 diabetes based on a case–control study of type-2 diabetes. Our results indicate that there is no mediating effect from the age at menopause on the association between the genetic variants and type-2 diabetes.

Keywords

Mediation analysis; censored mediator; case–control study; indirect effects; semiparametric accelerated failure time model

CORRESPONDING AUTHOR: Address for correspondence and reprints: Dr. Jian Wang, Department of Biostatistics, Unit 1411, The University of Texas MD Anderson Cancer Center, 1400 Pressler Dr, FCT4.6084, Houston, TX 77030, U.S.A., Phone: (713) 794-4210, Fax: (713) 563-4243, jianwang@mdanderson.org.

Declaration of Conflicting Interests

The authors declare that there is no conflict of interest.

1. Introduction

Mediation analysis is an approach for assessing the direct effect (*DE*) and indirect effect (*IE*) of an initial variable on an outcome variable through a mediator.¹ This approach has been widely used in many research fields, such as psychology, social science, health science, genetic epidemiology, prevention research, and political communication research.^{2–5} In some applications, the mediators cannot be observed completely and are subject to right-censoring. For example, in our motivating study on the association between genetic variants and the risk of type-2 diabetes in women, the potential mediator, age at menopause, is right-censored for women who have not gone through menopause.⁶ Another example is a lung cancer study in which the age at which an individual quits smoking, which is right-censored for current smokers, can also be a potential mediator for the association between genetic variants and lung cancer risk.

Most of the literature of mediation analysis given censored data has focused on censored outcomes. Among the existing approaches for mediation modeling with a censored mediator, the complete-case analysis, which uses only the subjects with complete data, does not fully utilize all observed information and is statistically inefficient.^{6,7} A naïve method that uses all observed values regardless of the censoring status may result in biased estimation and misleading inference.^{6–10} Wang and Zhang¹¹ proposed the Tobit mediation model, in which the structural equation method with a truncated multivariate normal distribution is used to account for the censored data. Wang and Shete⁶ used a multiple imputation approach to deal with a censored mediator, which, compared to other approaches, was shown to provide more accurate estimations for the coefficients of different paths as well as the *IE* and proportion of the total effect mediated by the mediator in the model. However, the previous work^{6,11} on mediation analysis with a censored mediator focused on the scenario in which the outcome variable is a continuous variable. To our knowledge, there is no statistical framework for mediation analyses when the outcome variable is binary and the mediator is a censored variable. In addition to lacking considerations of binary outcomes, the aforementioned methods^{6,11} require a parametric distribution assumption on the mediators, and may not be robust in terms of model mis-specification.

Our motivating study is a candidate gene case–control study of type-2 diabetes, in which we consider the association among genetic variants, a woman's age at menopause and her risk of developing type-2 diabetes. Type-2 diabetes is a chronic disease, caused by a complex interplay of environmental and genetic factors.^{12–17} Some studies have suggested an association between a woman's age at menopause and type-2 diabetes^{18–20}; while others have shown such association to be weak or not significant.^{21–23} Meanwhile, an association between genetic variants and a woman's age at menopause has been supported by genetic epidemiologic data.^{24,25} Therefore, we hypothesize that the age at menopause, which is censored for women who have not gone through menopause, is a potential mediator of the association between genetic variants and type-2 diabetes. We test this hypothesis by assessing the dual pathways from genetic variants to type-2 diabetes: (1) via a direct effect, and (2) through the age at menopause. Such a conceptual mediation model is presented in Figure 1 (A). In this motivating study, an analysis of the mediation model includes both a

binary outcome variable (i.e., type-2 diabetes status) and a censored mediator (i.e., age at menopause).

In addition to binary outcomes and censored mediators, another challenge of our motivating example is to use the data from a case-control study in which the control group is not truly representative of the general population. It has been shown that, without any special consideration, standard methods will lead to biased estimations for the association between the initial variable (e.g., genetic variants) and the mediator (e.g., age at menopause).^{5,26–30} This is because the observed values of the mediator are not sampled from the distribution of the mediator in the general population due to the biased selection of the controls.^{4,5} Therefore, such a mediation analysis based on standard methods that ignore the sampling bias may lead to biased estimates of the *IEs* and the proportions mediated.

Considering these methodological challenges, we propose an efficient yet robust approach to analyze the mediation model in a case-control study with a censored mediator, which we denote as “mediation analysis with a censored mediator in a case-control study (MAC-CC).” Specifically, we use the semiparametric accelerated failure time (AFT) model with an unspecified error distribution to assess the relationship between the initial variable and the mediator (path *a* in Figure 1 (A)). Such a relationship is calculated using a modified weighted least square method, denoted as the double-weighted least square method. Besides handling the right-censoring by assigning a weight, the double-weighted least square method also accounts for the sampling mechanism of the case-control design by including an additional weight determined by the prevalence of the disease (i.e., outcome variable) in the general population, which is usually available for common diseases (e.g., type-2 diabetes in the motivating application). The tail distribution of the censored mediator is estimated from the AFT model residuals by the Kaplan-Meier estimator. Based on the estimated coefficient *a* and the tail distribution of the mediator, a pseudo-likelihood function³¹ was employed to assess the relationships between the mediator and the outcome (path *b* in Figure 1 (A)) and between the initial variable and the outcome (path $\tilde{c}$ in Figure 1 (A)).

In mediation analysis, the measures commonly used to assess the *IE* include the difference in regression coefficients (Figure 1 (A)); total effect *c* – direct effect $\tilde{c}$ and the product of the regression coefficients (*ab*), which require both the mediator and outcome variables to be continuous and normally distributed and assume linearity among the variables in the model setting. When the mediator is a censored variable and the outcome is binary, the commonly used *IE* definitions are not applicable. Therefore, in this study, we employed the counterfactual definitions of the natural direct and indirect effects used in causal inference research, which have been widely applied in mediation analysis, especially in scenarios involving nonlinearities and interactions.^{32–36} We discuss the assumptions for the identification of the natural direct and indirect effects in mediation analyses in the Methods section.

In this study, we compared the performance of the proposed MAC-CC with those of existing methods, the naïve and complete-case approaches. Specifically, the naïve approach uses the entire data set regardless of the censoring status when estimating the coefficients for paths *b* and $\tilde{c}$; while the complete-case approach excludes censored individuals when estimating the

coefficients for all paths. Neither the naïve nor the complete-case approach accounts for the case-control sampling design, which leads to biased estimation. For comparing the performances of the three different approaches, we employed the same formulas for the direct and indirect effects obtained via the counterfactual framework.

In Section 2, we introduce the mediation model and develop the pseudo-likelihood approach for estimating the regression coefficients for different paths. We assess the empirical performance of the proposed MAC-CC via simulation studies in Section 3, describe a data analysis for the motivating application in Section 4, and provide a discussion in Section 5.

2. Methods

For a continuous outcome Y , and mediator T , a single mediator model is depicted in Figure 1(B), which can be specified using the following three regressions^{1,3,37,38}:

$$\begin{aligned} Y &= c_0 + cX + \varepsilon_1 \\ T &= a_0 + aX + \varepsilon_2 \\ Y &= b_0 + bT + \tilde{c}X + \varepsilon_3 \end{aligned}$$

where X , T and Y denote the initial variable, mediating variable, and outcome variable, respectively, and ε_1 , ε_2 , and ε_3 are the error terms in the model. The three equations establish a basic mediation model,³⁷ where (1) the first equation shows that X is associated with Y , which may be mediated (path c); (2) the second equation shows that X is associated with T (path a); and (3) the last equation shows that T affects Y while controlling for X (path b). Based on the mediation model, the total effect of X on Y (path c) is decomposed into two parts: the IE (through paths a and b) and the DE (through path $\tilde{c}$); see Figure 1 (B). In our study, the outcome variable Y is a binary variable and the mediator T is a time-to-event variable subject to right-censoring. The aforementioned standard mediation model cannot be directly applied and needs to be extended to accommodate the data features.

Let T be the time to the event of interest, C be the right-censoring time for the mediator, and Z be the other covariates. Since we are interested in the genetic variant as the initial variable, we consider X as a binary variable, assuming a dominant or recessive genetic model. Given a data set of n individuals, we observe $(y_i, m_i, \delta_i, x_i, z_i^T)$, $i = 1, \dots, n$, where $m_i = \min(t_i, c_i)$ and $\delta_i = I(t_i < c_i)$. Given an individual i , in a mediation model, the relationships between Y and the mediator T and initial variable X (paths b and $\tilde{c}$) can be expressed using a logistic regression model:

$$\text{Logit}\{\Pr(y_i = 1 | t_i, x_i)\} = b_0 + bt_i + \tilde{c}x_i + \gamma^T z_i, \quad (1)$$

where b_0 , b , $\tilde{c}$ and γ are the regression coefficients. In the presence of right-censoring, the likelihood function for the observed data $(y_i, m_i, \delta_i, x_i, z_i^T)$ for an individual i can be written as³¹

$$\begin{aligned}
& L(y_i, m_i, \delta_i, x_i, z_i) \\
&= \left\{ \Pr(y_i | m_i, x_i, z_i) f(m_i | x_i, z_i) \right\}^{\delta_i} \left\{ \int_{m_i}^{\infty} \Pr(y_i | t, x_i, z_i) dF(t | x_i, z_i) \right\}^{1 - \delta_i} \\
&= \left\{ \frac{\exp(b_0 + bm_i + \tilde{c}x_i + \gamma^T z_i)^{y_i}}{1 + \exp(b_0 + bm_i + \tilde{c}x_i + \gamma^T z_i)} f(m_i | x_i, z_i) \right\}^{\delta_i} \left\{ \int_{m_i}^{\infty} \frac{\exp(b_0 + bt + \tilde{c}x_i + \gamma^T z_i)^{y_i}}{1 + \exp(b_0 + bt + \tilde{c}x_i + \gamma^T z_i)} dF(t | x_i, z_i) \right\}^{1 - \delta_i},
\end{aligned}
\tag{2}$$

where $f(t | x_i, z_i)$ and $F(t | x_i, z_i)$ denote the conditional density and cumulative distribution function, respectively, of the mediator (T) given the initial variable X and the covariates Z . Specifically, the relationship between the initial variable and the mediator (path a) can be expressed by the AFT model^{39–41}:

$$\log(t_i) = a_0 + ax_i + \tilde{\gamma}^T z_i + \varepsilon_i = A_i^T \theta + \varepsilon_i, \quad (3)$$

where $A_i = (1, x_i, z_i^T)^T$, $\theta = (a_0, a, \tilde{\gamma}^T)^T$ and ε_i represents the independently and identically distributed random errors with mean zero and an unspecified distribution. We use the AFT model to relate X and T in the mediation model since it provides an easy way to interpret the effect of the change in the initial variable on the change in the length of time to the event of interest.⁶

There are two parts for the likelihood in equation (2): $\left\{ \Pr(y_i | m_i, x_i, z_i) f(m_i | x_i, z_i) \right\}^{\delta_i}$ is for fully observed individuals; while $\left\{ \int_{m_i}^{\infty} \Pr(y_i | t, x_i, z_i) dF(t | x_i, z_i) \right\}^{1 - \delta_i}$ is for individuals with right-censored values for the mediator. Also note that functions $\Pr(y_i | m_i, x_i, z_i)$ and $\Pr(y_i | t, x_i, z_i)$ contain the regression coefficients for paths b and $\tilde{c}$, while $f(m_i | x_i, z_i)$ and $f(t | x_i, z_i)$ contain the coefficient for path a . To assess the coefficients of interest, we adapted the semiparametric approach proposed by Kong and Nan³¹ for regression models with a censored covariate. The estimation procedure involves two stages, where the coefficient for path a and the AFT model error distribution are estimated in stage 1 and then the coefficients for paths b and $\tilde{c}$ are assessed in stage 2, which is described below.

In stage 1, we adopt the double-weighted least square estimator to assess the AFT regression coefficient a . Specifically, given observations $(y_i, m_i, \delta_i, x_i, z_i^T)$, $i = 1, \dots, n$, where $m_i = \min(t_i, c_i)$ and $\delta_i = I(t_i < c_i)$, the classical least squares principle can be generalized by minimizing the weighted least squares,

$$L(\theta) = \frac{1}{2} \sum_{i=1}^n \frac{\delta_i w_i}{\hat{G}(m_i)} \{ \log(m_i) - A_i^T \theta \}^2, \quad (4)$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimator for the censoring survival function. Note that the whole weight includes two components: $\delta_i/\hat{G}(m_i)$, which handles the right-censoring, and w_i , which is the sampling weight to account for the sampling mechanism of the case-control study. The sampling weight can be determined by the outcome prevalence in the general population using the reciprocals of the sample fractions.^{28,30} Specifically, if we use $w_i = 1$ for individuals who have the outcome phenotype (i.e., cases; $Y = 1$), the sampling weight for individuals who do not have the outcome phenotype (i.e., controls; $Y = 0$) can be calculated as a ratio $n_1(1-K)/n_0K$, where n_1 and n_0 are the number of cases and controls, and K is the prevalence of the outcome in the general population, which is typically available for common diseases. We can obtain the double-weighted least square estimator by setting the derivative of equation (4) to zero and solving the equation. The double-weighted least square estimator $\hat{\theta} = (\hat{a}_0, \hat{a}, \hat{\gamma}^T)^T$ is given in a closed form as

$$\hat{\theta} = \left\{ \sum_{i=1}^n \frac{\delta_i w_i A_i A_i^T}{\hat{G}(m_i)} \right\}^{-1} \sum_{i=1}^n \frac{\delta_i w_i A_i \log(m_i)}{\hat{G}(m_i)}. \quad (5)$$

Based on the coefficients estimated from equation (5) and the AFT model error distribution obtained from stage 1, in stage 2, we assess the coefficients for paths b and $\tilde{c}$ using a pseudo-likelihood function. Given a random sample of n individuals, the log-pseudo-likelihood is given as

$$PL(\varphi) = \frac{1}{n} \sum_{i=1}^n w_i \left[\delta_i \log \left\{ \Pr_{\varphi}(y_i | m_i, A_i) \right\} + (1 - \delta_i) \log \int_{m_i - A_i^T \hat{\theta}}^{\tau} \Pr_{\varphi}(y_i | t + A_i^T \hat{\theta}, A_i) d\hat{\eta}_{\hat{\theta}}(t) \right], \quad (6)$$

where $\varphi = (b_0, b, \tilde{c}, \gamma^T)^T$, which is the set of parameters to be estimated; w_i is the sampling weight to account for the case-control study design, τ is the largest observed event time on a residual scale and $\hat{\eta}_{\hat{\theta}}(\cdot)$ is the AFT model error distribution, estimated from the censored

residuals of equation (3) by the Kaplan-Meier estimator given the estimates $\hat{\theta} = (\hat{a}_0, \hat{a}, \hat{\gamma}^T)^T$ obtained in stage 1. In equation (6), the first part, $\delta_i \log \{ \Pr_{\varphi}(y_i | m_i, A_i) \}$, is calculated using individuals with observed values for the mediator; while the second part, $(1 - \delta_i) \log \int_{m_i - A_i^T \hat{\theta}}^{\tau} \Pr_{\varphi}(y_i | t + A_i^T \hat{\theta}, A_i) d\hat{\eta}_{\hat{\theta}}(t)$, is calculated from right-censored individuals.

The conditional probabilities $\Pr_{\varphi}(y_i | m_i, A_i)$ and $\Pr_{\varphi}(y_i | t + A_i^T \hat{\theta}, A_i)$ are formulated using equations (1) and (3), respectively.

We maximized the log-pseudo-likelihood in equation (6) to obtain the estimators

$\hat{\varphi} = (\hat{b}_0, \hat{b}, \hat{c}, \hat{\gamma}^T)^T$ for paths b and $\tilde{c}$. For this purpose, we applied the minimization algorithm proposed by Nelder and Mead⁴² using the results from the complete-case analysis as the initial values. The Nelder and Mead algorithm is implemented by the “optim” function in the R package “stats”.⁴³

Under the assumption of correct specification of the mediation models, the consistency of $\hat{\theta}$ follows by the uniqueness of the minimum for the double-weighted least squares and the consistency of the estimated weights. Then we can derive the asymmetric normality of $\hat{\theta}$ using arguments similar to those of Shen et al.⁴⁴ Given the desirable asymptotic properties of $\hat{\theta}$ from the first stage, we can follow the arguments of Kong and Nan³¹ to obtain the consistency and asymptotic normality of $\hat{\varphi}$ from the second stage estimation.

Indirect Effect with a Censored Mediator and a Binary Outcome

As mentioned, the commonly used measures of the *IEs* (i.e., ab and $c\tilde{c}$) are only applicable for the mediation model with a continuous and normally distributed outcome variable and mediator. When the mediation model involves a censored mediator and a binary outcome, as discussed in this study, the commonly used *IE* measures are not applicable. Therefore, we used the counterfactual framework^{32–34,45–47} to address this issue of the assessment of *IEs* in such a mediation model.

Let Y_x and T_x respectively be the values of the outcome Y and mediator T that would have been observed had the initial variable X been set to x . Let Y_{xt} be the value of the outcome that would have been observed had T and X been set to t and x . Based on the counterfactual framework, the natural indirect effect (called indirect effect [*IE*] in this study), conditional on the covariates Z , compares the effects of the mediator T at values of T_x and T_{x^*} on the outcome Y when the initial variable X is set to x , which is defined as follows³²:

$$IE = E(Y_{xT_x} | Z) - E(Y_{xT_{x^*}} | Z) = \Pr(Y_{xT_x} = 1 | Z) - \Pr(Y_{xT_{x^*}} = 1 | Z). \quad (8)$$

And the natural direct effect (called direct effect [*DE*] in this study), conditional on Z , assesses the effect of the initial variable X on the outcome variable Y by setting the mediator T to the value it would have been if the initial variable had the reference value of x^* , which is defined as follows:

$$DE = E(Y_{xT_{x^*}} | Z) - E(Y_{x^*T_{x^*}} | Z) = \Pr(Y_{xT_{x^*}} = 1 | Z) - \Pr(Y_{x^*T_{x^*}} = 1 | Z). \quad (9)$$

Furthermore, the total effect (*TE*), conditional on Z , compares the average outcome Y if the initial variable X had been set to x versus x^* , which is defined as follows:

$$TE = E(Y_x|Z) - E(Y_{x^*}|Z) = \Pr(Y_x = 1|Z) - \Pr(Y_{x^*} = 1|Z). \quad (10)$$

Note that, given such definitions, DE and IE can still hold the decomposition property that TE can be decomposed into DE and IE , as shown below.

$$\begin{aligned} TE &= E(Y_x|Z) - E(Y_{x^*}|Z) = E(Y_{xT_x}|Z) - E(Y_{x^*T_{x^*}}|Z) \\ &= \left\{ E(Y_{xT_x}|Z) - E(Y_{xT_{x^*}}|Z) \right\} + \left\{ E(Y_{xT_{x^*}}|Z) - E(Y_{x^*T_{x^*}}|Z) \right\} \\ &= IE + DE \end{aligned}$$

In this study, we focused on the definitions of DE and IE using the risk difference scale.^{32–34} Therefore, IE is the difference in probability of being in a phenotypic state (e.g., type-2 diabetes) for individuals with different genotypes (e.g., AA and Aa vs. aa for the dominant model) that can be attributed to the mediator (e.g., age at menopause). Whereas DE is the difference in probability of being in a phenotypic state for individuals with different genotypes that can be attributed to a direct path or other paths not included in the mediation model. TE is the difference in probability of being in a phenotypic state because of different genotypes. Also, as we consider a binary initial variable X , we assess the difference in probability of being in a phenotypic state between $X=1$ vs. $X=0$ (reference value).

To identify IE , DE and TE in the mediation analysis using the observed data, we require certain assumptions about the absence of unmeasured confounding variables.^{32,34,35,48–50} In particular, conditional on measured confounders Z , we must assume there are no unmeasured confounding variables for (A1) the X - Y relation: $X \perp Y_{xt} | Z$; (A2) the T - Y relation conditional on X : $T \perp Y_{xt} | X, Z$; and (A3) the X - T relation: $X \perp T_x | Z$. We also need to assume that, conditional on measured confounders Z , there is no unmeasured confounding variable that is affected by the initial variable X and which itself affects the T - Y relation: $T_{x^*} \perp Y_{xt} | Z$ (A4). In addition to the assumptions about the absence of unmeasured confounders, we need the standard assumption of consistency (A5): $T_x = T$ when $X=x$ and $Y_{xt} = Y$ when $X=x$ and $T=t$,^{48,50,51} which ensures that the outcome value is not changed if the values of the initial variable and mediator are set to the values they would naturally take.⁵⁰

The identifiability of IE and DE under certain assumptions was established previously.^{32,34,36} Derivations for calculating IE and DE are shown in the Supporting Information. If the assumptions described above hold, IE and DE are identified and can be assessed using the observed data as follows:

$$\begin{aligned} IE &= \Pr(Y_{xT_x} = 1|Z) - \Pr(Y_{xT_{x^*}} = 1|Z) \\ &= \int \Pr(Y = 1|t, x, z) dF(t|x, z) - \int \Pr(Y = 1|t, x, z) dF(t|x^*, z) \end{aligned}$$

and

$$\begin{aligned} DE &= \Pr(Y_{xT_{x^*}} = 1 | Z) - \Pr(Y_{x^*T_{x^*}} = 1 | Z) \\ &= \int \Pr(Y = 1 | t, x, z) dF(t | x^*, z) - \int \Pr(Y = 1 | t, x^*, z) dF(t | x^*, z) \end{aligned}$$

where $F(t | x, z)$ is the cumulative distribution function of the mediator given the initial variable X and the covariates Z .

Therefore, TE can be assessed as

$$\begin{aligned} TE &= IE + DE = \left\{ \Pr(Y_{xT_x} = 1 | Z) - \Pr(Y_{xT_{x^*}} = 1 | Z) \right\} + \left\{ \Pr(Y_{xT_{x^*}} = 1 | Z) - \Pr(Y_{x^*T_{x^*}} = 1 | Z) \right\} \\ &= \int \Pr(Y = 1 | t, x, z) dF(t | x, z) - \int \Pr(Y = 1 | t, x^*, z) dF(t | x^*, z) \end{aligned}$$

If the models in equations (1) and (3) are correctly specified, based on the estimated AFT model error distribution for the mediator and the estimated coefficients for different paths in the mediation model, we can assess IE , DE , and TE based on the estimated AFT model error distribution $\hat{\eta}_{\hat{\theta}}(\cdot)$, as well as the estimated coefficients $\hat{\varphi} = (\hat{b}_0, \hat{b}, \hat{c}, \hat{\gamma}^T)^T$ and $\hat{\theta} = (\hat{a}_0, \hat{a}, \hat{\gamma}^T)^T$ as obtained above.

Furthermore, based on the estimated TE and IE , we can assess the proportion mediated (PM) of X on Y mediated by the mediator as $\hat{PM} = \hat{IE} / \hat{TE}$. We report the proportion mediated only when there is evidence of an IE (i.e., the IE is significant).^{52,53}

3. Simulation

Simulation Approach

We conducted simulations to investigate the performance of the proposed MAC-CC approach and compare it to the naïve and complete-case approaches for analyzing a mediation model with a binary outcome and a censored mediator given data from a case-control study. Specifically, for each individual i , the initial variable x_i (e.g., dominant model for genetic variants) was generated based on the binomial distribution using a pre-specified minor allele frequency, which was assumed to be 0.3. Given x_i , the time-to-event mediator t_i was generated using an AFT model $\log(t_i) = a_0 + ax_i + e_i$, where $e_i \sim \text{Normal}(0, 1)$. In the simulation, we considered $a_0 = 6$ and $a = 0.3$. The right-censoring time c_i was generated independently from uniform distributions corresponding to different censoring percentages, ~20% and ~40%. 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 logistic regression model as in equation (3). We fixed $\tilde{c}$ as 0.5 and considered two different values for $b = 0.3$ and 0.4 to include different values for the IE s. Different values of b_0 were selected to generate data with different levels of disease

prevalence in the general population. In this way, we simulated a large amount of data on the population, from which we randomly sampled 500 cases and 500 controls.

When there is no *IE* for the mediator, as either a or b is zero (i.e., null hypothesis), we considered 6 scenarios with respect to different values of coefficient b and different censoring percentages (Tables 1 and 2). When there is an *IE* of the mediator, as both a and b are non-zero (i.e., alternative hypothesis), we considered 12 scenarios with respect to different values of coefficient b , different censoring percentages and different levels of disease prevalence (Tables 3 and 4). In the simulation studies, we also calculated the expected *IEs* and proportions mediated for different scenarios using the pre-specified parameters, including a_0 , a , b_0 , b and $\tilde{c}$, and the normal distribution for the conditional probability of the mediator. Note that when using the MAC-CC approach, the non-parametric Kaplan-Meier estimator from the censored residuals was used to calculate the conditional probability of the mediator in the assessment of *IEs* and proportions mediated instead of parametric distributions. We employed the bias-corrected and accelerated (BCa) bootstrap⁵⁴ to determine the confidence intervals (CIs) for the *IEs*. We compared the performance of the MAC-CC approach to those of the naïve and complete-case approaches.

Simulation Results

Tables 1 and 2 report the results for the null scenarios, that is, there is no *IE* through the mediator for the association between the initial variable and the outcome. In Table 1, we report the means and standard errors of the estimated coefficients for the different paths, a_0 , a , b , b_0 and $\tilde{c}$, obtained using the naïve, complete-case and MAC-CC approaches. The naïve approach provided accurate estimations for the coefficients a_0 , a and $\tilde{c}$ for all scenarios; but provided highly biased estimation for the coefficient b_0 . It provided accurate estimation for the coefficient of path b when it was zero (scenarios 1–4), but underestimated the coefficient of path b when it was non-zero (scenarios 5 and 6). The complete-case approach provided accurate estimations for the coefficients a_0 , a , b and $\tilde{c}$; but overestimated the coefficient b_0 , especially when b was zero (scenarios 1–4). In contrast, the MAC-CC approach provided accurate estimations for all coefficients for the different scenarios. For example, for scenario 6, when the true value of b_0 and b were -3 and 0.4 , respectively, the estimates obtained by the naïve approach were -0.859 and 0.117 , which are biased; while the estimates by the MAC-CC approach were -3.037 and 0.405 , which are close to the true simulation values. For scenario 1, when the true value of b_0 was -3 , the estimate by the complete-case approach was -0.3 , which is overestimated; while the estimate by the MAC-CC approach was -3.022 which is close to the pre-specified value of b_0 .

In Table 2, we report the means and standard errors of the estimated *IEs*, as well as the coverage probabilities of the 95% CIs when testing the *IEs*. For all scenarios, the three approaches performed similarly and provided accurate estimations that were close to 0 for *IE*. The coverage probability was calculated as the proportion of the 500 simulated replicates for which the 95% BCa CIs contained the expected value of *IE*. Under the null hypothesis of no *IE*, the naïve, complete-case and MAC-CC approaches provided reasonable coverage probabilities for *IE* that are close to the nominal value of 0.95 for all scenarios.

Tables 3 and 4 report the results for the alternative scenarios, that is, there is an IE through the mediator for the association between the initial variable and the outcome. In Table 3, we report the means and standard errors of the estimated coefficients for the different paths, a_0 , a , b_0 , b and $\tilde{c}$, obtained using all three approaches. For all scenarios, the naïve approach overestimated the a_0 , a , b_0 and $\tilde{c}$ coefficients and underestimated the b coefficient; the complete-case approach overestimated the a and b_0 coefficients; while the MAC-CC approach provided accurate estimates for all the coefficients a_0 , a , b_0 , b and $\tilde{c}$. For example, for scenario 1, when the true values of the coefficients a_0 , a , b_0 , b and $\tilde{c}$ were 6, 0.3, -6, 0.4, and 0.5, respectively, the estimates by the naïve approach were 6.149, 0.354, -1.278, 0.160 and 0.589, which are biased compared to the true values. The estimates by the complete-case approach were respectively 6.085, 0.353, -2.837, 0.4 and 0.505, for which the estimates of a and b_0 were overestimated. In contrast, the estimates by the MAC-CC approach were respectively 5.996, 0.307, -5.991, 0.396 and 0.509, which are close to the true simulation values. The magnitude of the bias in the estimation of the coefficients tends to increase when the censoring percentage increases or when the prevalence of the outcome decreases.

The results for the IE s and proportions mediated under the alternative hypothesis are reported in Table 4. As expected, the naïve approach provided biased estimates for both the IE and the proportion mediated for most of the scenarios; the complete-case approach consistently provided overestimated values for the IE but relatively accurate estimates for the proportion mediated; while the MAC-CC approach provided accurate estimations for both the IE and the proportion mediated for all scenarios. For example, for scenario 8, when the expected IE and proportion mediated were respectively 0.013 and 0.232, the estimated values obtained using the naïve approach were 0.008 and 0.053, which were highly underestimated; the estimated values obtained using the complete-case approach were 0.031 and 0.233, which overestimated the IE ; whereas those obtained using the MAC-CC approach were 0.012 and 0.235, which are close to the expected values of the IE and proportion mediated. When assessing the IE s, the complete-case approach always provided the largest standard errors, while the naïve and MAC-CC approaches provided similar standard errors. When assessing the proportion mediated, when the censoring percentage was ~40%, the complete-case approach provided larger standard errors compared to the proposed MAC-CC approach. This observation showed that the complete-case approach might provide consistent but less efficient estimates for the proportion mediated when the censoring percentage is higher. We also report the 95% coverage probabilities for the IE and proportion mediated based on all approaches. The naïve approach provided 95% coverage probabilities for the IE and proportion mediated that are less than the nominal coverage probability of 0.95 for most scenarios. The complete-case approach provided 95% coverage probabilities for the IE that are less than the nominal value for all scenarios, but probabilities for the proportion mediated that are close to the nominal value for most scenarios. In contrast, the MAC-CC approach provided probabilities for both IE and proportion mediated that are close to the nominal value for all scenarios. For example, for scenario 4, the 95% coverage probabilities for the IE and proportion mediated were respectively 0.67 and 0.31 by the naïve approach; 0.12 and 0.916 by the complete-case approach; and 0.944 and 0.938 by the MAC-CC approach.

4. Application to the Study of Genetic Variants, Age at Menopause and Type-2 Diabetes

The purpose of the real data application is to assess the mediating effect of a woman's age at menopause on the association between single nucleotide polymorphisms (SNPs) and type-2 diabetes risk, as described in the Introduction. We applied the naïve, complete-case and proposed MAC-CC approaches for mediation analysis using the data from a candidate gene case-control study for type-2 diabetes, based on the Multi-Ethnic Study of Atherosclerosis (MESA) cohort provided by the National Center for Biotechnology Information and downloaded from dbGaP.⁵⁵ As presented in Figure 1 (A), we considered a conceptual mediation model with SNPs as the initial variable (X), the age at menopause as the mediator (T), and type-2 diabetes status as the outcome variable (Y). Of note, a woman's age at menopause was considered to be censored if she had not gone through menopause.

Based on the MESA data, genotypes of 47,871 SNPs from candidate genes for 2,956 women were included in the mediation analysis. The censoring percentage for a woman's age at menopause was ~14.5% and the prevalence of type-2 diabetes in the female population was reported as ~11.2%.⁵⁶ We excluded SNPs with minor allele frequency < 0.05 according to the standard quality control procedure. We first assessed the associations between all SNPs and type-2 diabetes, as well as the associations between SNPs and age at menopause, assuming the dominant model for the SNPs. At a significance level of 0.01, we identified three SNPs, rs11771343, rs3924519 and rs11842863, associated with both type-2 diabetes and age at menopause. Specifically, the p -values of the three SNPs were 1.86×10^{-3} , 3.63×10^{-3} and 3.98×10^{-3} , respectively, for their association with type-2 diabetes; and 5.9×10^{-3} , 9.85×10^{-3} and 6.38×10^{-3} , respectively, for their association with a woman's age at menopause. We included the three SNPs in the mediation analysis, using the naïve, complete-case and MAC-CC approaches. To assess the significance for the IE s, we conducted BCa bootstrapping to calculate the 95% CIs. Age and/or ethnicity were adjusted in the mediation model.

Table 5 reports the results from the mediation analysis for the three SNPs, age at menopause and type-2 diabetes, including the estimated TE s, DE s and IE s and 95% BCa CIs for the IE s. The MESA case-control data showed a non-significant association between a women's age at menopause and risk of developing type-2 diabetes, which implies a non-significant IE . In this situation where there is no IE , all three approaches performed similarly, which is consistent with the simulation results. For example, for the SNP rs3924519, the IE s were 0.0006, 0.0005 and 0.0008 using the naïve, complete-case and MAC-CC approach, respectively. The 95% CIs of the IE s were respectively $[-0.0013, 0.0022]$, $[-0.0009, 0.002]$ and $[-0.0002, 0.0032]$ for the three approaches, which included zero. Similar results were obtained for the other two SNPs. These findings suggest that there was no mediating effect of the age at menopause on the association between the three SNPs, rs11771343, rs3924519 and rs11842863, and type-2 diabetes risk.

Sensitivity Analysis

The most common concern for mediation analysis in practice is ignoring possible confounders. It has been suggested that great effort should be made to collect data on variables that are potential confounders; and that if it is impossible to control for such confounders, the “no-unmeasured-confounder” assumptions (i.e., Assumptions A1-A4 in the Supporting Information Appendix) need to be evaluated using the investigators’ subject matter knowledge and sensitivity analysis techniques.^{32,46,57} In the real data application of genetic variants, Assumptions A1 (i.e., no unmeasured confounder for the SNP–type-2 diabetes relation) and A3 (i.e., no unmeasured confounder for the SNP–age at menopause relation) are plausible for the genetic variants (X) subject to no population stratification since race was adjusted for in the mediation analysis as a covariate.⁵⁸ To investigate Assumption A2 (i.e., no unmeasured confounder for the age at menopause–type-2 diabetes relation), we employed a recently proposed sensitivity analysis technique^{59,60} that makes relatively fewer assumptions compared to other sensitivity analysis approaches. This technique includes two parameters: γ denotes the maximum risk ratio relating the unknown confounder U and type-2 diabetes (Y) among the exposed individuals (i.e., $X = 1$) across different values of age at menopause (T); and λ denotes the maximum risk ratio relating U and SNP (X) across different levels of T . The sensitivity analysis shows that the IE estimate is unlikely to be altered given a wide range of the two parameters γ and λ due to confounding variables (Supporting Information Table S1).

We also explored Assumption A4 (i.e., no unmeasured SNP-induced confounder for the age at menopause–type-2 diabetes relation) by adapting a non-parametric sensitivity analysis technique proposed by VanderWeele and Chiba (2014) for the mediation model with a binary mediator.⁶¹ Using a sensitivity analysis parameter of a quarter of a standard deviation for the risk probability scale of type-2 diabetes as suggested by VanderWeele and Chiba,⁶¹ both the IE s and DE s were robust: -0.0005 , 0.001 and -0.0017 for the IE s and 0.044 , -0.037 and 0.0468 for the DE s for the respective SNPs rs11771343, rs11842863 and rs3924519. If a very large sensitivity analysis parameter (i.e., standard deviation) was used, the large DE s were still robust (0.0428 , -0.0363 , 0.046). The magnitude of the small IE s may be subject to violation of this assumption (-0.0016 , 0.0018 , -0.0025); however, the IE s were very small compared to the DE s and the signs of the point estimates retained the same directions. Thus, based on the sensitivity analysis, our qualitative conclusion would not change that age at menopause does not explain much of the observed association between the genetic variants and type-2 diabetes.

5. Discussion

In this study, we focused on a single mediator model, for which the mediator is a time-to-event variable subject to right-censoring and the outcome is a binary variable, using data from a case–control study. To address the challenges, we proposed the MAC-CC approach for such a mediation model, which employed the semiparametric AFT model with an unspecified error distribution, combined with a pseudo-likelihood function. We adapted the counterfactual definitions used in causal inference research for the assessments of the DE and IE , as well as the TE and proportion mediated. Based on simulations, we showed that,

compared to the naïve and complete-case approaches, the MAC-CC approach can provide accurate estimations for the coefficients of different paths (a , b and $\tilde{c}$) as well as the IE and proportion mediated for the mediation model where the mediator is a time-to-event variable and the outcome is a binary variable.

We applied the naïve, complete-case and MAC-CC approaches to a study of SNPs, a woman's age at menopause and type-2 diabetes risk to assess the mediating effect of the age at menopause on the SNP–type-2 diabetes association, using data from a case–control study of type-2 diabetes (MESA data). Three SNPs, rs11771343, rs11842863 and rs3924519, were included in the mediation analysis since they were associated with both a woman's age at menopause and type-2 diabetes. The real data application showed that there were no mediating effects of a woman's age at menopause on the associations of the three SNPs and type-2 diabetes.

Important assumptions are needed for mediation analysis since the mediating effect is considered to be a causal effect.^{52,62} In this study, in addition to the assumptions about unmeasured confounders for the derivations of IE and DE , we also assumed that (1) the mediation model is specified correctly, including correctly specified causal orders and causal directions; and (2) all variables involved in the mediation model do not have measurement errors. For the real data application of SNPs, a woman's age at menopause and type-2 diabetes, we conceptualized the mediation model, including the causal orders and causal directions, based on the literature.^{12–20,24,25} We assumed that there were no measurement errors for all variables in the mediation model, as well as no unmeasured confounders that affected the relationships among the SNPs, age at menopause, and type-2 diabetes in the mediation model (Assumptions A1-A4 in the Supporting Information Appendix). Sensitivity analysis for the “no-unmeasured-confounder” assumptions was conducted and discussed in the real data application section.

We also assumed that the censoring process for the mediator variable is independent of the mediator T , initial variable X , outcome Y and covariates Z . We further conducted a sensitivity analysis using simulation to investigate the assumption for the censoring process. In the simulation, we generated the time-to-event mediator T using an AFT model as described in the Methods section; but generated the right-censoring time C for the mediator T using different distributions conditional on X : a uniform distribution when $X = 0$ and a normal distribution when $X = 1$. Even in such a scenario, the MAC-CC approach provided accurate estimations for the coefficients, IE s and proportions mediated, as well as the corresponding coverage probabilities of the 95% CIs for all scenarios (Supporting Information Table S2). This sensitivity analysis suggests that the proposed method has some degree of robustness to the violation of the independence assumption.

We also investigated the performance of our MAC-CC approach for scenarios in which the censoring percentage is high (i.e., 50% and 80%) using simulations (Supporting Information Table S3). When the censoring percentage was 50%, the MAC-CC approach provided accurate estimations for the coefficients, IE s and proportions mediated. When the censoring percentage was quite high (80%), the IE s and proportions mediated were underestimated when using the MAC-CC approach. This is expected since it is well known that the inverse

weighted approach has limitations when the censoring percentage is very high.^{63,64} In our simulations with 80% censoring, the inverse weighted approach underestimated the coefficients a_0 and a , which in turn affected the estimations of b , IE and the proportion mediated. However, even in this situation, the MAC-CC approach outperformed the naïve and complete-case approaches. For example, for scenario 7, in which the expected IE was 0.006, the naïve and complete-case approaches provided either an underestimated (0.001) or overestimated (0.02) IE estimate; while the IE estimate obtained using the MAC-CC approach was closer to the expected value (0.003).

This study focused on scenarios for which the mediator is a time-to-event variable subject to right-censoring. However, if the mediator is a left-censored variable, the MAC-CC approach is still applicable. As discussed in previous work,³¹ one can consider transforming the left-censored variable to a right-censored variable using a monotone decreasing transformation function and then apply the MAC-CC approach. For the MAC-CC approach, the sampling weight that accounts for the sampling mechanism of the case-control study is calculated using the prevalence of the outcome variable in the general population, since the motivating example did not involve a stratification study design using certain covariates (e.g., race). In practice, since different subjects may have different levels of covariates, we suggest using the covariate-adjusted prevalence of the outcome if the study design involves stratification according to the covariates and such information is available.

In this study, IE was defined by setting X to x and DE was defined by setting the mediator T to the value it would have been if the initial variable had the reference value of x^* .

Alternatively, IE can be defined when X is set to x^* , that is $IE = E(Y_{x^*T_{x^*}} | Z) - E(Y_{x^*T_x} | Z)$;

while DE is defined by setting the mediator T to the value it would have been if the initial variable had the value of x , that is $DE = E(Y_{xT_x} | Z) - E(Y_{xT_{x^*}} | Z)$. In this case, TE can still be

decomposed into IE and DE such as $TE = E(Y_{x^*} | Z) - E(Y_x | Z) = IE + DE$. Note that the estimation method for assessing the IE s and proportion mediated is still applicable when the aforementioned alternative definitions of IE and DE are employed.

In a mediation analysis, the proportion mediated is employed to capture the portion of total effect that is due to the mediation effect, which is usually reported when there is evidence of a mediating effect (indirect effect).^{52,53} The measure of the proportion of total effect mediated by the mediator is applicable and interpretable only when the signs of indirect and direct effects are the same.⁴⁶ When there is no significant indirect effect through the mediator (i.e., indirect effect is close to 0), the sign of indirect effect can become arbitrary (i.e., positive or negative). In this case, the proportion mediated could be negative (i.e., total effect and indirect effect have opposite signs), which cannot be interpreted appropriately and meaningfully.⁶⁵ One possible solution is to take the absolute values of direct and indirect effects before calculating the proportions mediated.⁶⁶ We also conducted simulation studies to investigate the performance of the MAC-CC approach for scenarios in which the proportion mediated is high (e.g., 40%, 60% and 80%). Regardless of the magnitude of the proportion mediated, the biases of the estimated coefficients, IE s and proportions mediated by the MAC-CC were reasonable and the coverages of the IE s and proportions mediated

were close to the nominal level (Supporting Information Table S4). These simulations suggest that our proposed MAC-CC is quite stable in terms of changes in the proportion mediated.

There are some limitations to the proposed approach. First, the approach was developed for a binary initial variable (e.g., dominant or recessive genetic model), therefore, the analytics need to be further developed for a categorical or continuous initial variable (e.g., additive genetic model). Also, this approach considers only a single censored mediator; further work to include multiple censored mediators will be needed to extend this approach.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

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 R03CA192197. Jing Ning was supported in part by the NIH grant R01CA193878 and the Andrew Sabin Family Fellowship. Sanjay Shete was supported in part by the NIH grants R01DE022891 and R25DA026120; the Cancer Prevention Research Institute of Texas grants RP130123 and RP170259; and the Barnhart Family Distinguished Professorship in Targeted Therapy. Jian Wang, Jing Ning, and Sanjay Shete were supported in part by the NIH Cancer Center Support Grant P30CA016672. The MESA project has been conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA has been 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, NY: Erlbaum; 2008.
2. Shrout PE, Bolger N. Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol Methods*. 2002;7(4):422–445. [PubMed: 12530702]
3. Taylor AB, MacKinnon DP. Four applications of permutation methods to testing a single-mediator model. *Behav Res Methods*. 2012;44(3):806–844. [PubMed: 22311738]
4. 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(10):e47705. [PubMed: 23077662]
5. Wang J, Spitz MR, Amos CI, et al. 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(14):3458–3462. [PubMed: 20564069]
6. Wang J, Shete S. Estimation of indirect effect when the mediator is a censored variable. *Stat Methods Med Res*. 2017;962280217690414.
7. Wang HJ, Feng XD. Multiple imputation for M-regression with censored covariates. *J Am Stat Assoc*. 2012;107(497):194–204.
8. 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 Anal*. 2015;85:37–53. [PubMed: 25598564]

9. May RC, Ibrahim JG, Chu H. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Stat Med.* 2011;30(20):2551–2561. [PubMed: 21710558]
10. Tsimikas JV, Bantis LE, Georgiou SD. Inference in generalized linear regression models with a censored covariate. *Comput Stat Data Anal.* 2012;56(6):1854–1868.
11. Wang LJ, Zhang ZY. Estimating and testing mediation effects with censored data. *Struct Equ Modeling.* 2011;18(1):18–34.
12. Fuchsberger C, Flannick J, Teslovich TM, et al. The genetic architecture of type 2 diabetes. *Nature.* 2016;536(7614):41–47. [PubMed: 27398621]
13. 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(6):762–771. [PubMed: 26678054]
14. 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(9):981–990. [PubMed: 22885922]
15. Almind K, Doria A, Kahn CR. Putting the genes for type II diabetes on the map. *Nat Med.* 2001;7(3):277–279. [PubMed: 11231616]
16. 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]
17. 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]
18. 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(4):1012–1019. [PubMed: 23230098]
19. LeBlanc ES, Kaphahn 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.* 2017;24(1):64–72. [PubMed: 27465714]
20. Malacara JM, Huerta R, Rivera B, et al. Menopause in normal and uncomplicated NIDDM women: physical and emotional symptoms and hormone profile. *Maturitas.* 1997;28(1):35–45. [PubMed: 9391993]
21. Lopez-Lopez R, Huerta R, Malacara JM. Age at menopause in women with type 2 diabetes mellitus. *Menopause.* 1999;6(2):174–178. [PubMed: 10374226]
22. 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(12):4007–4014. [PubMed: 24170752]
23. Szmuiłowicz ED, Stuenkel CA, Seely EW. Influence of menopause on diabetes and diabetes risk. *Nat Rev Endocrinol.* 2009;5(10):553–558. [PubMed: 19687788]
24. Laven JS. Genetics of Early and Normal Menopause. *Semin Reprod Med.* 2015;33(6):377–383. [PubMed: 26569518]
25. 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(11):1294–1303. [PubMed: 26414677]
26. Li H, Gail MH, Berndt S, et al. Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genet Epidemiol.* 2010;34(5):427–433. [PubMed: 20583284]
27. Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol.* 2009;33(3):256–265. [PubMed: 19051285]
28. Richardson DB, Rzehak P, Klenk J, et al. Analyses of case-control data for additional outcomes. *Epidemiology.* 2007;18(4):441–445. [PubMed: 17473707]
29. Wang J, Shete S. Power and type I error results for a bias-correction approach recently shown to provide accurate odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet Epidemiol.* 2011;35(7):739–743. [PubMed: 21769937]
30. Wang J, Shete S. Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet Epidemiol.* 2011;35(3):190–200. [PubMed: 21308766]

31. Kong SC, Nan B. Semiparametric approach to regression with a covariate subject to a detection limit. *Biometrika*. 2016;103(1):161–174.
32. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. *Am J Epidemiol*. 2010;172(12):1339–1348. [PubMed: 21036955]
33. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143–155. [PubMed: 1576220]
34. Pearl J. Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence* 2001:411–420. ftp://ftp.cs.ucla.edu/pub/stat_ser/R273-U.pdf.
35. Huang YT, Pan WC. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*. 2016;72(2):402–413. [PubMed: 26414245]
36. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci*. 2010;25(1):51–71.
37. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51(6):1173–1182. [PubMed: 3806354]
38. Selig JP, Preacher KJ, Little TD. Modeling time-dependent association in longitudinal data: a lag as moderator approach. *Multivariate Behav Res*. 2012;47(5):697–716. [PubMed: 24771950]
39. Swindell WR. Accelerated failure time models provide a useful statistical framework for aging research. *Exp Gerontol*. 2009;44(3):190–200. [PubMed: 19007875]
40. Tein JY, MacKinnon DP. Estimating mediated effects with survival data. *New Developments in Psychometrics*. 2003:405–412.
41. Collett D *Modeling Survival Data in Medical Research*. Boca Raton, FL: CRC Press; 2003.
42. Nelder JA, Mead R. A simplex-method for function minimization. *Comput J*. 1965;7(4):308–313.
43. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.
44. Shen Y, Ning J, Qin J. Analyzing Length-biased Data with Semiparametric Transformation and Accelerated Failure Time Models. *J Am Stat Assoc*. 2009;104(487):1192–1202. [PubMed: 21057599]
45. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*. 2009;20(1):18–26. [PubMed: 19234398]
46. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods*. 2010;15(4):309–334. [PubMed: 20954780]
47. Li W, Zhou XH. Identifiability and estimation of causal mediation effects with missing data. *Stat Med*. 2017;36(25):3948–3965. [PubMed: 28783880]
48. Huang YT, Yang HI. Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology*. 2017;28(3):370–378. [PubMed: 28296661]
49. VanderWeele TJ. Causal mediation analysis with survival data. *Epidemiology*. 2011;22(4):582–585. [PubMed: 21642779]
50. Lange T, Hansen JV. Direct and indirect effects in a survival context. *Epidemiology*. 2011;22(4):575–581. [PubMed: 21552129]
51. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20(6):880–883. [PubMed: 19829187]
52. Kenny DA. Mediation. <http://davidakenny.net/cm/mediate.htm>. Accessed March 15, 2017.
53. Mordkoff JT. Quantitative methods in psychology. <http://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%202/MRC%20Meds.pdf>. Accessed June 18, 2018.
54. Efron B Better Bootstrap Confidence-Intervals. *J Am Stat Assoc*. 1987;82(397):171–185.
55. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39(10):1181–1186. [PubMed: 17898773]
56. 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.
57. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*. 2010;21(4):540–551. [PubMed: 20479643]

58. Vanderweele TJ. Varieties of sensitivity analysis for mediation. <http://www.da.ugent.be/cvs/pages/en/Presentations/Presentation%20Tyler%20VanderWeele.pdf>. Accessed June 18, 2018.
59. VanderWeele TJ. Mediation analysis: a practitioner's guide. *Annu Rev Public Health*. 2016;37:17–32. [PubMed: 26653405]
60. Ding P, Vanderweele TJ. Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Biometrika*. 2016;103(2):483–490. [PubMed: 27279672]
61. VanderWeele TJ, Chiba Y. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiol Biostat Public Health*. 2014;11(2):e9027. [PubMed: 25580387]
62. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annu Rev Psychol*. 2007;58:593–614. [PubMed: 16968208]
63. Dai HS, Bao YC. An inverse probability weighted estimator for the bivariate distribution function under right censoring. *Stat Probab Lett*. 2009;79(16):1789–1797.
64. Dai H, Wang H. *Analysis for Time-to-Event Data under Censoring and Truncation*. London, UK: Academic Press; 2016.
65. Hayes AF. Beyond Baron and Kenny: statistical mediation analysis in the new millennium. *Commun Monogr*. 2009;76(4):408–420.
66. Alwin DF, Hauser RM. Decomposition of effects in path analysis. *Am Sociol Rev*. 1975;40(1):37–47.

Figure 1. Mediation models. (A) Conceptual model for the study of the mediating effect of age at menopause on the association between genetic variants and type-2 diabetes risk. (B) Diagram for a single mediator model.

Table 1.

Means and standard errors of estimated coefficients for different paths, a_0 , a , b_0 , b and $\tilde{c}$, obtained using naïve, complete-case and MAC-CC approaches, based on 1,000 replicates, each with 500 cases and 500 controls, under null hypothesis that there is no indirect effect through the mediator. Different scenarios were considered based on different values of a , b and censoring percentage, with $a_0 = 6$, $b_0 = -3$ and $\tilde{c} = 0.5$.

Scenario	a	b	CP	Naïve approach				Complete-case approach				MAC-CC approach						
				a ₀ (se)	a (se)	b ₀ (se)	b (se)	$\tilde{c}$ (se)	a ₀ (se)	a (se)	b ₀ (se)	b (se)	$\tilde{c}$ (se)	a ₀ (se)	a (se)	b ₀ (se)	b (se)	$\tilde{c}$ (se)
1	0	0	21%	6.001 (0.05)	0.000 (0.07)	-0.285 (0.28)	0.000 (0.05)	0.505 (0.14)	5.934 (0.05)	0.000 (0.07)	-0.300 (0.45)	0.002 (0.07)	0.505 (0.15)	5.998 (0.07)	0.002 (0.09)	-3.022 (0.44)	0.002 (0.07)	0.507 (0.13)
2	0	0	40%	5.996 (0.06)	0.004 (0.09)	-0.282 (0.26)	0.000 (0.05)	0.499 (0.13)	5.832 (0.06)	0.003 (0.08)	-0.313 (0.49)	0.005 (0.08)	0.501 (0.17)	5.998 (0.08)	0.000 (0.11)	-3.014 (0.48)	0.002 (0.08)	0.503 (0.13)
3	0.3	0	21%	5.996 (0.05)	0.303 (0.07)	-0.287 (0.27)	0.000 (0.05)	0.506 (0.13)	5.934 (0.06)	0.302 (0.07)	-0.285 (0.43)	0.000 (0.07)	0.508 (0.15)	5.996 (0.07)	0.303 (0.10)	-2.992 (0.41)	-0.002 (0.07)	0.504 (0.13)
4	0.3	0	42%	6.005 (0.06)	0.295 (0.09)	-0.305 (0.25)	0.003 (0.05)	0.511 (0.13)	5.839 (0.06)	0.286 (0.08)	-0.312 (0.51)	0.004 (0.09)	0.517 (0.18)	6.000 (0.08)	0.297 (0.12)	-3.011 (0.47)	0.001 (0.08)	0.505 (0.14)
5	0	0.4	19%	6.028 (0.05)	0.004 (0.07)	-1.315 (0.30)	0.190 (0.05)	0.475 (0.13)	5.969 (0.05)	0.004 (0.07)	-2.668 (0.46)	0.401 (0.08)	0.487 (0.14)	5.999 (0.05)	0.002 (0.07)	-3.014 (0.45)	0.403 (0.07)	0.492 (0.13)
6	0	0.4	40%	6.028 (0.06)	0.002 (0.09)	-0.859 (0.26)	0.117 (0.05)	0.488 (0.13)	5.861 (0.06)	0.002 (0.08)	-2.724 (0.55)	0.408 (0.09)	0.507 (0.16)	5.999 (0.06)	-0.004 (0.09)	-3.037 (0.51)	0.405 (0.08)	0.502 (0.13)

CP: censoring percentage; se: standard error

Table 2.

Means and standard errors of estimated indirect effects (*IE*s), as well as 95% coverage probabilities of confidence intervals for the estimation of *IE*, obtained using naïve, complete-case and MAC-CC approaches, based on 1,000 replicates, each with 500 cases and 500 controls, under null hypothesis that there is no *IE* through the mediator. Different scenarios were considered based on different values of *a*, *b* and censoring percentage, with $a_0 = 6$, $b_0 = -3$ and $\tilde{c} = 0.5$.

Scenario	<i>a</i>	<i>b</i>	Exp <i>IE</i>	<i>CP</i>	Naïve approach		Complete-case approach		MAC-CC approach	
					<i>IE</i> (se)	95% cov	<i>IE</i> (se)	95% cov	<i>IE</i> (se)	95% cov
1	0	0	0.000	21%	0.000 (0.001)	0.99	0.000 (0.001)	0.99	0.000 (0.000)	1.00
2	0	0	0.000	40%	0.000 (0.001)	1.00	0.000 (0.002)	1.00	0.000 (0.001)	1.00
3	0.3	0	0.000	21%	0.000 (0.004)	0.94	0.000 (0.005)	0.95	0.000 (0.002)	0.94
4	0.3	0	0.000	42%	0.000 (0.003)	0.94	0.000 (0.006)	0.94	0.000 (0.002)	0.95
5	0	0.4	0.000	19%	0.000 (0.003)	0.93	0.000 (0.007)	0.93	0.000 (0.007)	0.95
6	0	0.4	0.000	40%	0.000 (0.003)	0.94	0.000 (0.008)	0.93	0.000 (0.008)	0.94

Exp: expected; CP: censoring percentage; se: standard error; cov: coverage probability

Table 3.

Means and standard errors of estimated coefficients for different paths, a_0 , a , b_0 , b and $\tilde{c}$, obtained using naïve, complete-case and MAC-CC approaches, based on 1,000 replicates, each with 500 cases and 500 controls, under alternative hypothesis that there is an indirect effect through the mediator. Different scenarios were considered based on different values of b_0 , b , censoring percentage and prevalence of the outcome, with $a_0 = 6$, $a = 0.3$ and $\tilde{c} = 0.5$.

Scenario	b_0	b	CP	prev	Naïve approach					Complete-case approach					MAC-CC approach				
					a_0 (se)	a (se)	b_0 (se)	b (se)	$\tilde{c}$ (se)	a_0 (se)	a (se)	b_0 (se)	b (se)	$\tilde{c}$ (se)	a_0 (se)	a (se)	b_0 (se)	b (se)	$\tilde{c}$ (se)
1	-6	0.4	21%	4%	6.149 (0.05)	0.354 (0.07)	-1.278 (0.30)	0.160 (0.05)	0.589 (0.13)	6.085 (0.05)	0.353 (0.07)	-2.837 (0.48)	0.400 (0.08)	0.505 (0.15)	5.996 (0.06)	0.307 (0.09)	-5.991 (0.50)	0.396 (0.08)	0.509 (0.13)
2	-5	0.4	21%	10%	6.133 (0.06)	0.340 (0.07)	-1.274 (0.28)	0.163 (0.05)	0.570 (0.13)	6.069 (0.06)	0.340 (0.07)	-2.806 (0.47)	0.399 (0.07)	0.491 (0.15)	5.998 (0.07)	0.305 (0.09)	-4.987 (0.47)	0.397 (0.07)	0.495 (0.13)
3	-4	0.4	21%	23%	6.091 (0.05)	0.320 (0.07)	-1.284 (0.29)	0.168 (0.05)	0.569 (0.13)	6.028 (0.05)	0.319 (0.07)	-2.809 (0.50)	0.404 (0.08)	0.498 (0.14)	5.996 (0.06)	0.305 (0.08)	-4.020 (0.49)	0.402 (0.08)	0.497 (0.13)
4	-5.5	0.3	21%	3%	6.118 (0.05)	0.338 (0.07)	-1.053 (0.29)	0.124 (0.05)	0.569 (0.14)	6.054 (0.05)	0.337 (0.07)	-2.208 (0.48)	0.303 (0.08)	0.505 (0.15)	6.000 (0.07)	0.301 (0.10)	-5.540 (0.48)	0.304 (0.08)	0.509 (0.14)
5	-4.5	0.3	21%	9%	6.106 (0.05)	0.327 (0.07)	-1.064 (0.27)	0.129 (0.05)	0.556 (0.13)	6.042 (0.05)	0.326 (0.07)	-2.220 (0.44)	0.307 (0.07)	0.502 (0.15)	5.999 (0.06)	0.297 (0.09)	-4.533 (0.45)	0.305 (0.07)	0.499 (0.13)
6	-3.5	0.3	21%	21%	6.077 (0.05)	0.312 (0.07)	-1.046 (0.29)	0.128 (0.05)	0.552 (0.13)	6.014 (0.05)	0.311 (0.07)	-2.174 (0.47)	0.304 (0.08)	0.495 (0.15)	6.000 (0.06)	0.296 (0.08)	-3.531 (0.46)	0.304 (0.07)	0.499 (0.13)
7	-6	0.4	42%	4%	6.157 (0.07)	0.348 (0.09)	-0.874 (0.24)	0.099 (0.04)	0.590 (0.13)	5.982 (0.07)	0.337 (0.09)	-2.860 (0.56)	0.406 (0.09)	0.485 (0.17)	6.001 (0.09)	0.302 (0.12)	-5.994 (0.54)	0.396 (0.08)	0.488 (0.14)
8	-5	0.4	42%	10%	6.136 (0.06)	0.331 (0.09)	-0.856 (0.25)	0.097 (0.04)	0.590 (0.13)	5.962 (0.06)	0.323 (0.08)	-2.859 (0.57)	0.407 (0.09)	0.495 (0.18)	6.000 (0.08)	0.293 (0.11)	-5.008 (0.54)	0.398 (0.08)	0.494 (0.14)
9	-4	0.4	42%	23%	6.093 (0.06)	0.317 (0.09)	-0.866 (0.24)	0.101 (0.04)	0.593 (0.12)	5.923 (0.06)	0.309 (0.08)	-2.881 (0.57)	0.415 (0.09)	0.505 (0.17)	5.996 (0.07)	0.300 (0.10)	-4.079 (0.52)	0.410 (0.08)	0.502 (0.13)
10	-5.5	0.3	42%	3%	6.120 (0.07)	0.336 (0.09)	-0.714 (0.23)	0.072 (0.04)	0.567 (0.13)	5.950 (0.06)	0.325 (0.09)	-2.185 (0.54)	0.301 (0.09)	0.492 (0.18)	6.007 (0.08)	0.295 (0.12)	-5.484 (0.51)	0.295 (0.08)	0.494 (0.14)
11	-4.5	0.3	42%	9%	6.105 (0.06)	0.326 (0.09)	-0.724 (0.24)	0.074 (0.04)	0.576 (0.13)	5.934 (0.06)	0.318 (0.08)	-2.210 (0.54)	0.305 (0.09)	0.505 (0.18)	5.999 (0.07)	0.297 (0.10)	-4.523 (0.53)	0.301 (0.08)	0.503 (0.14)
12	-3.5	0.3	42%	21%	6.074 (0.06)	0.319 (0.09)	-0.728 (0.22)	0.078 (0.04)	0.566 (0.13)	5.904 (0.06)	0.312 (0.08)	-2.154 (0.54)	0.302 (0.09)	0.494 (0.17)	5.995 (0.07)	0.308 (0.10)	-3.493 (0.50)	0.298 (0.08)	0.495 (0.13)

CP: censoring percentage; prev: prevalence of the outcome in the general population; se: standard error

Table 4.

Means and standard errors of estimated indirect effects (*IEs*) and proportions mediated (*PMs*), as well as 95% coverage probabilities of confidence intervals for the estimations of *IE* and proportion mediated, obtained using naïve, complete-case and MAC-CC approaches, based on 1,000 replicates, each with 500 cases and 500 controls, under alternative hypothesis that there is an *IE* through the mediator. Different scenarios were considered based on different values of b_0 , b , censoring percentage and prevalence of the outcome, with $a_0 = 6$, $a = 0.3$ and $\tilde{c} = 0.5$.

Scenario	Exp			Naïve approach				Complete-case approach				MAC-CC approach						
	b_0	b	IE	PM	CP	prev	IE (se)	95% cov	PM (se)	95% cov	IE (se)	95% cov	PM (se)	95% cov	IE (se)	95% cov	PM (se)	95% cov
1	-6	0.4	0.006	0.240	21%	4%	0.014 (0.005)	0.606	0.089 (0.036)	0.198	0.034 (0.009)	0.034	0.230 (0.086)	0.962	0.006 (0.002)	0.946	0.243 (0.088)	0.950
2	-5	0.4	0.013	0.232	21%	10%	0.013 (0.005)	0.942	0.090 (0.039)	0.248	0.032 (0.009)	0.344	0.229 (0.088)	0.946	0.013 (0.004)	0.938	0.241 (0.089)	0.966
3	-4	0.4	0.023	0.219	21%	23%	0.013 (0.005)	0.492	0.088 (0.037)	0.276	0.031 (0.009)	0.880	0.215 (0.079)	0.950	0.023 (0.007)	0.966	0.229 (0.082)	0.962
4	-5.5	0.3	0.004	0.190	21%	3%	0.010 (0.005)	0.670	0.071 (0.038)	0.310	0.025 (0.008)	0.120	0.179 (0.079)	0.916	0.004 (0.001)	0.944	0.196 (0.087)	0.938
5	-4.5	0.3	0.009	0.185	21%	9%	0.010 (0.004)	0.948	0.073 (0.035)	0.326	0.024 (0.007)	0.422	0.177 (0.074)	0.958	0.009 (0.003)	0.968	0.191 (0.079)	0.944
6	-3.5	0.3	0.016	0.175	21%	21%	0.010 (0.004)	0.680	0.069 (0.036)	0.358	0.023 (0.008)	0.888	0.171 (0.075)	0.944	0.016 (0.006)	0.938	0.181 (0.075)	0.948
7	-6	0.4	0.006	0.240	42%	4%	0.008 (0.004)	0.906	0.056 (0.030)	0.068	0.033 (0.011)	0.178	0.237 (0.110)	0.944	0.006 (0.002)	0.920	0.245 (0.113)	0.938
8	-5	0.4	0.013	0.232	42%	10%	0.008 (0.004)	0.744	0.053 (0.031)	0.090	0.031 (0.011)	0.546	0.233 (0.133)	0.944	0.012 (0.005)	0.940	0.235 (0.107)	0.930
9	-4	0.4	0.023	0.219	42%	23%	0.008 (0.004)	0.140	0.052 (0.029)	0.076	0.030 (0.010)	0.904	0.219 (0.102)	0.942	0.023 (0.008)	0.938	0.228 (0.093)	0.930
10	-5.5	0.3	0.004	0.190	42%	3%	0.006 (0.004)	0.918	0.042 (0.029)	0.140	0.024 (0.010)	0.318	0.194 (0.245)	0.928	0.004 (0.002)	0.950	0.190 (0.093)	0.946
11	-4.5	0.3	0.009	0.185	42%	9%	0.006 (0.004)	0.868	0.041 (0.029)	0.100	0.024 (0.009)	0.644	0.180 (0.103)	0.936	0.008 (0.004)	0.960	0.188 (0.090)	0.964
12	-3.5	0.3	0.016	0.175	42%	21%	0.006 (0.004)	0.354	0.043 (0.029)	0.166	0.023 (0.009)	0.918	0.180 (0.135)	0.940	0.017 (0.007)	0.962	0.185 (0.083)	0.956

Exp: expected; CP: censoring percentage; cov: coverage probability; prev: prevalence of the outcome in the general population; se: standard error

Table 5.

Estimations of total effects (*TEs*), direct effects (*DEs*), and indirect effects (*IEs*) for the SNPs associated with both type-2 diabetes and a woman's age at menopause in the real data analysis using different approaches. The 95% confidence intervals (CIs) for *IEs* were assessed using a bootstrap approach with 1000 bootstraps.*

CHR	SNP	Naïve approach			Complete-case approach			MAC-CC approach					
		TE	DE	IE	95% CI of IE	TE	DE	IE	95% CI of IE	TE	DE	IE	95% CI of IE
7	rs11771343	0.0456	0.0452	0.0003	[−0.0016,0.0022]	0.0567	0.0563	0.0004	[−0.0019,0.0026]	0.0442	0.0443	−0.0001	[−0.0015,0.0013]
8	rs3924519	−0.0357	−0.0362	0.0006	[−0.0013,0.0022]	−0.0402	−0.0407	0.0005	[−0.0009,0.0020]	−0.0365	−0.0373	0.0008	[−0.0002,0.0032]
13	rs11842863	0.0488	0.0500	−0.0012	[−0.0051,0.0013]	0.0584	0.0598	−0.0014	[−0.0053,0.0012]	0.0456	0.0471	−0.0015	[−0.0053,0.0009]

CHR: chromosome

* Proportions of total effects mediated by the mediator are not reported because it is not meaningful when the direct and indirect effects have opposite directions.