Submitted to the Annals of Applied Statistics

MASH: MEDIATION ANALYSIS OF SURVIVAL OUTCOME AND HIGH-DIMENSIONAL OMICS MEDIATORS WITH APPLICATION TO COMPLEX DISEASES

BY SUNYI CHI[1], CHRISTOPHER R FLOWERS[2], ZIYI LI[1], XUELIN HUANG[1,*] AND PENG WEI[1,†]

[1]*Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA,* [*]*xlhuang@mdanderson.org;* [†]*pwei2@mdanderson.org*

[2]*Department of Lymphoma, The University of Texas MD Anderson Cancer Center, Houston, USA*

Environmental exposures such as cigarette smoking influence health outcomes through intermediate molecular phenotypes, such as the methylome, transcriptome, and metabolome. Mediation analysis is a useful tool for investigating the role of potentially high-dimensional intermediate phenotypes in the relationship between environmental exposures and health outcomes. However, little work has been done on mediation analysis when the mediators are high-dimensional and the outcome is a survival endpoint, and none of it has provided a robust measure of total mediation effect. To this end, we propose an estimation procedure for Mediation Analysis of Survival outcome and High-dimensional omics mediators (MASH) based on a second-moment-based measure of total mediation effect for survival data analogous to the R^2 measure in a linear model. In addition, we propose a three-step mediator selection procedure to mitigate potential bias induced by non-mediators. Extensive simulations showed good performance of MASH in estimating the total mediation effect and identifying true mediators. By applying MASH to the metabolomics data of 1919 subjects in the Framingham Heart Study, we identified five metabolites as mediators of the effect of cigarette smoking on coronary heart disease risk (total mediation effect, 51.1%) and two metabolites as mediators between smoking and risk of cancer (total mediation effect, 50.7%). Application of MASH to a diffuse large B-cell lymphoma genomics data set identified copy-number variations for eight genes as mediators between the baseline International Prognostic Index score and overall survival.

1. Introduction Mediation analysis is performed to elucidate how an exposure influences an outcome through mediating variables. Researchers have used this approach widely in epidemiology and clinical

Keywords and phrases: high-dimensional mediators, mediation analysis, survival analysis, total mediation effect, variable selection.

studies, for example, to investigate the connection between environmental risk factors and disease outcomes via intermediate molecular phenotypes, such as gene expression or protein markers. To date, most mediation studies have focused on a single mediator or a few mediators, with much less attention paid to high-dimensional mediators [Daniel et al. (2015); VanderWeele et al. (2014)], particularly in the analysis of survival data [Huang et al. (2017)]. Mediation analysis of high-dimensional data has garnered increasing interest as a result of rapid technological advancements in high-throughput genomic profiling such as RNA sequencing. Although researchers have recently made progress in high-dimensional mediation analysis of continuous outcomes in the linear model setting [Liu et al. (2021); Sampson et al. (2018)], investigators have devised only a few methods for high-dimensional mediation analysis of survival data owing to the considerable challenges in variable selection, censoring, and estimation of the total mediation effect. Luo et al. (2020) proposed a procedure called HIMA for identifying the putative mediators and estimating the indirect effects of DNA methylation on the pathway from smoking to overall survival among lung cancer patients in the Cancer Genome Atlas (TCGA) lung cancer cohort. HIMA used the classic product-based mediation effect measure that could cancel the component-wise mediation effects in a multiple-mediator model owing to mediation direction disagreement, which is commonly encountered in high-dimensional genomic studies. In fact, real data application results of HIMA showed that the indirect effect by the product measure was negative due to the cancellation of both positive and negative component-wise mediation effects, resulting in that the total effect was less than the direct effect. In addition, the classic product-based and difference-based total mediation effect measures for linear models are only applicable in survival models for rare outcomes under the counterfactual framework [VanderWeele et al. (2011)].

As an alternative to the first-moment-based product effect measure, Fairchild et al. (2009) first proposed to use the R^2-based mediation effect size measure for a single mediator in the linear model setting. Recently, Yang et al. (2021) extended the use of R^2 measure to multiple- or high-dimensional-mediator models for a continuous outcome. Importantly, Yang et al. (2021) showed that the R^2 measure can capture the non-zero total mediation effect in the presence of bi-directional component-wise mediation effects, which are likely ubiquitous in high-dimensional omics data settings. Shi et al. (2022) further extended the R^2 measure to survival outcomes and compared five R^2 measures in a Cox proportional hazards model owing to the non-unique definition of R^2 in survival models. However, to the best of our knowledge, the R^2 measure has yet to be studied with high-dimensional mediator models for survival outcomes, which are common in prospective cohort studies, such as the Framingham Heart Study (FHS), and clinical studies, such as the TCGA project.

To fill this gap in knowledge described above, we propose the MASH method: Mediation Analysis of Survival outcome and High-dimensional omics mediators based on a three-step mediator selection procedure and the second-moment mediation effect size measure for survival models analogous to the R^2 measure in a linear model. In addition to extending the R^2 measure to high-dimensional survival data, we propose the use of a partial R^2 mediation measure to adjust for confounding effects.

To evaluate our mediation effect estimation in terms of bias and variance, as well as mediator selection accuracy, we conducted extensive simulation studies in various settings by varying the censoring rate, the number of putative mediators and the directions of mediation effects. We demonstrated that our approach has better performance than existing methods in terms of bias and variance in mediation effect estimation and is stable to data with high censoring rate. Furthermore, we applied the MASH to the metabolomics data of 1919 subjects in the FHS and a diffuse large B-cell lymphoma (DLBCL) genomics data set to address three problems: how smoking influences the risk of coronary heart disease through metabolites, how smoking influences the risk of cancer through metabolites, and how the baseline International Prognostic Index (IPI) impacts overall survival of DLBCL through copy-number variations.

2. Methods Mediation processes are framed in terms of intermediate variables between an independent variable and a dependent variable, with a minimum of three variables required in total. Let T denote the dependent variable – a survival outcome, X denote the independent variable, an exposure of interest and M denote the set of mediator variables that is supposed to transmit the effect of X to T. Figures 1A and 1B illustrate the relationships among the survival outcome, exposure, and mediators.

In our context in which all the mediators are measured at the same time and a direct causal relationship among them is less likely to exist, we assume that the mediators are not causally related to each other. Although we will propose a method to control for confounding variables in section 2.3, the following assumptions are made as in VanderWeele et al. (2014): (1) no unmeasured confounders between the exposure and the survival outcome, between the mediators and the survival outcome or between the exposure and the mediators, and (2) no exposure-induced confounding between the mediators and the survival outcome.

2.1. *Review of Mediation Models* We use the following Cox proportional hazards regression models to assess the role of mediators M in the pathway from exposure X to survival outcome T:

$$\lambda(t|X_i) = \lambda_0^X(t)exp(cX_i), \tag{1}$$

$$\lambda(t|X_i, M_i) = \lambda_0^{XM}(t)exp(rX_i + \sum_{j=1}^{p} b_j M_{ij}), \tag{2}$$

$$\lambda(t|M_i) = \lambda_0^{M}(t)exp(\sum_{j=1}^{p} d_j M_{ij}), \tag{3}$$

$$M_{ij} = a_j X_i + \epsilon_{ij}, \tag{4}$$

where $M_i = (M_{i1}, M_{i2}, \ldots, M_{ip})'$ is the p-dimensional mediator vector for subject $i = 1, 2, 3, ..., n$ and, without loss of generality, X and M_j are standardized to have mean 0 and variance 1. Let D denote the time to event and C denote the censoring time. The observed survival outcome is $T_i = min(D_i, C_i)$, and the failure indicator is $\delta_i = I(D_i \leq C_i)$ for $i = 1, 2, 3, ..., n$. Equations (1), (2), and (3) are Cox proportional hazards models describing the relationship (1) between T and X, (2) between T, X, and M, and (3) between T and M, respectively. $\lambda_0^X(t), \lambda_0^{XM}(t)$, and $\lambda_0^M(t)$ are their respective baseline hazard functions. In a Cox proportional hazards regression model, the hazard ratio is usually used as a measure of an independent variable's effect on a dependent variable. In equations (1), (2), and (3), c is the parameter relating the exposure to the outcome; r is the direct effect parameter relating the exposure to the outcome; b, defined as $(b_1, \ldots, b_p)^T$, is the parameter vector relating the mediators to the outcome with adjustment for the effect of the exposure; and $d = (d_1, \ldots, d_p)'$, is the parameter vector relating the mediators to the outcome. Equation (4) characterizes how exposure influences the mediators, where $a = (a_1, \ldots, a_p)'$ is the parameter vector relating the exposure to the mediators and residual $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{ip}) \sim MVN(0, E_{p\times p})$, for $i = 1, \ldots, n$. To formally define the causal effect for Cox model in counterfactual framework [Robins and Greenland (1992); Pearl (2001)], we follow the framework of VanderWeele et al. (2011) and Huang et al. (2017) using difference in log-hazard scale. Let $\lambda^{x,M(x)}(t)$ denote the hazard function of $T_{x,M(x)}$ when X was set to x and M was set to the value it would have taken if X was set to x. Let $\lambda^{x^*,M(x^*)}(t)$ denote the hazard function of $T_{x^*,M(x^*)}$ when X was set to x^* and M was set to the value it would have taken if X was set to x^*. Thus, the natural direct effect (NDE) is defined as the difference between $log\lambda^{x^*,M(x)}(t)$ and $log\lambda^{x,M(x)}(t)$, and the natural indirect effect (NIE) is defined as the difference between $log\lambda^{x^*,M(x^*)}(t)$ and $log\lambda^{x^*,M(x)}(t)$. The total effect (TE) of X on T can then be decomposed into two components: a direct effect of X on T (NDE) and an indirect effect of X on T through M (NIE). Under the rare outcome assumption, the NDE approximate to $(x^* - x)r$, the NIE approximate to $(x^* - x)\sum_{j=1}^{p} a_j b_j$, and the TE approximate to $(x^* - x)(r + \sum_{j=1}^{p} a_j b_j)$. Without loss of generality, we set $x^* = 1$ and $x = 0$. Therefore,

$NDE = r$, $NIE = \sum_{j=1}^{p} a_j b_j$, and $TE = r + \sum_{j=1}^{p} a_j b_j$.

2.2. *Measures of mediation effects* The standard difference measure (i.e., $c - r$) and product measure (i.e., $\sum_{j=1}^{p} a_j b_j$) for the indirect effect coincide for a continuous outcome with c, r, a_j, b_j similarly defined in a linear model as in models (1), (2) and (3) (MacKinnon (2008)). They are also comparable based on simulations in single-mediator analysis without considering censoring and ties in log survival time models and log hazard time models, but not in Cox proportional hazards models [Tein and MacKinnon (2003)]. In single-mediator Cox model, the difference method uses $c - r$ as a measure of the indirect effect and the product method uses ab. VanderWeele et al. (2011) showed that if all of the models are correctly specified and the outcome is rare, these two are approximately equal in single-mediator Cox model. However, neither the product measure nor the difference measure for the proportional hazards model has any sort of clear causal interpretation as a measure of mediation effect VanderWeele et al. (2011). In multiple-mediator Cox models, the difference and product measures are, respectively, defined as $c - r$ and $\sum a_j b_j$, which are approximately equal only under rare events/high censoring rates (VanderWeele et al. (2011),Shi et al. (2022)). Furthermore, in high-dimensional mediation analysis the component-wise mediation effects ($a_j b_j$'s) could cancel out due to the likely presence of opposite mediation directions, resulting in close to zero or negative product measure. However, existing mediation analysis methods for multiple mediators and survival outcomes still use the product measure. Luo et al. (2020) proposed a high-dimensional mediation analysis framework for survival outcomes using the product measure that led to negative product measure and the the estimated total effect less than the direct effect in their real data applications. On the other hand, the R^2 mediation effect measure, which was originally proposed in a single-mediator model by Fairchild et al. (2009), does not suffer from the issue of misleading cancellation when the mediation effects have both positive and negative components. Also, it has good performance with complex structured data, such as highly correlated mediators (Yang et al. (2021)). Therefore, here we propose a explained-variation measure ($R^2_{mediated}$) based on second moments to estimate mediation effects of high-dimensional mediators for survival outcomes. This measure does not require rare event assumption and has more robust estimation in multi-mediator analysis than the standard product measure as to be shown later on.

$R^2_{mediated}$, the measure of mediation effect, is defined by the amount of variation in outcome T that is explained by exposure X through mediators M. In Cox proportional hazards models, this measure is com-

puted as follows:

$$R^2_{mediated} = R^2_{T,M} + R^2_{T,X} - R^2_{T,MX}, \tag{5}$$

where $R^2_{T,M}$ is the proportion of variation in outcome T that is explained by the mediators, $R^2_{T,X}$ is the variation in T that is explained by the exposure, and $R^2_{T,MX}$ is the proportion of variation in T that is explained by both the mediators and the exposure. Thus, $R^2_{T,X}$ quantifies the total effect whereas $R^2_{T,MX} - R^2_{T,M}$ is the direct effect. And the R^2 measure of indirect effect is the total effect minus the direct effect, which is given in equation (5). The shared over simple effect (SOS), defined as $\frac{R^2_{mediated}}{R^2_{T,X}}$ by Lindenberger and Pötter (1998) is the standardized exposure-related variation in the outcome that is shared with the mediators.

Although R^2 is clearly defined in linear regression models, applying the notion of explained variation to survival data is not straightforward because of nonnormality of the errors and censoring of the dependent variable. There have been many versions of R^2-like measures proposed for survival data. Royston (2006) outlined the properties that a good explained variation measure for survival models should have: 1) approximate independence of the amount of censoring; 2) reduction to (or a close relationship with) the R^2 usually obtained via "equivalent" linear regression analysis of the same data set, if possible; 3) the nesting property for two models: M1 $\subset$ M2 (with $\subset$ denoting nesting) and $R^2(M1) \leq R^2(M2)$; 4) increase in R^2 as the strength of the association increases; and 5) availability of confidence intervals. Shi et al. (2022) compared five measures of explained variation for survival data in mediation models and suggested the use of R^2_w by Kent and O'Quigley (1988) in multiple-mediator analysis in the Cox proportional hazards models. Specifically, R^2_w is defined as $\frac{var(Z\beta)}{1+var(Z\beta)}$, where Z is the vector of independent variables and β is the coefficient vector in a Cox model. Plugging R^2_w in equation (5) results in the $R^2_{mediated}$ for our proposed high-dimensional mediation analysis for survival outcomes.

R^2_w is an approximation of explained variation used to measure the information gain [Kent and O'Quigley (1988)]. It is compatible with the R^2 measure in linear regression models and satisfies the properties outlined by Shi et al. (2022). Moreover, Schemper and Stare (1996) noted that R^2_w is unaffected by censoring and has good performance across simulation settings comparing measures of explained variation in survival analysis. Importantly, R^2_w is robust to use in our proposed high-dimensional mediation analysis framework based on our simulation studies and real data applications.

$R^2_{mediated}$ can be shown as a function of a, b, c, d, and r as follows:

$$R^2_{mediated} = \frac{(d^2)^T(a^2) + d^T E d}{1 + (d^2)^T(a^2) + d^T E d} + \frac{c^2}{1 + c^2} - \frac{(r + b^T a)^2 + b^T E b}{1 + (r + b^T a)^2 + b^T E b}, \tag{6}$$

where $d^2 = (d_1^2, \ldots, d_p^2)$ and $a^2 = (a_1^2, \ldots, a_p^2)$. The derivation of equation (6) is presented in the supplementary materials Section 2.

In the scenario of all a's and b's equal to 0, both product measure and $R^2_{mediated}$ are equal to 0. However, the product measure could be equal to 0 when individual mediator effects exist. Even when none of a_j's or b_j's is zero, the product measure $b^T a$ could be zero or close to zero due to cancellation of positive and negative $a_j b_j$'s. On the other hand, $R^2_{mediated}$ doesn't suffer from this problem. However, as demonstrated in Supplemental Materials Section 2, $R^2_{mediated}$ can be non-zero when there is no true mediator (i.e., both a_j and b_j are non-zero) but certain types of non-mediators exist (i.e., either a_j or b_j is zero but not both). To mitigate the potential bias induced by those non-mediators, we propose a mediator selection procedure as introduced in the next section.

2.3. *New method: MASH* To perform mediation analysis of high-dimensional omics mediators such as metabolomics and genome-wide copy-number alterations, we must select the true mediators in the pathway from the exposure to the outcome (Fig. 1B). Let $\mathbf{S_0}$ denote a set of potential mediators, $\mathbf{M}$ denote a set of true mediators, $\mathbf{U}$ denote a set of variables associated with the outcome but not the exposure (Fig. 1C), and $\mathbf{V}$ denote a set of variables associated with the exposure but not the outcome (Fig. 1D). None of the $\mathbf{U}$ or $\mathbf{V}$ variables are true mediators. We emphasize that although inclusion of U in mediation analysis for continuous outcomes does not affect the estimation of the R^2 mediation effect [Yang et al. (2021)], our derivation in Supplementary Material Section 2 and results of our simulation study shown in Supplementary Material Table S4 demonstrated that inclusion of either U or V could bias the estimation of the R^2 mediation effect on survival outcomes. Therefore, we propose a three-step mediator selection procedure as illustrated in Figure 2 to identify true mediators between the exposure and outcome and mitigate potential bias induced by non-mediators by excluding $\mathbf{U}$, $\mathbf{V}$, and noise variables (i.e., those not associated with the exposure or the outcome).

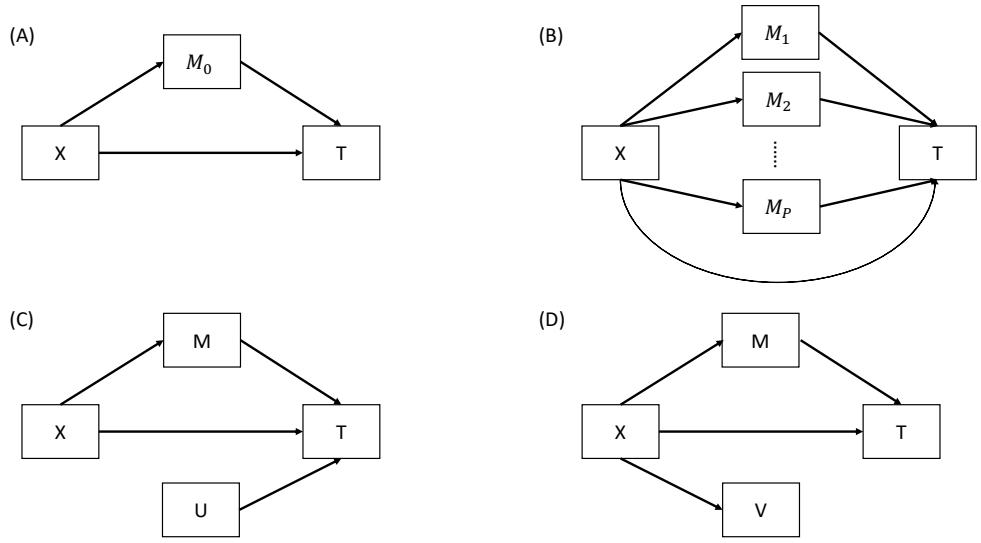

FIG 1. **Diagrams for potential mediation settings.** X is the independent variable, T is the survival outcome, and M is a set of true mediator. U is a set of variables associated with Y but not X, and V is a set of variables associated with X but not Y. (A) A single-mediator model. (B) A multiple-mediator model. (C) A model with non-mediator U. (D) A model with non-mediator V.

Traditional statistical methods fail when the number of mediators p is much larger than the sample size n. Sure independence screening is a large-scale screening method for an ultra-high dimensional feature space with the property that all the true variables survive after variable screening with the probability tending to 1 [Fan and Lv (2008)]. Principled sure independence screening by Zhao and Li (2012) is an extension of sure independence screening for censored survival data using the Cox proportional hazards model, which resembles the marginal ranking method proposed by Fan et al. (2010). This method avoids the requirement of choosing a size of subset to retain after screening by specifying the desired false-positive rate. Therefore, as shown in Figure 2, with MASH, principled sure independence screening based on Cox proportional hazards model (2) is first used to reduce the dimension and exclude V-type non-mediators and noise variables. In this step, a subset of potential mediators S_1 is selected under a fixed false-positive rate of 0.05; those are the mediators likely to be associated with the outcome. Variable selection within the subset S_1 via the minimax concave penalty (MCP) is then conducted to further exclude putative mediators not associated with the outcome based on the Cox proportional hazards model (2). Ten-fold cross validation for MCP-based

survival models is used over a grid of values for the regularization parameter λ. Next, the false discovery rate (FDR) adjusted p value based on the Benjamini-Hochberg procedure with a critical value of 0.2 for the model in equation (4) is used to test the marginal association of each selected potential mediator with the exposure to exclude U-type non-mediators and noise variables. Finally, the mediation effect $R^2_{mediated}$ and SOS are estimated in the estimation dataset using the selected mediators in the variable selection dataset. Of note, as shown in Figure 2, the data are split into two halves with one half used as a variable selection set to select the true mediators and the other half used as an independent dataset to estimate the mediation. The sure independence screening in step (a) is only applied for ultra-high dimensional data, for example, simulation settings 1-16 as demonstrated in Section 3. We skipped step (a) in the metabolomics data applications in Section 4, where n was around 2000 and p was around 200.

2.4. *Partial R^2 to control confounding effects* As illustrated by VanderWeele (2016) failure to control for exposure-outcome, exposure-mediator and mediator-outcome confounding effects in mediation analysis can substantially bias estimates of mediation effects, whereas adjusting for measured confounders as covariates in mediation models can reduce the bias. Therefore, we propose a partial R^2 measure to adjust for potential confounders in mediation analysis. Let Z denote baseline covariates such as age, which may act as a potential confounder. Our models adjusting for confounders are as follows:

$$\lambda_i(t|X_i, Z_i) = \lambda_0^X(t)exp(cX_i + fZ_i), \tag{7}$$

$$\lambda_i(t|X_i, M_i, Z_i) = \lambda_0^{XM}(t)exp(rX_i + gZ_i + \sum_{j=1}^{p} b_j M_{ij}), \tag{8}$$

$$\lambda_i(t|M_i, Z_i) = \lambda_0^M(t)exp(\sum_{j=1}^{p} d_j M_{ij} + hZ_i), \tag{9}$$

$$M_{ij} = a_j X_i + l_j Z_i + \epsilon_{ij}, \tag{10}$$

In addition to controlling for mediator-outcome confounding effects, we also propose models adjusting for exposure-outcome confounders, which include equations (4), (7), (8), (9), and (11) as follows:

$$Z_{ij} = l_j X_i + \epsilon^*_{ij}. \tag{11}$$

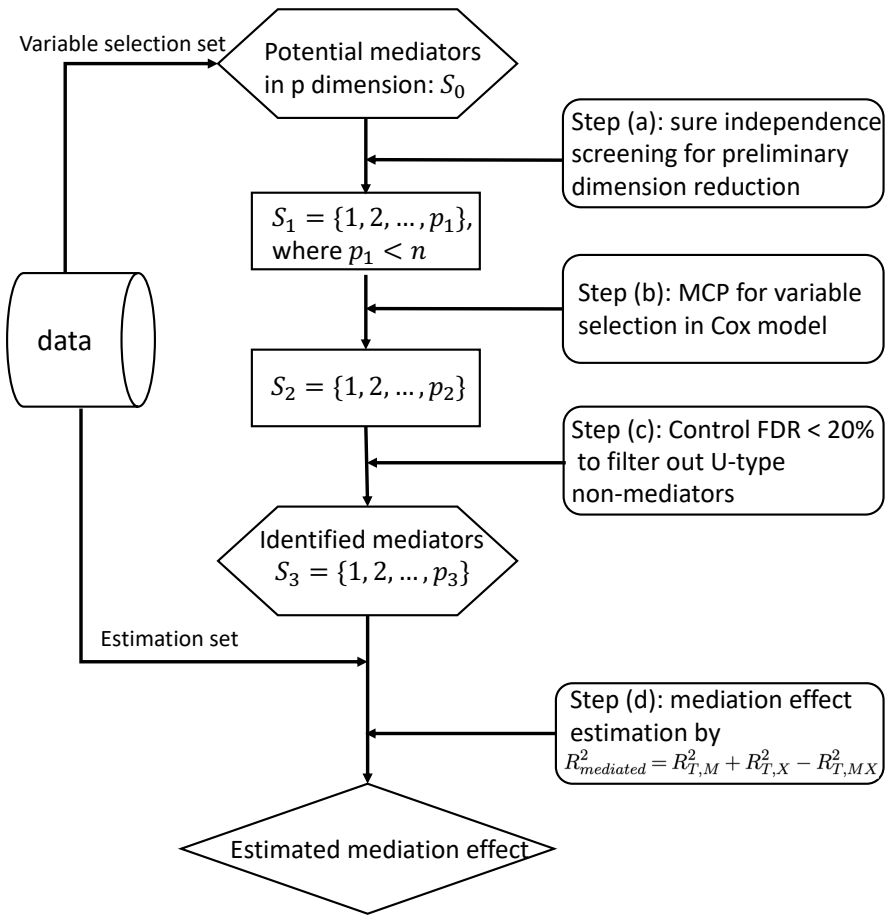

FIG 2. **Overall workflow of MASH.** The workflow consists of four main steps: (a) sure independence screening for preliminary screening, (b) MCP for variable selection, (c) FDR control at 0.2 level for variable selection, and (d) mediation effect estimation using R^2-like measurements.

In general, partial R^2 is defined as the proportion of variation in the outcome variable that cannot be explained in a reduced model but can be explained by the predictors specified in a full model. To adjust for the covariates in the indirect effect of exposure on the outcome through mediators, partial R^2 is adapted based on our original formulas to calculate the mediation effect as shown below.

$$R^2_{mediated|Z} = R^2_{T,M|Z} + R^2_{T,X|Z} - R^2_{T,MX|Z}, \tag{12}$$

$$SOS = \frac{R^2_{mediated|Z}}{R^2_{T,X|Z}}, \tag{13}$$

where, according to the definition of partial R^2, $R^2_{T,M|Z}$, $R^2_{T,X|Z}$, and $R^2_{T,MX|Z}$ are

$$R^2_{T,M|Z} = \frac{R^2_{T,MZ} - R^2_{T,Z}}{1 - R^2_{T,Z}}, \tag{14}$$

$$R^2_{T,X|Z} = \frac{R^2_{T,XZ} - R^2_{T,Z}}{1 - R^2_{T,Z}}, \tag{15}$$

$$R^2_{T,MX|Z} = \frac{R^2_{T,MXZ} - R^2_{T,Z}}{1 - R^2_{T,Z}}. \tag{16}$$

Therefore, $R^2_{mediated|z}$ can be interpreted as the variation of the outcome variable that can be explained by the exposure variable through mediators adjusting for confounder(s) Z.

3. Results In this section, we performed a series of simulation studies to assess the performance of the proposed MASH R^2 mediation analysis procedure for time-to-event outcome in comparison with alternative procedures. We evaluated the MASH in two aspects—the estimated total mediation effect value and the true mediator selection and compared it with the state-of-art HIMA method [Luo et al. (2020)].

3.1. *Simulation Design* We generated data using equations (2) and (4) . For subject $i = 1, ..., n$, the exposure X was generated from the standard normal distribution $N(0, 1)$, where the true mediators were generated using equation (4) and noise mediators were generated from $N(0, 1)$. The error term ϵ_{ij} was generated from $N(0, 1)$. We assumed a common form of baseline hazard $\lambda_0(t) = \lambda t^{\eta-1}$ following a Weibull distribution with $\lambda = 2$ and $\eta = -5$. The censoring time was generated from $U(0, c_0)$ with a constant c_0, which was chosen to control the censoring rate. Furthermore, we conducted simulation studies in 17 settings, mimicking real data and varying the censoring rate, coefficients of mediators in true models, number of true mediators, and number of potential mediators. For each simulation setting, we generated 500 replicates, and the sample size n was 2000 to mimic our real data application to the FHS. To evaluate the performance of MASH in estimating the mediation effect, we calculated pseudo-true mediation effect measures with the known true mediators and mediation models. The pseudo-true R^2 mediation effect is not the exact true value because the true parameters c and d are unknown, so the R^2 of models (2) and (4) ($R^2_{T,X}$ and $R^2_{T,M}$) are estimated values, which may introduce bias and variance in model fitting. For fair comparison, pseudo-true value for ab mediation effect is also calculated as the benchmark for HIMA. To circumvent the undue

influence of model fitting uncertainty, we calculated the pseudo-true mediation effect measures using a very large sample size (n = 200,000).

We performed simulation studies in each setting with censoring rates of 10%, 30%, and 50%. We found that the value of the estimated mediation effect increased with the censoring rate. Current censoring methods in simulation studies in the literature typically use a uniform or exponential distribution to generate the censoring time and adjust the parameters of these distributions to control the censoring rate. With a varying censoring rate, data will not have the same effective sample size. Therefore, comparing the results using censored data with those using uncensored data is unfair. Thus, we used a pseudo-true R^2 mediation effect with comparable censoring rates (10%, 30%, and 50%). The bias for our estimated R^2 mediation effect (shown in Fig. 3) increased very slightly as the censoring rate increased.

3.2. *Simulation results* To explore the performance of the MASH procedure in different scenarios, we conducted simulation studies in settings with different numbers of true mediators (m = 1, 5, and 10) and a fixed number of potential mediators (p = 1000), in settings with ultra-high-dimensional data (p = 1000, 5000, and 10,000, and n = 2000) and a fixed number of true mediators (m = 10), and in settings with different values of coefficients a and b when the numbers of true mediators m and potential mediators p were fixed. Simulation results for all settings are shown in the Supplementary Material Table S1-S7. The results demonstrated that our estimation of the mediation effect using MASH was very stable and performed well in terms of bias, variance and variable selection across the different numbers of true mediators m and a fixed p when we varied the values of parameters a and b and number of potential mediators while keeping m and p fixed. With ultra-high dimensional data (settings with p = 1000, 5000, and 10,000), MASH still performed well as shown in Tables S2, S3 and S6. With a small sample size (n = 300 and 500), the performance of our method could be biased as shown in Table S7. However, the bias is still very small if n = 500 and censoring rate is less or equal to 30%. Therefore, we would suggest using our method with reasonably large samples to avoid bias. More precisely, we need to consider the effective sample size for survival analysis, which is the number of non-censored subjects.

Figure 3 shows boxplots of the estimated total mediation effect of MASH and HIMA in two settings with n = 2000, m = 5, and p = 1000. The coefficients in the two settings were different. In setting 1, a = 0.4, and the coefficients of mediators b was 0.25 for all five true mediators. When the directions of component-wise mediation effects differ across multiple mediators, a potential problem is that the effects in different directions could cancel each other out in the product measure of the total mediation effect. Therefore, Figure 3 also shows results of setting 8 with five mediators that had different component-wise mediation effects

directions with coefficients b = (-1, -1, 0.5, 0.5, 0.5). Figures 3A and 3C show MASH results, and Figure 3B and 3D shows HIMA results. In comparing MASH and HIMA in setting 1 (Fig. 3A and 3B), setting 8 (Fig. 3C and 3D) and all other settings described in the supplementary materials, we found that estimating the total mediation effect using MASH produced a smaller relative bias and smaller variance than did that using HIMA. Also, MASH was more stable than using the HIMA as the censoring rate increased as the product measure is only approximately true for a survival outcome in rare disease setting[VanderWeele et al. (2011)]. In addition, the bias in estimating b in Cox model increases as censoring rate increases. Therefore, the bias of the total indirect effect estimation in summation of ab further increases as the summation of censoring effects. However, R^2 is an overall measure of explained variation using three Cox models, which is less dependent on individual parameter estimation accuracy and little affected by censoring [Schemper and Stare (1996)]. For setting 8, the total indirect effects using HIMA were negative, with the conflict directions of the component-wise mediation effects canceled out in the product measure when the total effect of exposure on outcome was positive. Overall, MASH outperformed HIMA in high-dimensional mediation analysis for survival outcomes across varying simulation settings.

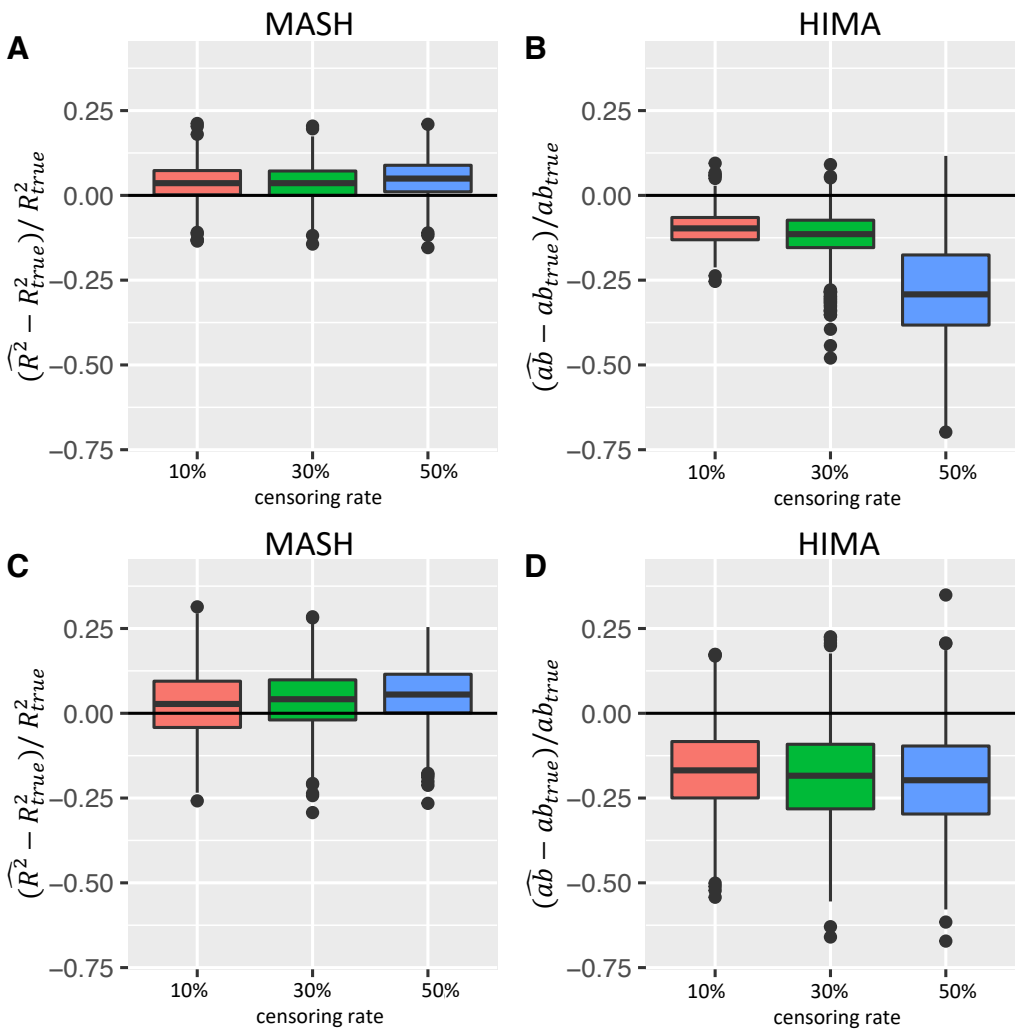

FIG 3. **Boxplots of the relative bias across simulation replications of MASH and HIMA.** (Simulation setting: n = 2000, m = 5, p = 1000; a_j = 0.4, b_j = 0.25 for $j = 1, 2, ..., 5$ in setting 1 and a_j = 0.65 for $j = 1, 2, ..., 5$, b = (-1, -1, 0.5, 0.5, 0.5) in setting 8.) The x-axis corresponds to the censoring rate at (10%, 30%, 50%); Y-axis corresponds relative bias across simulation replications. (A and B) mediation effects of all true mediators were in the same direction (setting 1); (C and D) mediation effects of true mediators were in different directions (setting 8). MASH results are shown in A and C, and HIMA results are shown in B and D.

To evaluate our proposed partial R^2 method adjusting for confounders, we performed a simulation study using equations (9) and (11) as data-generating models. We used a sample size of 2000 and p = 1000 potential mediators, 5 of which were true mediators. We also used censoring rates of 10%, 30%, and 50% and $U(0, 1)$ as the censoring time distribution. We set the parameters a = 0.38, b = 0.5, r = 2.5, and l = 0.1 for all true mediators. The mediator-outcome confounder Z and noise mediators were generated by the

standard normal distribution. We compared our proposed partial R^2 method with the unadjusted R^2 method in this simulation setting, varying the coefficients of covariate g (0.5, 0.75, and 1). The simulation results shown in Supplementary Material Table S5 demonstrated that the partial R^2 measure performed well in estimating the mediation effect controlling for confounding effects.

The existence of the U- or V-type non-mediators U and V in identified mediators set S_3 could bias our estimation of the total mediation effect; thus, we considered two settings where potential mediators set S_0 included ten non-mediators U or V in addition to noise variables. The results are shown in the Supplementary Material Table S4. We used a sample size of 2000 and 1000 potential mediators, 10 of which were the true mediators. We still used censoring rates of 10%, 30%, and 50% and $U(0,1)$ as the censoring time distribution. We also set the parameters a = 1, b = 0.5, and r = 2.5 for the true mediator. We set the parameters a = 1 and b = 0 for 10 non-mediators V and a = 0 and b = 0.5 for 10 non-mediators U. The results demonstrated that our proposed MASH consistently performs well in estimating the total mediation effect. In contrast, we observed substantial bias in the estimated $R^2_{mediated}$ without screening non-mediators, underscoring the importance of mediator selection prior to estimating $R^2_{mediated}$.

In addition to the mediation effect estimation, we assessed the accuracy of mediator variable selection using MASH. The true positive rate (percentage of true mediators correctly selected) and false positive rate (percentage of non-mediators incorrectly selected) for all settings are presented in Supplementary Material Table S6. The true positive rates were greater than 99% among all settings, and the highest false positive rate for all settings was 0.7%. Therefore, the performance of MASH is satisfactory in terms of selecting true mediators and filtering out non-mediators (controlling the false positive rate).

4. Real Data Applications To demonstrate the proposed MASH method's versatility with different omics mediators and study designs, we applied it to the metabolomics data as mediators in the prospective cohort study FHS with time to coronary heart disease (CHD) and smoking-related cancer as the outcomes, as well as genome-wide copy-number variations as mediators with overall survival as the outcome in a cancer genomics dataset.

4.1. *Application of MASH to the FHS* The FHS is a long-term cohort study of cardiovascular disease that has been ongoing since the 1940s and includes three generations: the Original Cohort, the Offspring Cohort, and the Third Generation Cohort [Splansky et al. (2007)]. In our application of MASH to the FHS, we used the metabolomics data of 1919 individuals in the Offspring Cohort from whom blood samples were

collected at Exam 5. We were interested in exploring the roles of metabolites in the pathway from smoking to the risk of CHD or cancer.

4.1.1. *CHD* CHD is the leading cause of death globally. Smoking is a major risk factor for CHD risk [HHS (2010)]. Although researchers have studied many metabolites and showed that they are significantly associated with the risk of incident CHD [e.g. Wang et al. (2019); Cavus et al. (2019); Li et al. (2017)], their mediation role in the smoking-CHD relationship has yet to be elucidated. Therefore, we applied MASH to the plasma-based metabolomics data from individuals in the FHS Offspring Cohort. After excluding subjects with more than 20% missing values in smoking status or metabolite data, we applied the k-nearest neighbors (KNN) algorithm [Altman (1992)] to impute the remaining missing data. After preprocessing, we had 1902 individuals, 307 of whom were diagnosed with CHD during follow-up after Exam 5 when plasma was collected for metabolomics profiling, with 190 metabolites as potential mediators. The exposure was tobacco smoking status at Exam 5 as a binary variable (never/ever). The outcome was the number of days from Exam 5 to CHD diagnosis or the last follow-up visit. The median number of follow-up days was 8672.5. We evaluated the effect of smoking on CHD risk as mediated by metabolites, adjusting for age, gender and body mass index (BMI). We randomly split the data into 50% as variable selection set and 50% as estimation set as illustrated in Fig. 2. We used the first half to select mediators and used the second half to estimate the mediation effect. The number of mediators p was smaller than the sample size n in the FHS data set, so we skipped the preliminary principled sure independence screening step here. We applied the MCP in Figure 2 step (b) to exclude noise variables and V-type non-mediators, and controlled the FDR at 20% in step (c) to filter out U-type non-mediators. We also applied the HIMA method as comparison.
The results are presented in Table 1. We identified five metabolites as mediators of the effect of smoking on CHD risk including cotinine, carnitine, C48:0 triacylglycerol, C20:4 cholesteryl ester, and monosaccharides. The total mediation effect of the metabolites was 0.017 (SE = 0.016) by the $R^2_{mediated}$ measure. Because the values of $R^2_{mediated}$ measures depend on the total effect, we focused on more interpretable SOS measures in real data applications, which are standardized by the total effect and can be interpreted as the percentage of the total effect that is explained by the mediators. We estimated that 51.1% (SE = 0.215) of the total effect was mediated by metabolites based on the SOS measure. We also evaluated the mediation effect of each selected mediators in single-mediator models and presented their mediation effects as estimated using the partial R^2 method, adjusting for age, gender and BMI. In our single-mediator model, we estimated that cotinine, carnitine, C48:0 triacylglycerol, C20:4 cholesteryl ester, and monosaccharides

explained 21.5%, 20.3%, 16.1%, 15.5%, and 9.6% of the total effect, respectively, by the SOS measure.

TABLE 1

Application of proposed MASH to FHS data: smoking, metabolites and CHD risk in comparison with HIMA

MASH	$R^2_{mediated}$	SOS	HIMA	ab	Indirect/Total effect
Multiple-mediator model					
All	0.017	0.511	All	0.083	0.047
Single-mediator models					
Cotinine	0.016	0.215	C48:0 TAG	0.022	0.012
Carnitine	0.007	0.203	Carnitine	0.020	0.012
C48:0 TAG	0.005	0.161	Histidine	0.014	0.008
C20:4 CE	0.005	0.155	Creatinine	0.012	0.007
monosaccharides	0.003	0.096	C34:2 DAG	-0.013	-0.007

The sample size was 1902 including 307 subjects who were diagnosed with CHD after Exam 5. The number of metabolites (potential mediators) was 190, where 5 metabolites were selected as true mediators by MASH. The outcome was time to CHD diagnosis censored by the last follow-up visit time. Age, gender and BMI were considered confounding variables. We estimated the total mediation effect using the multiple-mediator model and the single-mediator effect of identified mediators using the single-mediator model. The five mediators were selected in the multiple-mediator model, and their estimated mediation effect in single-mediator model was larger than 0.001 by $R^2_{mediated}$. CE, cholesteryl ester; DAG, diacylglycerols;TAG, triacylglycerol.

Cotinine is an alkaloid found in tobacco and is the predominant metabolite of nicotine. It is a highly sensitive and specific marker of active and passive exposure to tobacco [Benowitz (1996)]. Researchers reported similarity in the relative risk of CHD among individuals with substantial exposure to passive smoking and those with light active exposure to it [Whincup et al. (2004)]. Also, studies have demonstrated that the ability of urine cotinine to improve assessment of cardiovascular disease risk assessment is similar to that provided by self-reported smoking status [Kunutsor et al. (2018)]. Our results further showed that cotinine is a mediator in the pathway from smoking to the risk of CHD.

Previous studies also found that L-carnitine can effectively improve the cardiac function rating and thus improve cardiac function [Zhao et al. (2020)] and was associated with a 27% reduction in the all-cause mortality rate, a 65% reduction in ventricular arrhythmias, and a 40% reduction in anginal symptoms in patients experiencing acute myocardial infarction [DiNicolantonio et al. (2013)]. Carnitine is also a metabolite found at higher levels in current smokers than in non-smokers [Xu et al. (2013)]. Our results are consistent with the results of these studies and revealed the important mediation role of carnitine in the impact of smoking on the risk of CHD.

Cholesteryl ester present in human plasma is positively correlated with high-density lipoprotein (HDL) levels [Subbaiah et al. (2012)]. Many clinical and epidemiological studies have clearly demonstrated that the HDL cholesterol level is inversely associated with the risk of CHD and is a critical and independent

component of predicting this risk [Wang et al. (2017); Mundra et al. (2018)]. In addition, elevated plasma triacylglycerol concentrations have been associated with increased risk of CHD and associated with other CHD risk factors, namely, reduced HDL cholesterol concentrations and increased low-density lipoprotein (LDL) particles [Roche and Gibney (2000)]. Previous studies suggested that smoking is associated with total cholesterol, low-density lipoprotein cholesterol, and triglyceride levels [Gossett et al. (2009)]. Our mediation analysis results provide further understanding of how these risk factors work together to contribute to the incidence of CHD.

The three most common monosaccharides are glucose, fructose, and galactose. Sugars are naturally components of the tobacco leaf and are also commonly added to cigarettes by tobacco companies. Glucose and fructose are the most abundant natural sugars in dried tobacco leaves Clarke et al. (2006). Many studies found that increased intake of sugar and blood sugar level were associated with an increased CHD risk Howard et al. (2002). We confirmed the mediating role of monosaccharides in the effect of cigarette smoking on CHD risk using MASH.

As shown in Table 1, the indirect effect estimated using HIMA was 0.083, and the ratio of the indirect effect to the total effect was 0.047. HIMA selected 14 mediators, 3 of which were selected by MASH: carnitine, C48:0 TAG and monosaccharides. The 11 other mediators selected by HIMA were histidine, creatinine, adma, aminoadipate, amp, salicylurate, C36:1pc, C16:1sm, C34:2dag, C52:3tag, and C58:12tag. We observed conflicting directions of the indirect effect in product measure, which caused cancellation of the estimated total indirect effect by HIMA.

4.1.2. *Smoking-Related Cancer* Smoking can cause cancer and, further, prevent the body from fighting it [HHS (2020a)]. Specifically, it can cause cancer at almost all of the most common sites, including the bladder, colon, kidney, lung, and pancreas [HHS (2020b)]. Many studies have suggested that metabolites are important biomarkers associated with the risk of cancer [Mazzilli et al. (2020)]. In addition, smoking is associated with many metabolites, such as cotinine, O-cresol sulfate and hydroxycotinine [Cross et al. (2014)]. The risk of colon, lung, pancreatic, urinary bladder, and kidney cancer was shown to be associated with cigarette smoking in many previous studies [ACS (2021); CDC (2020)]. Therefore, the mediation role of metabolites in the relationship between tobacco smoking and cancer is of great interest. To examine this, we applied the MASH method to the FHS Offspring Cohort with smoking as the exposure, metabolomics at Exam 5 as the potential mediators and incident smoking-related cancer during the follow-up after Exam 5 as the outcome, adjusting for age, gender and BMI as covariates. We were left with 1919 individuals and 190 metabolites after removing subjects and metabolites with more than 20% missing values. One hundred

eighty-five individuals in the data set were diagnosed with the aforementioned smoking-related cancer during follow-up. As illustrated in Figure 2, we randomly spit the data set into two halves for variable selection and estimation, respectively. We applied the MCP to exclude noise variables and V-type non-mediators. We controlled the FDR at 20% in Figure 2 step (c) to filter out U-type non-mediators.

The results are presented in Table 2. We identified two metabolites that mediated the effect of smoking on the risk of cancer: cotinine and glycodeoxycholic acid. The total mediation effect of the metabolites was 0.055 (SE = 0.026) by the $R^2_{mediated}$ measure. We estimated that 50.7% (SE = 0.184) of the total effect was mediated by the two metabolites according to the SOS measure. In our single-mediator models, we estimated that cotinine explained 46.8% of the total effect, and glycodeoxycholic acid explained 4.2% of the total effect by the SOS measure. Previous studies demonstrated that cotinine, as a biomarker of current smoking status, was associated with the incidence of bladder cancer [Thong et al. (2016)], lung cancer [Ellard et al. (1995)], colorectal cancer [Cross et al. (2014)], and other smoking-related cancers. Furthermore, it was reported that glycodeoxycholic acid, a conjugated bile acid, stimulated tumor growth [Dai et al. (2013)], and a recent study suggested that it promotes colon cancer development [Kühn et al. (2020)]. On the other hand, HIMA selected 0 mediator, which means that the estimated mediation effects of metabolites in the pathway from smoking to cancer using HIMA was 0. Therefore, MASH produced more reasonable results of mediation effect estimation than HIMA in the applications of real data with high censoring rates.

TABLE 2
Application of proposed MASH to FHS data: smoking, metabolites and cancer risk

Mediators	$R^2_{mediated}$	SOS
Multiple-mediator model		
All	0.055	0.507
Single-mediator model		
Cotinine	0.051	0.468
Glycodeoxycholic	0.005	0.042

The sample size was 1919 including 185 individuals who were diagnosed with smoking-related cancer after Exam 5. The number of metabolites (potential mediators) was 190, where 2 metabolites were selected as true mediators by MASH. The outcome was time to cancer diagnosis censored by the last follow-up visit time. Age, gender and BMI were considered as confounding variables. We estimated total mediation effect using the multiple-mediator model and the single-mediator effect of identified mediators using the single-mediator model. The two mediators were selected in the multiple-mediator model, and their estimated mediation effect in the single-mediator model was larger than 0.001 by $R^2_{mediated}$.

4.2. *Prognosis for DLBCL* Diffuse large B-cell lymphoma (DLBCL) is the most common hematological malignancy and is characterized by a striking degree of genetic and clinical heterogeneity [Reddy et al.

(2017); Xu et al. (2022)]. The International Prognostic Index (IPI), a clinical scoring system developed by oncologists to aid in predicting the prognosis for lymphoma, assigns one point to each negative prognostic factor (age > 60 years, serum lactate dehydrogenase level above the upper limit of normal, Ann Arbor stage III/IV disease, Eastern Cooperative Oncology Group performance status 2, and more than one site with extranodal involvement) and categorizes patients into four risk groups based on the total score: 0/1 = low risk, 2 = low-intermediate risk, 3 = high-intermediate risk, and 4/5 = high risk. Researchers developed the IPI at a time when DLBCL patients received chemotherapy-only regimens, and had estimated 5-year overall survival rates ranging from 26% to 73% depending on the risk category [INHLPFP (1993)]. We were interested in determining how the clinical exposure (IPI score) affects the overall survival of DLBCL patients through copy-number alterations. Reddy et al. (2017) published a study of 1001 DLBCL patients with copy-number alteration profiling of 140 genes known to be associated with cancer development and prognosis. We used 754 observations in this data set for the present study, after removing 247 patients with missing IPI scores or outcomes (overall survival in years). The censoring rate was 0.683. In our application of MASH, we randomly spit the data set into two halves for variable selection and estimation. In mediator selection, we applied the MCP to exclude noise variables and V-type non-mediators and controlled the FDR at 20% in Figure 2 step (c) to filter out U-type non-mediators.

Our results shown in Table 3 demonstrated that the total mediation effect of the copy-number alterations was 0.063 (SE = 0.041) by the $R^2_{mediated}$ measure. And we estimated that the total mediation effect of the copy-number alterations explained 19.7% (SE = 0.130) of the total effect by the SOS measure. We identified eight genes that mediated the effect of IPI score on the overall survival using our MASH procedure. In our single-mediator models, we estimated that $FOXP1, POU2F2, ANKRD17, LIN54, CD70, DNMT3A, S1PR2$, and $CD79B$ explained 8.3%, 7.4%, 6.0%, 4.0%, 3.3%, 1.9%, 0.6% and 0.6% of the total effect by the SOS measure, respectively. In another study, investigators identified $ANKRD17$ [Reddy et al. (2017)], $POU2F2$ [Reddy et al. (2017)], $CD79B$ [Reddy et al. (2017)], and $S1PR2$ [Baldari et al. (2016)] as driver genes for DLBCL. Researchers also found that the genes $DNMT3A$ [Reddy et al. (2017)], $CD70$ [Reddy et al. (2017)], $CD79B$ [Schmitz et al. (2018)], and $FOXP1$ [Barrans et al. (2004)] are prognostic factors for DLBCL overall survival. $LIN54$, a protein coding gene, is a component of the LIN, or $DREAM$ complex, which is an essential regulator of cell cycle genes known to be associated with ovarian cancer. While looking into the gene expression and copy number variation of $LIN54$, we found that they were both borderline significant for overall survival of BLDCL patients (Supplemental materials Section 4).

Lastly, HIMA selected 0 mediator, which means that the estimated mediation effect of copy number vari-

ation in the pathway from clinical IPI score to DLBCL using HIMA was 0. Therefore, MASH produced more reasonable results of mediation effect estimation than HIMA and outperformed HIMA in the three applications of real data with high censoring rates.

TABLE 3

Application of proposed MASH to DLBCL prognosis: IPI score, copy-number alterations and overall survival of DLBCL patients

Mediators	$R^2_{mediated}$	SOS
Multiple-mediator model		
All	0.063	0.197
Single-mediator model		
FOXP1	0.027	0.083
POU2F2	0.024	0.074
ANKRD17	0.019	0.060
LIN54	0.013	0.040
CD70	0.011	0.033
DNMT3A	0.006	0.019
S1PR2	0.002	0.006
CD79B	0.002	0.006

The sample size was 754 and censoring rate was 68.3%. The number of genes of copy-number alterations (potential mediators) for each patient was 140, where 8 genes were selected as true mediators by MASH. The outcome was overall survival. We estimated total mediation effect using the multiple-mediator model and the single-mediator effect of identified mediators using the single-mediator model. The eight mediators were selected in the multiple-mediator model, and their estimated mediation effect in the single-mediator model was larger than 0.001.

5. Discussion We have developed MASH, a novel method of mediation analysis, for high-dimensional mediators and time-to-event outcomes to estimate the total mediation effect and identify mediators. MASH critically extends the existing R^2-based mediation analysis for continuous outcomes to survival outcomes in the high-dimensional setting. Our approach incorporates sure independence screening, MCP variable selection, and false discovery rate control for high-dimensional mediator variable selection and R^2-based measures for mediation effect estimation. It has good performance in mediation effect estimation and mediator selection, which we have shown via simulation studies and multiple real data applications. We would like to remind the readers that including U- or V-type non-mediators in estimating $R^2_{mediated}$ could result in substantial bias as illustrated by Supplemental Materials Table S4, while our proposed mediator selection procedure can effectively mitigate the bias.

Extending the R^2 measure defined in linear regression to survival models is not straight-forward owing to the nonunique definition of R^2 in the latter. Based on our simulations, the R^2_w-based measure should be chosen for survival models because the influence of censoring on it is less than that on other R^2 measures. Besides the Cox proportional hazards model, we have also explored an accelerated failure time (AFT) model for analysis of right censored data and found that the Cox model is more stable in R^2 measure estimation

with a wide range of censoring rates.

Many questions of interest about mediation analysis for high-dimensional survival data remain to be addressed in the future. Mediation analysis of multiple types of high-dimensional omics mediators such as gene expression data in addition to metabolomics data, is biologically interesting, yet challenging for selecting mediators from multi-omics data with potentially complex dependence structures. Although we have proposed using partial R^2 measure for MASH to control confounding in mediation analysis, models allowing for exposure-mediator interactions warrant further investigation in the future.

Acknowledgements This work was supported by the National Institutes of Health (NIH) grants R01HL116720. P.W. was partially supported by NIH grant P50CA217674 and Cancer Prevention and Research Institute of Texas (CPRIT) Grant RP230166. X.H. was partially supported by NIH grant U54CA096300, U01CA152958 and P50CA100632 and the Dr. Mien-Chie Hung and Mrs. Kinglan Hung Endowed Professorship. The authors acknowledge the Texas Advanced Computing Center at The University of Texas at Austin for providing HPC resources. The authors declare that there are no conflicts of interest. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This article was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. The manuscript was edited by Don Norwood and Sarah Bronson, ELS, of Editing Services, Research Medical Library at The University of Texas MD Anderson Cancer Center. The authors thank two anonymous reviewers, an Associate Editor and Editor Dr. Beth Ann Griffin for their many constructive comments, which helped improve the manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "MASH: Mediation analysis of survival outcome and high-dimensional omics mediators with application to complex disease."

The supplementary material contains the complete results of our simulation study discussed in the text.

REFERENCES

ALTMAN, NAOMI S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46** 175–185.

AMERICAN CANCER SOCIETY (2021). Cancer Prevention & Early Detection Facts & Figures 2021-2022. Atlanta, Ga: American Cancer Society.

BALDARI, C. T. (2016). S1PR2 deficiency in DLBCL: a FOXy connection. *Blood* **127(11)** 1380–1381.

BARRANS, S. L., FENTON, J. A., BANHAM, A., OWEN, R. G. and JACK, A. S. (2004). Strong expression of FOXP1 identifies a distinct subset of diffuse large B-cell lymphoma (DLBCL) patients with poor outcome. *Blood* **104** 2933–2935.

BENOWITZ, N. L. (1996). Cotinine as a biomarker of environmental tobacco smoke exposure. *Epidemiol Rev.* **18** 188-204.

CAVUS, E., KARAKAS, M., OJEDA, F. M., KONTTO, J., VERONESI, G., FERRARIO, M. M. et al. (2019). Association of Circulating Metabolites With Risk of Coronary Heart Disease in a European Population: Results From the Biomarkers for Cardiovascular Risk Assessment in Europe (BiomarCaRE) Consortium. *JAMA cardiology* **4(12)** 1270–1279.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2020). Health Effects of Cigarette Smoking.

CHI, S., FLOWERS, C.,LI, Z., HUANG, X. and WEI, P. (2021). Supplement to "MASH: Mediation analysis of survival outcome and high-dimensional omics mediators with application to complex diseases."

CLARKE, M. B., BEZABEH, D. Z. and HOWARD, C. T. (2006). Determination of carbohydrates in tobacco products by liquid chromatography-mass spectrometry/mass spectrometry: a comparison with ion chromatography and application to product discrimination. *Journal of agricultural and food chemistry* **54(6)** 1975–1981.

CROSS, A. J., BOCA, S., FREEDMAN, N. D., CAPORASO, N. E., HUANG, W. Y., SINHA, R., SAMPSON, J. N. and MOORE, S. C. (2014). Metabolites of tobacco smoking and colorectal cancer risk. *Carcinogenesis* **35(7)** 1516–1522.

DAI, J., WANG, H., DONG, Y., ZHANG, Y. and WANG, J. (2013). Bile acids affect the growth of human cholangiocarcinoma via NF-kB pathway. *Cancer investigation* **31(2)** 111–120.

DANIEL, R. M., DE STAVOLA, B. L., COUSENS, S. N., VANSTEELANDT S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* **71(1)** 1–14.

DINICOLANTONIO, J. J., LAVIE, C. J., FARES, H., MENEZES, A. R. and O'KEEFE, J. H. (2013). L-carnitine in the secondary prevention of cardiovascular disease: systematic review and meta-analysis. *Mayo Clinic proceedings* **88(6)** 544–551.

ELLARD, G. A., DE WAARD, F. and KEMMEREN, J. M. (1995). Urinary nicotine metabolite excretion and lung cancer risk in a female cohort. *British journal of cancer* **72(3)** 788–791.

FAIRCHILD, A.J., MACKINNON, D.P., TABORGA, M.P., TAYLOR, A.B. (2009). R2 effect-size measures for mediation analysis. *Behavior Research Methods* **41(2)** 486–498

FAN, J., LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society* **70(5)** 849–911.

FAN, J., FENG, Y., WU, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. *Institute of Mathemat ical Statistics Collections* 70-86.

GOSSETT, L. K., JOHNSON, H. M., PIPER, M. E., FIORE, M. C., BAKER, T. B., STEIN, J. H. (2009). Smoking intensity and lipoprotein abnormalities in active smokers. *Journal of Clinical Lipidology* **3** 372–378.

HELLER, G. (2012). A measure of explained risk in the proportional hazards model. *Biostatistics* **13(2)** 315-325.

HOWARD, B. V. and WYLIE-ROSETT, J. (2002). Sugar and cardiovascular disease: A statement for healthcare professionals from the Committee on Nutrition of the Council on Nutrition, Physical Activity, and Metabolism of the American Heart Association. *Circulation* **106(4)** 523–527.

HUANG, Y. T., YANG, H. I. (2017). Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology* **28(3)** 370–8.

INTERNATIONAL NON-HODGKIN'S LYMPHOMA PROGNOSTIC FACTORS PROJECT (INHLPFP) (1993). A predictive model for aggressive non-Hodgkin's lymphoma. *N Engl J Med.* **329(14)** 987-994.

KENT, J. T., and O'QUIGLEY, J. (1988). Measures of Dependence for Censored Survival Data. *Biometrika* **75(3)** 525–4.

KÜHN, T., STEPIEN, M., LÓPEZ-NOGUEROLES, M., DAMMS-MACHADO, A., SOOKTHAI, D., JOHNSON, T., ROCA, M., HÜSING, A., MALDONADO, S. G., CROSS, A. J., MURPHY, N., FREISLING, H., RINALDI, S., SCALBERT, A., FEDIRKO, V., SEVERI, G., BOUTRON-RUAULT, M. C., MANCINI, F. R., SOWAH, S. A., BOEING, H., ... KAAKS, R. (2020). Prediagnostic Plasma Bile Acid Levels and Colon Cancer Risk: A Prospective Study. *Journal of the National Cancer Institute* **112(5)** 516–524.

KUNUTSOR, S. K., SPEE, J. M., KIENEKER, L. M., GANSEVOORT, R. T., DULLAART, R. P. F., VOERMAN, A. J., TOUW, D. J., BAKKER, S. J. L. (2018). Self-Reported Smoking, Urine Cotinine, and Risk of Cardiovascular Disease: Findings From the PREVEND (Prevention of Renal and Vascular End-Stage Disease) Prospective Cohort Study. *J Am Heart Assoc.* **7(10)** e008726.

LI, Y., ZHANG, D., HE, Y. et al. (2017). Investigation of novel metabolites potentially involved in the pathogenesis of coronary heart disease using a UHPLC-QTOF/MS-based metabolomics approach. *Sci Rep.* **7** 15357.

LINDENBERGER, U. and PÖTTER, U. (1998). The complex nature of unique and shared effects in hierarchical linear regression: Implications for developmental psychology. *Psychological Methods* **3(2)** 218–230.

LIU, Z., SHEN, J., BARFIELD, R., SCHWARTZ, J., BACCARELLI, A., LIN, X. (2021). Large-Scale Hypothesis Testing for Causal Mediation Effects with Applications in Genome-wide Epigenetic Studies. *Journal of the American Statistical Association.* DOI: https://doi.org/10.1080/01621459.2021.1914634

LUO, C., FA, B., YAN, Y., WANG, Y., ZHOU, Y., ZHANG, Y., and YU, Z. (2020). High-dimensional mediation analysis in survival models. *PLoS computational biology* **16(4)** e1007768.

MAZZILLI, K. M., MCCLAIN, K. M., LIPWORTH, L., PLAYDON, M. C., SAMPSON, J. N., CLISH, C. B., GERSZTEN, R. E., FREEDMAN, N. D. and MOORE, S. C. (2020). Identification of 102 Correlations between Serum Metabolites and Habitual Diet in a Metabolomics Study of the Prostate, Lung, Colorectal, and Ovarian Cancer Trial. *The Journal of nutrition* **150(4)** 694–703.

MUNDRA, P. A., BARLOW, C. K., NESTEL, P. J., BARNES, E. H., KIRBY, A., THOMPSON, P., SULLIVAN, D. R., ALSHEHRY, Z. H., MELLETT, N. A., HUYNH, K., JAYAWARDANA, K. S., GILES, C., MCCONVILLE, M. J., ZOUNGAS, S., HILLIS, G. S., CHALMERS, J., WOODWARD, M., WONG, G., KINGWELL, B. A., SIMES, J., ... LIPID Study Investigators (2018). Large-scale plasma lipidomic profiling identifies lipids that predict cardiovascular events in secondary prevention. *JCI insight* **3(17)** e121326.

PEARL, J. (2001). Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (pp. 411– 420). San Francisco, CA: Morgan Kaufmann

REDDY, A., ZHANG, J., DAVIS, N. S., MOFFITT, A. B., LOVE, C. L., WALDROP, A., LEPPA, S., PASANEN, A., MERIRANTA, L., KARJALAINEN-LINDSBERG, M. L., NØRGAARD, P., PEDERSEN, M., GANG, A. O., HØGDALL, E., HEAVICAN, T. B., LONE, W., IQBAL, J., QIN, Q., LI, G., KIM, S. Y., ... DAVE, S. S. (2017). Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171** 481–494.e15.

ROBINS, J. M., GREENLAND, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology* **3(2)** 143–155.

ROCHE, H. M., GIBNEY, M. J. (2000). Effect of long-chain n-3 polyunsaturated fatty acids on fasting and postprandial triacylglycerol metabolism. *Am J Clin Nutr.* **71(1)** 232S-7S.

ROYSTON, P. (2006). Explained Variation for Survival Models. *The Stata Journal* **6(1)** 83-96.

SAMPSON, J.N., BOCA, S.M., MOORE, S.C., HELLER, R. (2018). FWER and FDR control when testing multiple mediators. *Bioinformatics.* **34(14)** 2418–24.

SCHMITZ, R., WRIGHT, G. W., HUANG, D. W., ... STAUDT, L. (2018). Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N Engl J Med.* **378(15)** 1396-1407.

SHI, B., HUANG, X., WEI, P. (2022). Comparison of Effect Size Measures for Mediation Analysis of Survival Outcomes with Application to the Framingham Heart Study. *arXiv: 2205.03303.*

SPLANSKY, G. L., COREY, D., YANG, Q., ATWOOD, L. D., CUPPLES, L. A., BENJAMIN, E. J., D'AGOSTINO, R. B., SR, FOX, C. S., LARSON, M. G., MURABITO, J. M., O'DONNELL, C. J., VASAN, R. S., WOLF, P. A. and LEVY, D. (2007). The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *American journal of epidemiology* **165** 1328–1335.

SCHEMPER, M., STARE, J. (1996). Explained variation in survival analysis. *Statistics in medicine* **15(19)** 1999–2012.

SUBBAIAH, P. V., JIANG, X. C., BELIKOVA, N.A., AIZEZI, B., HUANG, Z. H., REARDON, C.A. (2012). Regulation of plasma cholesterol esterification by sphingomyelin: effect of physiological variations of plasma sphingomyelin on lecithin-cholesterol acyltransferase activity. *Biochim Biophys Acta.* **1821(6)** 908-913.

TEIN, J.-Y. and MACKINNON, D. P. (2003). Estimating Mediated Effects with Survival Data. In: Yanai H, Rikkyo AO, Shigemasu K, Kano Y and Meulman JJ (eds) *New Developments on Psychometrics* (pp. 405-412). Tokyo, Japan: Springer-Verlag Tokyo Inc.

THONG, A. E., PETRUZELLA, S., ORLOW, I., ZABOR, E. C., EHDAIE, B., OSTROFF, J. S., BOCHNER, B. H. and BARNES, H. F. (2016). Accuracy of Self-reported Smoking Exposure Among Bladder Cancer Patients Undergoing Surveillance at a Tertiary Referral Center. *European urology focus* **2(4)** 441–444.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2010). How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2020). A Report of the Surgeon General. How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2010 [accessed 2020 January 27].

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2020). The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014 [accessed 2020 January 27].

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

VANDERWEELE, T. J., VANSTEELANDT, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods* **2(1)** 95–115.

VANDERWEELE, T. J. (2016). Mediation Analysis: A Practitioner's Guide. *Annu Rev Public Health* **37** 17-32.

WANG, H. H., GARRUTI, G., LIU, M., PORTINCASA, P. and WANG, D. Q. (2017). Cholesterol and Lipoprotein Metabolism and Atherosclerosis: Recent Advances In reverse Cholesterol Transport. *Annals of hepatology* **16** s27–s42.

WANG, Z., ZHU, C., NAMBI, V., MORRISON, A.C., FOLSOM, A.R., BALLANTYNE, C.M., BOERWINKLE, E., YU, B. (2019). Metabolomic Pattern Predicts Incident Coronary Heart Disease. *Arterioscler Thromb Vasc Biol.* **39(7)** 1475-1482.

WHINCUP, P. H., GILG, J. A., EMBERSON, J. R., JARVIS, M. J., FEYERABEND, C., BRYANT, A., WALKER, M. and COOK, D. G. (2004). Passive smoking and risk of coronary heart disease and stroke: prospective study with cotinine measurement. *BMJ (Clinical research ed.)* **329(7459)** 200–205.

XU, T., HOLZAPFEL, C., DONG, X. et al. (2013) Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC Med.* **11** 60.

XU, P. P., HUO, Y. J. and ZHAO, W. L. (2022). All roads lead to targeted diffuse large B-cell lymphoma approaches. *Cancer cell* **40(2)** 131–133.

YANG, T., NIU, J., CHEN, H., WEI, P. (2021). Estimation of Mediation Effect for High-dimensional Omics Mediators. *BMC Bioinformatics* **22** 414.

ZHANG C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38(2)** 894-942.

ZHAO, G., ZHANG, H., WANG, Y., GAO, X., LIU, H. and LIU, W. (2020). Effects of levocarnitine on cardiac function, urinary albumin, hs-CRP, BNP, and troponin in patients with coronary heart disease and heart failure. *Hellenic journal of cardiology : HJC = Hellenike kardiologike epitheorese* **61(2)** 99–102.

ZHAO, S. D., LI Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J Multivar Anal.* **105(1)** 397-411.

MACKINNON, D. P. (2008). Introduction to statistical mediation analysis. Taylor Francis Group/Lawrence Erlbaum Associates.