Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Mediation analysis using incomplete information from publicly available data sources

Andriy Derkach*[1] | Elizabeth D. Kantor[1] | Joshua N Sampson[2] | Ruth M Pfeiffer[2]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 485 Lexington Ave, New York, NY 10017, US
[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20850, USA

Correspondence
* Email: derkacha@mskcc.org

Summary

Our work was motivated by the question whether, and to what extent, well-established risk factors mediate the racial disparity observed for colorectal cancer (CRC) incidence in the United States (US). Mediation analysis examines the relationships between an exposure, mediator and an outcome. All available methods require access to a single complete data set with these three variables. However, because population-based studies usually include few non-White participants, these approaches have limited utility in answering our motivating question. Recently, we developed novel methods to integrate several data sets with incomplete information for mediation analysis. These methods have two limitations: (i) they only consider a single mediator, and (ii) they require a data set containing individual-level data on the mediator and exposure (and possibly confounders) obtained by independent and identically distributed sampling from the target population. Here, we propose a new method for mediation analysis with several different data sets that accommodates complex survey and registry data, and allows for multiple mediators. The proposed approach yields unbiased causal effects estimates and confidence intervals with nominal coverage in simulations. We apply our method to data from US cancer registries, a US-population-representative survey and summary level odds-ratio estimates, to rigorously evaluate what proportion of the difference in CRC risk between non-Hispanic Whites and Blacks is mediated by three potentially modifiable risk factors (CRC screening history, body mass index, regular aspirin use).

KEYWORDS:
Survey sampling, data integration, direct and indirect effects, summary level information, registry data

1 | INTRODUCTION

Mediation analysis is a popular statistical tool for understanding the causal relationship between an exposure, a mediator, and an outcome. It is widely applicable in many disciplines, including epidemiology, clinical research, psychology, economics, and neuroscience. Traditional parametric mediation models are based on an exposure, a single mediating variable and the outcome, all measured in the same data set[1,2]. Modern mediation analysis has recently adopted a causal counterfactual framework to evaluate mediating effects without strong parametric assumptions[3,4,5,6]. In parallel, advanced statistical methods for mediation analysis have been proposed for high-dimensional settings to identify the true set of mediators and to describe the underlying mediation model[7,8,9,10,11,5]. Mediation methods have also been extended to accommodate non-linear models with various types of outcomes, as well as discrete and ordinal mediators[11,12,13]. However, all these methods have the critical limitation that they require the outcome, mediator and exposure to be measured on the same individuals. It is increasingly desirable in many fields of research, including epidemiological studies, to utilize combinations of data sets, with various levels of information[14,15,16], in part due to the dramatic increase in data availability. In the setting of mediation analysis for epidemiologic investigations, there are frequently two types of data: 1) a potentially large data set with exposures and outcomes derived from one or more observational studies (e.g. case-control or cohort studies from large consortia) or from population-based surveys; and 2) data sets that contain detailed information linking exposures and putative mediators (e.g., health care databases and well-designed prospective epidemiological studies). There are an increasing number of statistical methods for analyzing and combining data sets with partial information, but these are not suited for mediation analysis with summary level data[17,18], and approaches that can handle summary level data assume perfect mediation, i.e., the entire effect of an exposure on outcome is through the mediators[19]. Therefore, we recently developed novel methods to perform mediation analysis for a single mediator with several "incomplete" data sets, i.e. data sets containing partial information, and not a single one containing all primary variables[20].

Here we build upon these methods for mediation analysis with several different data sets, each with partial information, to accommodate survey and registry data, and multiple mediators. Our work was motivated by the question whether, and to what extent, well-established risk factors mediate the racial disparity observed for colorectal cancer (CRC) incidence in the United States (US). Racial disparities in CRC incidence and mortality are widely documented[21,22,23] Multiple factors associated with CRC risk have been consistently identified, including family history of CRC, high body mass index (BMI), a history of cigarette smoking, regular aspirin use, and CRC screening and polyp history among others[24,25,26]. However, because population-based studies usually include only few participants from racial minority groups, there is limited evidence on the extent to which differences in the prevalence of these factors explain racial differences in CRC incidence.

To answer the motivating question, we propose a new method for mediation analysis with several different data sets that accommodates data from US cancer registries, a US population-representative survey and summary level odds ratio estimates of selected CRC risk or protective factors from a matched population-based case-control study. We build on the theoretical results of our earlier methods[20], which assumed that all individual level data sets contained information on identically and independently sampled study participants from the same target population, and only considered a single mediator. We use the same theoretical framework to prove identifiability of all regression model parameters under the new setting. We derive a new set of estimating equations that incorporate sampling weights, and we prove consistency and asymptotic normality of the corresponding estimates. We propose two new approaches for a variance calculation that accommodate the complex survey sampling design of one of the data sets. The first approach derives the joint asymptotic variance of the estimates of the regression parameters and empirical distribution function of the covariates, mediators and exposures while accommodating complex survey sampling. The second approach incorporates uncertainty in the estimate of the empirical distribution though a bootstrap approach. We then use these methods to rigorously evaluate what proportion of the difference in the probabilities of developing CRC over a given time period between non-Hispanic Whites and Blacks is mediated by three potentially modifiable risk factors (CRC screening history, body mass index (BMI), and regular aspirin use).

After a brief overview, we describe the data used for the analysis, present the statistical model and methods for estimating the parameters needed for the mediation analysis from incomplete data sources (Section 2). In several simulations we demonstrate that the new method provides unbiased estimates of the mediation statistics and analytic standard errors that are close to their empirical values (Section 3). We then analyze the CRC data example and give estimates of the percentages of mediated effects for various age-groups for US Whites and Blacks (Section 4). Lastly, we conclude with a short discussion in Section 5.

2 METHODS

We present our statistical methodology in the context of the motivating example, namely estimating the mediating effects of well-established risk factors on the racial disparity observed for colorectal cancer (CRC) incidence. Our novel statistical method combines age-specific disease incidence rates, race-specific distributions of three modifiable risk factors (the mediators) and their association estimates with disease risk (odds ratios, ORs) to estimate adjusted age-specific differences in the probabilities of disease development between two racial groups, and the percent of difference mediated by select risk factors. Specifically, for our CRC example we combine the following information: 1) estimates of sex- and age-specific CRC incidence rates for Whites and Blacks using the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) cancer registry database; 2) estimates of the prevalence of the mediating factors (CRC screening history; BMI, ≤ 24.9, $25-30$, or $> 30kg/m^2$;

regular aspirin use) for both racial groups obtained from the US National Health Interview Survey (NHIS); and 3) ORs and their covariance matrices for the mediating factors estimated from non-Hispanic Whites enrolled in a population-based age-matched case-control study[24].

2.1 | Statistical Methods

Let Y denote a binary indicator that is 1 if a person is diagnosed with CRC within one year from baseline, E the exposure indicator that is 1 if a subject is non-Hispanic White, and 0 if the subject is non-Hispanic Black and let $\boldsymbol{M} = \left(M_1, \ldots, M_4\right)$ denote the mediators that include CRC screening history in the last ten years, BMI ($\leq$ 24.9 kg/m^2, 25-30 kg/m^2, or >30 kg/m^2, with $\leq$ 24.9 kg/m^2 used as the reference category), and aspirin use (regular user and nonuser; nonusers are the reference category). Covariates are denoted by $\boldsymbol{C} = \left(C_1, C_2, C_3\right)$ and include the age categories coded with dummy variables ($40 - 49, 50 - 59, 60 - 69, \geq 70$ years; the ≥ 70 age group is the reference group).

Model for cancer registry data

To evaluate racial disparities in the CRC incidence rates in the SEER population, where we only observe race ($\boldsymbol{E}$) and age ($\boldsymbol{C}$), we assume the following logistic regression with Blacks as the reference category ($E = 0$)

$$P(Y = 1 \,|\, E, \boldsymbol{C}) = \frac{\exp(\beta_0^* + \beta^* E + \sum_i \alpha_i^* C_i + \zeta^* C_1 E)}{1 + \exp(\beta_0^* + \beta^* E + \sum_i \alpha_i^* C_i + \zeta^* C_1 E)}. \tag{2.1}$$

We also included an interaction term $C_1 E$ to capture the nonlinear impact of age, noticeable in the age group 40-49 years, and denote the vector of all parameter estimates in model (2.1) by $\hat{\boldsymbol{\eta}}^* = (\hat{\beta}_0^*, \hat{\beta}^*, \hat{\boldsymbol{\alpha}}^*, \hat{\zeta}^*)$. The mediators $\boldsymbol{M}$ were not measured in the SEER data set and thus are not included in the model (2.1). Partial results for the estimates in model (2.1) are shown in Table 5 (2nd column of the table).

Model for the summary odds ratios

The summary ORs for White men and women[24] estimated the association of the mediators with CRC risk in a population based case-control study. As controls were matched to cases based on age (used in four categories) and study center, which is a nuisance parameter, conditional logistic regression was used to estimate ORs. Here, we assume that the following logistic model was postulated

$$P(Y = 1 \,|\, E = 1, \boldsymbol{M}, \boldsymbol{C}) = \frac{\exp(\beta_0^{**} + \sum_i \alpha_i^{**} C_i + \sum_j \gamma_j^{**} M_j)}{1 + \exp(\beta_0^{**} + \sum_i \alpha_i^{**} C_i + \sum_j \gamma_j^{**} M_j)}. \tag{2.2}$$

where both β_0^{**}, the intercept, and α_i^{**} ($i = 1, 2, 3$), the ORs corresponding to the age groups cannot be estimated from a conditional logistic regression model and thus $\boldsymbol{\alpha}^{**}$ is not provided to us. We denote the vector of all parameters in model (2.2) by $\boldsymbol{\eta}^{**} = (\beta_0^{**}, \boldsymbol{\gamma}^{**}, \boldsymbol{\alpha}^{**})$. Partial results for the estimates in model (2.2) are shown in the 3rd column of Table 5.

Population level outcome model

We assume the following relationship between the outcome Y, the exposure E, mediators $\boldsymbol{M}$, and additional covariates $\boldsymbol{C}$:

$$P(Y = 1 \,|\, E, \boldsymbol{M}, \boldsymbol{C}) = \frac{\exp(\beta_0 + \beta E + \sum_i \alpha_i C_i + \sum_j \gamma_j M_j + \zeta C_1 E)}{1 + \exp(\beta_0 + \beta E + \sum_i \alpha_i C_i + \sum_j \gamma_j M_j + \zeta C_1 E)}. \tag{2.3}$$

This model does not incorporate any interactions between mediators and exposure or other covariates, based on a lack of evidence of such interactions for CRC in the literature[24]. Incorporating interactions between risk factors and race is discussed in Section 2.2. We let $\boldsymbol{\eta} = (\beta_0, \beta, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \zeta)$ denote all parameters in (2.3).

Combining information from the three different data sources to estimate the mediation parameters

We now propose a new approach to estimate $\boldsymbol{\eta} = (\beta_0, \beta, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \zeta)$, the parameters in model (2.3) using $\hat{\boldsymbol{\eta}}^* = (\hat{\beta}_0^*, \hat{\beta}^*, \hat{\boldsymbol{\alpha}}^*, \hat{\zeta}^*)$ and $\hat{\boldsymbol{\gamma}}^{**}$, the reported association estimates from models (2.1) and (2.2) respectively, and data on $F(E, \boldsymbol{M}, \boldsymbol{C})$, the joint distribution of mediators, covariates and race obtained from the NHIS survey. Here, we extend and build on our previous work[20], where we provided theoretical justifications for the identifiability of $\boldsymbol{\eta}$ and for consistency of the proposed estimator. Briefly, under mild conditions[27,28], the expectations of the score vectors derived from the working models (2.1) and (2.2) are used to convert their estimates into a system of equations with unique solutions that correspond to the true parameters $\boldsymbol{\eta}$ of the model (2.3). The expectations are

$$\mathbb{E}\left\{\left[Y - \frac{\exp(\hat{\theta}^*)}{1+\exp(\hat{\theta}^*)}\right]\begin{bmatrix}1\\ E\\ \boldsymbol{C}\\ C_1 E\end{bmatrix}\right\} = \mathbf{0}, \text{ and} \tag{2.4}$$

$$\mathbb{E}\left\{\left[Y - \frac{\exp(\hat{\theta}^{**})}{1+\exp(\hat{\theta}^{**})}\right]\begin{bmatrix}1\\ \boldsymbol{M}\\ \boldsymbol{C}\end{bmatrix} | E=1\right\} = \mathbf{0}, \tag{2.5}$$

where $\hat{\theta}^* = \hat{\beta}_0^* + \hat{\beta}^* E + \sum_i \hat{\alpha}_i^* C_i + \hat{\zeta}^* C_1 E$, $\hat{\theta}^{**} = \hat{\beta}_0^{**} + \sum_i \hat{\alpha}_i^{**} C_i + \sum_j \hat{\gamma}^{**} \boldsymbol{M}_j$ and these expectations are taken under $P(Y = 1\,|\,E, \boldsymbol{M}, \boldsymbol{C})$ in (2.3) and $F(E, \boldsymbol{M}, \boldsymbol{C})$, the joint distribution of the racial groups, mediators and covariates. To estimate $\boldsymbol{\eta}$, we created the following system of equations based on the score vectors used to estimate the parameters from models (2.1) and (2.2):

$$\sum_{k=1}^{N} w_k \left[\frac{\exp(\theta_k)}{1+\exp(\theta_k)} - \frac{\exp(\hat{\theta}_k^*)}{1+\exp(\hat{\theta}_k^*)}\right]\begin{bmatrix}1\\ E_k\\ \boldsymbol{C}_k\\ C_{k1}E_k\end{bmatrix} = 0, \text{ and} \tag{2.6}$$

$$\sum_{k=1}^{N} w_k I(E_k = 1) \left[\frac{\exp(\theta_k)}{1+\exp(\theta_k)} - \frac{\exp(\hat{\theta}_k^{**})}{1+\exp(\hat{\theta}_k^{**})}\right]\begin{bmatrix}1\\ \boldsymbol{M}_k\\ \boldsymbol{C}_k\end{bmatrix} = 0, \text{ and} \tag{2.7}$$

where $\theta_k = \beta_0 + \beta E_k + \sum_i \alpha_i C_{ki} + \sum_j \gamma_j M_{kj} + \zeta C_{k1} E_k$ and w_k is the survey sampling weight for subject k in the NHIS survey. Note that in addition to all parameters from (2.3) we also estimated the nuisance parameters $(\hat{\beta_0}^{**}, \hat{\boldsymbol{\alpha}}^{**})$ from (2.2) using the equations in (2.7). Heuristically, identifiability of the parameters can be inferred as the number of equations in (2.6) and (2.7) is equal to the number of unknown parameters in the vectors β and $(\hat{\beta_0}^{**}, \hat{\boldsymbol{\alpha}}^{**})$. Rigorous proofs of identifiability and consistency of the estimates follow the same steps as in our previous work[20].

The derivation of the variance of $\hat{\boldsymbol{\eta}}$ requires accounting for the complex survey sampling scheme of the NHIS, a household interview survey. We propose a novel strategy for estimating the joint distribution of the parameters by calculating score vectors for each subject in the NHIS data set and treating them as a vector of the subject's features so that variance calculations can be conducted using available software for the analysis of surveys, e.g. the *R* package *survey* package. Appendix A contains details on the derivation of the variance of $\hat{\boldsymbol{\eta}}$.

Remarks:

1. Motivated by the observed differences in the estimated effects $\hat{\boldsymbol{\eta}}^*$ and $\hat{\boldsymbol{\eta}}^{**}$ from the working models (2.1) and (2.2), we fit separate population level outcome models (2.3) to men and women. However, for some applications it may be preferred to incorporate sex into model (2.3) as an additional covariate, to improve efficiency of the estimates. Heterogeneity by sex could then be accommodated using interactions, which are identifiable as shown previously[20].

2. In our data example we followed a standard of practice and discretized some patient's characteristics such as BMI and age. However, the approach outlined here also allows to model $\boldsymbol{C}$ and $\boldsymbol{M}$ as continuous variables. Consistency of the estimates with continuous exposures, mediators and covariates was previously presented in Derkach et al. [20]. While our new approach for the estimation of variance of $\hat{\boldsymbol{\eta}}$ outlined in Appendix A is still applicable, computing confidence intervals for the causal estimates and the percentages of causal effects mediated through the mediators requires modifications discussed later in this section.

3. Our approach allows population model (2.3) to assume a different structure than working models (2.1) and (2.2). E.g. in our example one could incorporate age as a continuous variable using a linear term. In fact, following our framework[20] and noticing that the system of equations (2.6 and 2.7) designate three equations (i.e. degrees of freedom) to model age ($\boldsymbol{C}$), one can easily demonstrate that age can be modeled flexibly using any functional form with three-degree freedoms, e.g. cubic polynomials.

Next, using $\hat{\boldsymbol{\eta}}$, we propose methods to estimate the proportion of the causal effects of race (E) mediated through all three risk factors ($\boldsymbol{M}$) simultaneously, without imposing any ordering or causal pathway. We make the same four assumptions as in Vansteelandt and Daniel[29] to ensure identifiability of causal effects under settings with multiple mediators. Estimates of the indirect effect of individual mediators can be obtained using the same equations as in[29] and the techniques given below.

The total effect (TE) of changing the exposure value from $E = 0$ (non-Hispanic Black) to $E = 1$ (non-Hispanic White) decomposes into the natural direct effect (NDE) and the natural indirect effect (NIE) through three risk factors, while controlling for age ($\boldsymbol{C}$),

$$TE_c = \mathbb{E}\left\{Y\left[1, \boldsymbol{M}(1)\right]|\boldsymbol{C} = \boldsymbol{c}\right\} - \mathbb{E}\left\{Y\left[0, \boldsymbol{M}(0)\right]|\boldsymbol{C} = \boldsymbol{c}\right\} = NDE_c + NIE_c, \tag{2.8}$$

where

$$NDE_c = \mathbb{E}\left\{Y\left[1, \boldsymbol{M}(0)\right]|\boldsymbol{C} = \boldsymbol{c}\right\} - \mathbb{E}\left\{Y\left[0, \boldsymbol{M}(0)\right]|\boldsymbol{C} = \boldsymbol{c}\right\} \tag{2.9}$$

and

$$NIE_c = \mathbb{E}\left\{Y\left[1, \boldsymbol{M}(1)\right]|\boldsymbol{C} = \boldsymbol{c}\right\} - \mathbb{E}\left\{Y\left[1, \boldsymbol{M}(0)\right]|\boldsymbol{C} = \boldsymbol{c}\right\}. \tag{2.10}$$

The percentage of the causal effect of race, E, mediated through $\boldsymbol{M}$ is calculated as the ratio

$$R_c = \frac{NIE_c}{TE_c} \tag{2.11}$$

for each age group. To estimate the expectations in (2.8)-(2.10), we used a non-parametric approach. We assumed that the distribution of $(E, \boldsymbol{M}, \boldsymbol{C})$ is given by point masses $\boldsymbol{p} = \left(p_1, \dots, p_L\right)$ at L unique values. Then, under model (2.3) and assumptions outlined in Imai et al.[4], the expectations in (2.8)-(2.10) are estimated using the conditional distribution of $\boldsymbol{M}$ given $E = e$ (0 or 1), where 0 denotes Blacks, and age group $\boldsymbol{c}$, as

$$\hat{\mathbb{E}}\left\{Y\left[e', \boldsymbol{M}(e)\right]|\boldsymbol{C} = \boldsymbol{c}\right\} = \frac{1}{T_e}\sum_{l=1}^{L}\frac{\exp(\hat{\beta}_0 + \hat{\beta}e' + \sum_j \hat{\gamma}_j M_{lj} + \sum_i \hat{\alpha}_i C_{li} + \hat{\zeta} C_{l1} e')}{1 + \exp(\hat{\beta}_0 + \hat{\beta}e' + \sum_j \hat{\gamma}_j M_{lj} + \sum_i \hat{\alpha}_i C_{li} + \hat{\zeta} C_{l1} e')} I(E_l = e, \boldsymbol{C}_l = \boldsymbol{c})\hat{p}_l, \tag{2.12}$$

where $T_e = \sum_{l=1}^{L} I(E_l = e, \boldsymbol{C}_l = \boldsymbol{c})\hat{p}_l$ estimates the joint probability of $E = e$ and $\boldsymbol{C} = \boldsymbol{c}$.

Confidence intervals for the causal estimates and the percentages of causal effects mediated through the three mediators (2.8)-(2.11) are calculated using a parametric bootstrap. Further details are given in Appendix B.

Remark: When the exposure, covariates or mediators are continuous, the dimension of vector $\boldsymbol{p}$ becomes large, and computing the asymptotic covariance matrix of $\left(\boldsymbol{\eta}, p_1, \dots, p_{L-1}\right)$ using the approach in Appendix B becomes challenging. To estimate confidence intervals of the causal parameters in this case, we propose to bootstrap the third data set, the NHIS survey, following an approach by Rao-Wu-Yue-Beaumont[30] which can handle complex survey designs. A disadvantage of this approach is its computational burden, as the entire NHIS survey has to be bootstrapped, even though our analysis uses only a small part of it. Further details on this algorithm are given in Appendix C.

2.2 | Sensitivity Analysis with Interaction Terms

Current approaches for mediation analysis often allow the effect of the mediator to be modified by the actual value of the exposure. For the colorectal cancer example we consider in Section 2.1, the interaction term between race and modifiable risk factors is not identifiable. However, under special settings, when estimates of effects of the mediators for each race group are reported, the interaction is identifiable, and it can be estimated by extending the system of equations (2.6) and (2.7). For our example information on the associations between the factors and CRC is not available for both racial groups. We suggest a sensitivity analysis by incorporating an interaction term defined as a proportion of the main effects of mediators on the outcome. We propose an extend outcome model (2.3), by incorporating the interaction term as function of $\boldsymbol{k} = (k_1, ..., k_j)$, user specified vector of parameters (assumed to be known):

$$P(Y = 1\,|E, \boldsymbol{M}, \boldsymbol{C}) = \frac{\exp(\beta_0 + \beta E + \sum_i \alpha_i C_i + \sum_j \gamma_j M_j + (\sum_j k_j \gamma_j M_j)(1 - E) + \zeta C_1 E)}{1 + \exp(\beta_0 + \beta E + \sum_i \alpha_i C_i + \sum_j \gamma_j M_j + (\sum_j k_j \gamma_j M_j)(1 - E) + \zeta C_1 E)}. \tag{2.13}$$

Similarly to the main method in Section 2.1, the expectations of the score vectors derived from the working models (2.1) and (2.2) are also used to create the system of equations (2.4) and (2.5) with unique solutions that correspond to the true parameters $\boldsymbol{\eta}$ of the model in (2.13). We suggest to use several values of $\boldsymbol{k}$ to determine a degree of the influence of potential interactions between the exposure and mediators on the causal effect estimates.

3 | SIMULATION STUDIES

We evaluated our procedures for estimating the regression parameters and causal effects in several simulation settings that closely mimicked our motivating example with a single binary exposure E, a vector of four binary mediators $\boldsymbol{M} = \left(M_1, \dots, M_4\right)$, and a vector of three binary covariates $\boldsymbol{C} = \left(C_1, C_2, C_3\right)$. We also compared efficiency of parameter and mediation effect estimates obtained from our approach to those estimated based on a single (survey) data set containing $(E, \boldsymbol{M}, \boldsymbol{C})$ and also measurements on the outcome Y.

3.1 | Data Generation

We assumed that the distribution of $(E, \boldsymbol{M}, \boldsymbol{C})$ was given by point masses $\boldsymbol{p} = \left(p_1, \dots, p_L\right)$ estimated from NHIS data for men. We generated two large cohorts by first generating a vector $(E, \boldsymbol{M}, \boldsymbol{C})$ using $\boldsymbol{p}$ and a binary outcome Y from model (2.3), where $\beta_0 = \log(0.01/0.99) = -4.6$, $\beta_1 = \log(1.3) = 0.26$, $\zeta = 0$, $\boldsymbol{\alpha} = \left(\alpha_1, \alpha_2, \alpha_3\right) = (0.26, 0.26, 0.26)$ and $\boldsymbol{\gamma} = \left(\gamma_1, \gamma_2, \gamma_3, \gamma_4\right) = (0, 0, 0, 0), (0.26, 0, 0, 0.26)$ and $(0.26, 0.26, 0.26, 0.26)$. For the first data set representing cancer registry data, we generated a cohort of size $N_1 = 2 \cdot 10^5$ and for the second data set representing an external case-control study with information on the modifiable risk factors, we sampled a nested case-control study with $N_2^0 = 5000$ controls and $N_2^1 = 5000$ cases from an independent cohort of the same size ($N = 10^5$). We considered two scenarios for the third data set containing information on the vector $(E, \boldsymbol{M}, \boldsymbol{C})$. In the first scenario, we assumed that the third data set was a random sample from a large cohort by generating $(E_k, \boldsymbol{M}_{\boldsymbol{k}}, \boldsymbol{C}_{\boldsymbol{k}})$, $k = 1, \dots, 10000$ from the estimated vector of point masses $\boldsymbol{p}$. In the second scenario, the third data set was the whole NHIS survey data set, without added variation. To compare efficiency of our approach to the 'best' case scenario in which one data set also contains outcome measurements, we additionally generated vector of outcomes Y_k, $k = 1, \dots, 10000$ using model (2.3) for the third data set. We used this complete data set to estimate $\boldsymbol{\eta}$ based on a logistic regression model, and the causal effects based on equations (2.8)-(2.11) and (2.12).

For each value of $\boldsymbol{\beta}$ and choice of the third data set we simulated 10,000 empirical studies. Then we applied our method to estimates $\hat{\boldsymbol{\eta}}^*$ from model (2.1) obtained from the first data set, $\hat{\boldsymbol{\gamma}}^{**}$ from estimating model (2.2) from the second data set, and using the third data set as the reference data set. The variance of the regression coefficients and confidence intervals were estimated as described in Appendices A and B using the R package *survey*[31].

3.2 | Simulation Results

Table 1 shows that our proposed method unbiasly estimated the true parameters of model (2.3) and the NDE and NIE for $\boldsymbol{C} = (0, 1, 0)$ from data sets with incomplete information for both types of the third data set (scenario 1 and 2). The asymptotic 95% confidence intervals (CIs) for the estimates of the regression coefficients had close to nominal coverage. The causal effects for other values of $\boldsymbol{C}$ were also unbiased (results not shown). However, the confidence intervals were conservative for some settings (Table 1). Specifically, when NIE was 0, the empirical coverage of the CIs was 96.9% and 99% for a randomly sampled third data set and the NHIS survey, respectively. The conservative nature of the CIs seen here is consistent with previous observations when testing the mediating effects of individual biomarkers[1,32].

When NIE was not equal to 0, the coverage of the 95% CIs was close to the nominal value for scenario 1 for the third data set. Generally, CIs had slightly conservative coverage in simulations when the third data set was the NHIS survey. This is because our procedure for the CI computation accounts for the uncertainty in the estimate of empirical distribution function when the third data set is a random sample, but not when we used the NHIS survey data in each simulation study, which we considered fixed.

Lastly, we compared the standard deviations of estimates of regression coefficients and causal effects obtained from our integration approach to those of estimates obtained from one single data set that had also outcome measurements ('best case scenario'). Supplemental Table S1 shows that our approach has much better efficiency compared to the 'best case scenario'. These gains in efficiency of our approach stems from gleaning information from the marginal estimates $\hat{\boldsymbol{\eta}}^*$ and $\hat{\boldsymbol{\gamma}}^{**}$. These results also agree with and complement our previous simulation results[20] where we assumed that study 1 or study 2 had all measurements instead of study 3. These scenarios are not applicable here, as SEER does not have detailed medical measurements and the nested case-control study[24] had a very small number of Black participants.

Table 1 Simulation results. The average value of an estimator over 10,000 simulations is reported with the coverage (Cov) of the true value by the 95% confidence interval computed using 1) the asymptotic variance for the regression parameter(i.e., Wald based CIs) and 2) using parametric bootstrap for causal effects (i.e., NDE_c and NIE_c) for a true $\beta = 0.26$. NDE_c and NIE_c are estimated for $\boldsymbol{C} = (0, 1, 0)$ (i.e. 50-59 years old males) with true values $NDE_c = \{0.38\%, 0.47\%, 0.57\%\}$ and $NIE_c = \{0\%, 0.06\%, 0.07\%\}$ for the three values of $\boldsymbol{\gamma}$ in the table below. The average values of the estimates of effects of covariates $\boldsymbol{C}$ were also close to true value with coverage close to nominal value (results are not shown).

True value $\boldsymbol{\gamma}$	β (Cov)	γ_1 (Cov)	γ_2 (Cov)	γ_3 (Cov)	γ_4 (Cov)	NDE (Cov)	NIE (Cov)
Scenario 1: the data set containing $(\boldsymbol{E}_k, \boldsymbol{M}_k, \boldsymbol{C}_k)$ was obtained by random sampling							
(0,0,0,0)	0.26 (95.0%)	0 (95.0%)	0 (94.9%)	0 (94.9%)	0 (94.6%)	0.38% (95.1%)	0.00% (96.9%)
(0.26,0,0,0.26)	0.26 (95.4%)	0.26 (95.3%)	0 (94.7%)	0 (95.2%)	0.26 (94.(9%)	0.47% (95.0%)	0.06% (95.7%)
(0.26,0.26,0.26,0.26)	0.26 (94.9%)	0.26 (94.7%)	0.26 (95.0%)	0.26 (95.3%)	0.26 (94.8%)	0.57% (94.(9%)	0.07% (95.4%)
Scenario 2: the data set containing $(\boldsymbol{E}_k, \boldsymbol{M}_k, \boldsymbol{C}_k)$ was the NHIS survey							
(0,0,0,0)	0.26 (95.1%)	0.26 (94.7%)	0 (94.(9%)	0 (95.2%)	0 (95.3%)	0.38% (95.3%)	0.00% (99%)
(0.26,0,0,0.26)	0.26 (95.4%)	0.26 (94.7%)	0 (95.0%)	0 (95.3%)	0.26 (94.8%)	0.47% (95.3%)	0.06% (100%)
(0.26,0.26,0.26,0.26)	0.26 (95.6%)	0.26 (95.2%)	0.26 (94.8%)	0.26 (94.9%)	0.26 (94.8%)	0.57% (95.6%)	0.07% (94.8%)

4 | MOTIVATING EXAMPLE: MEDIATION OF RACIAL DIFFERENCES FOR COLORECTAL CANCER

We analyzed men and women separately, as the risk factor associations differ between the two sexes. We restricted our analyses to men and women aged 40-80 years, who self-reported as non-Hispanic White or non-Hispanic Black and used race as the "exposure (E)" in our mediation framework. The outcome is a diagnosis of CRC, defined as cancer of the proximal or distal colon or rectum, within one year from variable ascertainment. A summary of the data sources used in this analysis data is presented in Table 2, and detailed information about each data source is presented below.

Table 2 Overview of the data sources used in motivating example.

Data source	**Available information**
SEER 17 Registries (https://seer.cancer.gov/)	Numbers of newly diagnosed CRC cases and individuals at risk in 2018, by race, sex, age at diagnosis, i.e. data on (Y, E, C). Model (2.1) was fit to estimate the parameters $\boldsymbol{\eta}^*$.
Published Literature[24]	Summary log-ORs $\hat{\boldsymbol{\gamma}}^{**}$ (with corresponding covariance matrix) for the associations of mediators (M) with CRC risk (Y) from fitting model (2.2) to a matched case-control study were shared with us. Results were reported for White men and women.
National Health Interview Survey (NHIS) *https://www.cdc.gov/nchs/nhis/nhis_2018_data_release.htm*	Individual level data on mediators and covariates for both racial groups $(\boldsymbol{M}, \boldsymbol{C}, \boldsymbol{E})$ and on survey sampling design variables (sampling weights, strata information).

Cancer registry data on race-specific colorectal cancer prevalence

For both non-Hispanic Whites and non-Hispanic Blacks, we used data on the number of individuals at risk and the number of newly diagnosed CRCs in 2018, by sex and age at diagnosis from the NCI's Surveillance, Epidemiology, and End Results (SEER; https://seer.cancer.gov/) 17 Registries, that cover 26% of the US population, including 23.6% percent of non-Hispanic Whites and 25.5% of non-Hispanic Blacks.

National Health Interview Survey (NHIS) data on race-specific joint mediator distributions

Data on the mediating factors, prevalence of CRC screening history (defined as a sigmoidoscopy and/or colonoscopy in the last ten years), BMI, and regular aspirin use (defined as taking aspirin 3 or more times a week) for Whites and Blacks were obtained from the National Health Interview Survey (NHIS) 2018, a cross-sectional, national survey that measures the health of the US civilian non-institutionalized population, *https://www.cdc.gov/nchs/nhis/nhis _2018_data_release.htm*. We used data and corresponding survey weights from 25,417 American adults from which 11,960 Whites and Blacks ages 40-79 years were used in the final analysis. Frequencies with 95% CIs were estimated by using the *survey* R package[31] to account for the survey sampling design, and are reported overall and separately for each age interval.

Estimates of the association of the mediators with CRC risk

In many settings only published association estimates are available. Here we utilized data used to develop an absolute risk model to predict the probability of developing CRC over a given time period using several well-established risk and protective factors for CRC[24]. This model combines SEER incidence and competing mortality rates with relative risks for risk factor associations, estimated by computing ORs from data on non-Hispanic White men and women from two age-matched case-control studies. Here, we combined proximal and distal CRC cases with rectal cancer cases from the two age matched case-control studies and fit conditional logistic regression models to accommodate the age-group matching separately to men ($N = 1675$ controls and 1407 cases) and women ($N = 1435$ controls and 1097 cases). We estimated ORs for CRC screening history, BMI and regular aspirin use and the corresponding covariance matrices from sex-specific models, and used this summary level information in our mediation analysis.

4.1 | Main Results

For both non-Hispanic Whites and non-Hispanic Blacks in the SEER 17 cancer registries in 2018, we calculated the one-year CRC incidence based on new CRC diagnoses shown in Table 3. Differences in the incidence were estimated by a fitting logistic regression model with CRC status as the outcome and race as the exposure for each age interval. Corresponding crude ORs and 95% CIs are reported in Table 3 as well. Most men and women were in the age range 40-79 years, with each 10-year age interval containing roughly 27-31% of individuals, and only approximately 16% of individuals were 70-79 years old. Compared to Whites, Blacks had higher one-year CRC crude incidence for almost all age intervals, 0.09% vs 0.07% for males aged 50-59, 0.16% vs 0.11% for males aged 60-69, and 0.21% vs 0.17% for males aged 70-79. However, males between 40 and 49 years had similar rates, 0.03% for both, Blacks and for Whites. Table 4 shows the prevalence of the three modifiable risk factors, CRC screening history in the last ten years, being overweight or obese (BMI≥25), and regular aspirin use in men, estimated using the 2018 NHIS data. Supplemental Table S2 shows these data for women. The age distribution in NHIS was similar to that in SEER, and only approximately 16% of persons were 70-79 years old.

Overall, a significantly lower proportion of Black men reported CRC screening than Whites (46% vs. 52.3%, p-value=0.01) and regular use of aspirin (29.2% vs. 33.6%, p-value =0.03). There was no statistically significant difference in BMI by racial/ethnic group (Table 4). Based on the age-specific analysis, we observed that the difference in rates of colonoscopy/sigmoidoscopy was primarily driven by a significant difference of 12.3% (95% CI: 3.8-20.8, p-value=0.005) in males aged 50 to 59 years. This difference diminished in older age groups. Lastly, there was no substantial difference in the prevalence of these risk factors between White and Black men 40-49 years old. In contrast to men, Black women overall were more likely to be obese than White women (53.7% vs. 35.9%, p-value<0.001). Differences in BMI between the Black and White women were persistent independent of age. Black women overall were less likely to report CRC screening (44.4% vs. 51.5%, p-value=0.002) (Supplement Table S2). However, Black women aged 70-79 years used aspirin more often than White women (60% vs. 46.5%, p-value<0.001). This observed sex-specific heterogeneity of the risk factor distribution motivated us to conduct separate analyses for men and women.

Table 5 and Supplemental Table S3 report the ORs for the association between race, the mediators (the modifiable risk factors), and risk of CRC for the reduced models (2.1 and 2.2) and full model (2.3) using estimates from two incomplete models (fit to the SEER registry data and utilizing the ORs from the external study). The association of race with CRC risk decreased negligibly after adjusting for screening, being overweight or obese, and aspirin use (OR=0.79 vs. 0.77 for men and 0.79 vs.

Table 3 Population at risk and number of CRC cases (with crude incidence in parenthesis) in 2018 for White and Black men and women in each age category from the SEER 17 database. ORs and corresponding 95% CIs were estimated from logistic regression with race as the exposure (E) using Blacks as the reference group, and a CRC diagnosis during the 2018 year as the outcome (Y) fitted to each age-group.

Age Interval	N (%)	N of NH-Black cases (%)	N of NH-White cases (%)	ORs (95 % CI)
Men (N= 12,316,564)				
40-49	3,500,756 (27)	202 (0.03)	956 (0.03)	1.05 (0.9, 1.22)
50-59	4,018,761 (31)	569 (0.09)	2493 (0.07)	0.83 (0.76, 0.91)
60-69	3,548,680 (27)	730 (0.16)	3371 (0.11)	0.7 (0.64, 0.75)
>70	2,071,981 (16)	454 (0.21)	3221 (0.17)	0.82 (0.74, 0.90)
Women (N=12,984,887)				
40-49	3,529,517 (25)	224 (0.03)	733 (0.03)	0.83 (0.72, 0.96)
50-59	4,149,151 (30)	482 (0.07)	1809 (0.05)	0.8 (0.73, 0.89)
60-69	3,865,496 (28)	639 (0.11)	2547 (0.08)	0.7 (0.64, 0.76)
>70	2,456,980 (18)	458 (0.15)	2817 (0.13)	0.87 (0.78, 0.96)

0.78 for women), indicating that most of racial differences in CRC incidence are not explained by these factors. However, the difference in rates between younger and older men and women increased after the adjustment of these factors. Lastly, we note that effect sizes of these three risk factors in the new joint model did not change compared to the external study ORs (i.e. $\hat{\gamma}^{**} = \hat{\gamma}$). This is because the external ORs are estimated from model (2.2), which is equivalent to fitting model (2.3) to Whites only, and with no interactions between the mediators and the other covariates in model (2.3) the corresponding main effects remain the same. The results from the mediation analysis (Table 6 and Supplemental Table S4) are consistent with the reported results for the full model with 95% CIs calculated using the 2.5% and 97.5% quantiles of a bootstrap distribution function from a parametric bootstrap with 10,000 bootstrap repetitions (see Appendix B). For men and women, racial differences persisted for every age category, except for the TE for men, which was not statistically different from 0 for ages 40-50. With increasing age, both TE and NDE also increased, with the largest NDE of 0.05% (95% CI: 0.035-0.060%) and 0.035% (95% CI: 0.019-0.053%) for men and women ages 70-79 years. Overall, the mediators (CRC screening history, being overweight or obese, and regular aspirin use) did not explain a substantial proportion of the difference in the rates between the two races. Only for males ages 50-59 years did we see a significant proportion, 23% (95% CI: 9-38%), of the effect meditated by these factors. Women in the same age group had a slightly lower proportion, 14% (95% CI: -2-30%) mediated. We also calculated 95% CIs using the bootstrap approach propsed in Appendix C. The confidence intervals that were based on asymptotic normality (see Table 5) and the bootstrap ones (Supplemental Table S5) were almost identical, as were the confidence intervals for the causal effects (see Table 6 and Supplemental Table S6). However, the bootstrap approach was computationally noticeably slower as it requires re-sampling the entire NHIS study.

4.2 | Sensitivity analysis

In the first sensitivity analysis we varied the effects of the modifiable risk factors on CRC risk based on the recently published literature[33,34,35,36,37] to evaluate the robustness of our conclusions. None of these investigations considered all four factors studied here simultaneously; however, most of reported effects were close to those used in our analysis. In Supplement Table S7, we report the range of the estimated causal effects with 95%CIs based on parametric bootsrap that accounted only for uncertainty due to empirical distribution estimated using NHIS data set. These results confirm our conclusion that the racial differences in CRC incidence are not fully explained by these four factors. For example, for males ages 50-59 years we did see a large increase in the proportion of the effect meditated by these factors; from 23% (95% CI: 9-38%) reported here to 44% (95% CI: 14-64%) when the association parameter for screening was assumed to be $OR = 0.3$ instead of $OR = 0.68$ (Table 5), but still did not completely explaining entire effect. Similar conclusions hold also for women.

Table 4 Prevalence of modifiable risk factors (M) in Black and White men in the general US population overall and for each age category estimated from NHIS 2018 survey. NHIS samples are weighted to US population. Total NHIS sample size was $N = 5525$, with $N_B = 701$ and $N_A = 4824$ Black and White men.

Risk Factors	Prevalence in NH-Black	Prevalence in NH-Whites	Difference in prevalence (95% CI)
Colonoscopy/sigmoidoscopy within last ten years	46.0%	52.30%	-6.2% (-11.0, -1.5)
BMI			
Healthy weight ($\leq 24.9 kg/m^2$)	21%	20.70%	0.3% (-3.6, 4.1)
Overweight ($25 - 30 kg/m^2$)	39.50%	43.0%	-3.5% (-8.1, 1.1)
Obese ($> 30 kg/m^2$)	39.50%	36.30%	3.2% (-1.4, 7.8)
Regular aspirin use	29.20%	33.60%	-4.5% (-8.5, -0.5)
Age interval: 40-49 representing 26% of the general US population			
Colonoscopy/sigmoidoscopy within last ten years	20.10%	15.90%	4.2% (-3.2, 11.6)
BMI			
Healthy weight ($\leq 24.9 kg/m^2$)	18.70%	19.80%	-1.1% (-8.1, 6)
Overweight ($25 - 30 kg/m^2$)	39.30%	43.90%	-4.6% (-13.6, 4.4)
Obese ($> 30 kg/m^2$)	41.90%	36.30%	5.6% (-3.6, 14.9)
Regular aspirin use	13.10%	11.50%	1.6% (-5.1, 8.3)
Age interval: 50-59 representing 29% of the general US population			
Colonoscopy/sigmoidoscopy within last ten years	41.90%	54.20%	-12.3% (-20.8, -3.8)
BMI			
Healthy weight ($\leq 24.9 kg/m^2$)	19.80%	19.20%	0.5% (-6.8, 7.9)
Overweight ($25 - 30 kg/m^2$)	37.70%	44.30%	-6.7% (-15.5, 2.2)
Obese ($> 30 kg/m^2$)	42.50%	36.40%	6.1% (-2.6, 14.9)
Regular aspirin use	24.70%	26.50%	-1.8% (-9.4, 5.8)
Age interval: 60-69 representing 29% of the general US population			
Colonoscopy/sigmoidoscopy within last ten years	67.40%	69%	-1.6% (-9.2, 6)
BMI			
Healthy weight ($\leq 24.9 kg/m^2$)	22.90%	21.50%	1.4% (-5.6, 8.4)
Overweight ($25 - 30 kg/m^2$)	42.40%	41.50%	1% (-7.1, 9)
Obese ($> 30 kg/m^2$)	34.70%	37.10%	-2.4% (-10, 5.3)
Regular aspirin use	40.40%	45.60%	-5.1% (-13, 2.8)
Age interval: 70-79 representing 16% of the general US population			
Colonoscopy/sigmoidoscopy within last ten years	77.10%	73.5%	3.6% (-5.4, 12.7)
BMI			
Healthy weight ($\leq 24.9 kg/m^2$)	26.10%	23.30%	2.8%(-5.9, 11.4)
Overweight ($25 - 30 kg/m^2$)	39.40%	42.10%	-2.8% (-13.2, 7.7)
Obese ($> 30 kg/m^2$)	34.60%	34.50%	0.1% (-9.8, 9.8)
Regular aspirin use	57.9%	57.3%	0.6% (-10, 11.1)

Table 5 Estimates of effects of race (E), modifiable risk factors ($\boldsymbol{M}$) and age (C) on the risk of colorectal cancer in men from two data sources (SEER and external case-control study) and our estimates of the effects for the full joint model (2.3). ORs and corresponding 95% CIs were estimated for logistic regression (2.3) from ORs obtained from reduced models (2.1 and 2.2) using estimating equations (2.4) and (2.5).

Variables	ORs estimated from SEER	Estimates of ORs provided from external study[24]	Estimated ORs for the full model (2.3)
Race: NH White vs NH Black	0.77 (0.73,0.81)	–	0.79 (0.72,0.86)
Colonoscopy/sigmoidoscopy within last ten years	–	0.68 (0.63,0.74)	0.68 (0.63,0.74)
BMI: Overweight ($25 - 30kg/m^2$) vs Healthy weight ($\leq 24.9kg/m^2$)	–	0.98 (0.88,1.08)	0.98 (0.88,1.08)
BMI: Obese ($> 30kg/m^2$) vs Healthy weight ($\leq 24.9kg/m^2$)	–	1.3 (1.16,1.47)	1.3 (1.16,1.47)
Regular aspirin use	–	0.89 (0.82,0.96)	0.89 (0.82,0.96)
Age: 40-49 vs 70-79	0.14 (0.12,0.17)	Not reported	0.11 (0.09,0.13)
Age: 50-59 vs 70-79	0.42 (0.40,0.44)	Not reported	0.37 (0.35,0.39)
Age: 60-69 vs 70-79	0.64 (0.62,0.68)	Not reported	0.62 (0.59,0.65)
Race*Age: 40-50	1.36 (1.16,1.59)	Not reported	1.33 (1.12,1.58)

Table 6 Estimates of causal mediation effects in men. The natural direct effect (NDE) and natural indirect effect (NIE) are estimated for each age interval. The ratio was calculated as NIE/(NIE+NDE). 95% confidence intervals (CIs) were estimated based on the percentiles of the bootstrap distribution function using 10,000 parametric bootstrap replicates.

Age Interval	Number(%) of Black CRC cases (%)	Number(%) of White CRC cases (%)	NDE (95% CI)	NIE (95% CI)	Ratio (95% CI)
40-50	202 (0.032%)	956 (0.032%)	0.00% (-0.004,0.006)	0.00% (-0.001,0.001)	-0.03 (-2.85, 2.73)
50-60	569 (0.089%)	2493 (0.074%)	-0.02% (-0.025,-0.015)	-0.01% (-0.011,-0.002)	0.23 (0.09,0.38)
60-70	730 (0.16%)	3371 (0.11%)	-0.03% (-0.039,-0.023)	0.00% (-0.007,0.005)	0.03 (-0.19,0.20)
>70	454 (0.21%)	3321 (0.17%)	-0.05% (-0.060,-0.035)	0.00% (-0.010,0.016)	-0.08 (-0.49,0.17)

Lastly, we evaluated if potential differences in the effect sizes of risk factors between the two racial groups could explain the disparities in incidence using the sensitivity analysis proposed in Section 2.2. We varied the percentage k representing the difference in effects of the mediators on outcome between the two race groups from -50% to 50%. However, the results were robust even when we assumed the presence of strong interactions between the exposure and the mediators (see Supplemental Figures 1 and 2).

5 | DISCUSSION

We developed a novel approach to estimate the mediation effects of several modifiable risk factors on racial differences in CRC incidence for men and women aged 40 years and older that integrates data from publicly available sources and published association estimates. Our innovative method allows one to combine data on 1) CRC incidence rates of non-Hispanic Whites and Blacks from SEER cancer registries, 2) data on the distribution of race and the potential mediators from the US population representative NHIS survey, and 3) estimates of the ORs for three potential mediators from an external case-control study[24]. The proposed method yields unbiased estimates of model parameters and causal effects, and both asymptotic and parametric bootstrap confidence intervals with close to nominal coverage.

In our previous work[20] we demonstrated identifiability and consistency of the regression model parameters for settings discussed in this paper under the assumption that all three data sets are obtained from the same underlying population by simple random sampling. Here, we relax this assumption by allowing the data set with joint information on the exposure and mediators to come from a complex survey sample, the NHIS, a household interview survey. We propose a novel strategy for accommodating the complex survey sampling in the asymptotic variance calculation. As an alternative to using asymptotic confidence intervals of the causal effects we propose a new bootstrap, that combines a parametric component with a non-parametric piece based on the Rao-Wu-Yue-Beaumont method[30] to accommodate the complex survey sampling design of the NHIS. This approach is computationally intensive, as ones needs to re-sample the entire survey, even though only a small part is used for the analysis. Nonetheless, this bootstrap is required for continuous exposures and or mediators, when their joint distribution function cannot be reliably estimated from the empirical distribution, and makes our method broadly applicable. Our novel bootstrap performed well in finite samples.

We expect there will be a growing interest in integrating partial information from multiple data sources into mediation models. In some settings individual level data with an outcome are provided. In our CRC study, the large SEER data set could have been integrated by developing new a likelihood based approach similar to Derkach et al.[20], which would have improved efficiency for estimating the natural direct effect but not natural indirect effect(see Supplemental Materials of Derkach et al.[20]); however, this approach is more computationally challenging. One limitation of our work is that we assumed that the distribution of (Y, E, M, C) is the same in the three data sources. This assumption is reasonable for our motivating CRC example, as we combined data from nationally and population representative sources. However, for some applications this assumption might be violated. In future work we will account for this heterogeneity, possibly utilizing ideas from statistical methods based on calibration equations[38], and inverse probability weighting[39,40]. Other extensions include response-dependent and two-phase studies that account for the sampling designs[41,42,43,44].

Our approach provides an attractive exploratory tool to evaluate whether, and to what extent, well-established risk factors explain the racial disparity observed for CRC incidence in the US. We found that CRC screening history within the last ten years, BMI, and regular use of aspirin did not explain a large proportion of this racial differences in 2018. The exception were men aged 50 to 59, years for whom an estimated 17% of the difference was explained by these three factors. The general lack of significant results may be partially explained by considerable uncertainty in the OR estimates and a limited number of participants in NHIS in each age category (Tables 3 and 3). However, the use of nationally representative sources of information (the SEER cancer registries and the NHIS survey) makes results broadly interpretable for the general US population. One source of bias is a potential difference in correlation between risk factors measured in NHIS and in the case-control study that gave rise to the summary ORs. For example, it is possible that healthier participants are more likely to be of normal weight, undergo CRC screening, and take aspirin daily have taken part in the epidemiological study. Further extensions might include incorporating correlations between risk factors in addition to ORs and their uncertainties. Somewhat reassuringly though, results from sensitivity analysis that used association estimates from other published studies led to very similar conclusions.

Results from mediation analyses need to be interpreted with caution, when studying exposures like race, as there are no reasonable hypothetical interventions on such exposures[45]. Under our approach adopted here, one can interpret the results as a way to quantify change due to intervening on the modifiable risk factors. A similar counterfactual framework was previously studied other social disparities[46,47,48,49]. The mediation analysis method adopted here controlled for measured confounders such as age and sex. That limits the interpretability of the results, as some of residual disparities in CRC may be attributable to unmeasured factors such as social economic variables at birth or at the time of the analysis or racial discrimination[48,49]. We thus view the results of our CRC example descriptively, as an aid in identifying inequities. However, our method is broadly applicable to study mediation effects for any exposure of interest in the setting of data sets with partial information.

In summary, we developed novel methods to integrate incomplete information from multiple data sources into a single regression model that we used to estimate causal mediation effects. Our methods have broad applications beyond racial disparities, to examine mediation when not all three variables, the outcome, the exposure and the mediators are available in same data set. We used our methods to estimate the causal effect of race and mediators on CRC risk in men and women aged 40 or older. Our results may aid scientists in determining what risk factors drive racial disparities in CRC and other outcomes and can also be helpful in designing future epidemiological studies that include multiple racial groups.

ACKNOWLEDGEMENTS

Andriy Derkach is supported by an institutional grant from the National Institutes of Health, National Cancer Institute (NCI) grant P30CA008748. Elizabeth D. Kantor is supported by P30 from the National Cancer Institute of the National Institutes of Health as well as Geoffrey Beene Cancer Research Center at Memorial Sloan Kettering Cancer Research Center.

SUPPORTING INFORMATION

Supplemental Tables and Figures, referenced in Section 4, are available as part of the online article. The data and code and data would be provided on the request.

□

APPENDIX

A ASYMPTOTIC COVARIANCE MATRIX OF $\hat{\boldsymbol{\eta}}$

The derivation of the asymptotic covariance matrix of the estimate of $\boldsymbol{\eta}$ follows similar steps as in Derkach et al.[20].

Let $\boldsymbol{S}\left(E_k, \boldsymbol{M}_{\boldsymbol{k}}, \boldsymbol{C}_{\boldsymbol{k}}; \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\eta}}^*, \hat{\boldsymbol{\eta}}^{**}\right) = \left(\boldsymbol{S}\left(E_k, \boldsymbol{M}_{\boldsymbol{k}}, \boldsymbol{C}_{\boldsymbol{k}}; \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\eta}}^*\right), \boldsymbol{S}\left(E_k, \boldsymbol{M}_{\boldsymbol{k}}, \boldsymbol{C}_{\boldsymbol{k}}; \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\eta}}^{**}\right)\right)$ be the score vector for kth subject that is element of sums in (2.6) and (2.7) with

$$\boldsymbol{S}\left(E_k, \boldsymbol{M}_{\boldsymbol{k}}, \boldsymbol{C}_{\boldsymbol{k}}; \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\eta}}^*\right) = \left[\frac{\exp(\hat{\theta}_k^*)}{1+\exp(\hat{\theta}_k^*)} - \frac{\exp(\hat{\theta}_k)}{1+\exp(\hat{\theta}_k)}\right] \begin{bmatrix} 1 \\ E_k \\ \boldsymbol{C}_k \\ C_{k1}E_k \end{bmatrix}, \text{ and}$$

$$\boldsymbol{S}\left(E_k, \boldsymbol{M}_{\boldsymbol{k}}, \boldsymbol{C}_{\boldsymbol{k}}; \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\eta}}^{**}\right) = I\left(E_k = 1\right) \left[\frac{\exp(\hat{\theta}_k^{**})}{1+\exp(\hat{\theta}_k^{**})} - \frac{\exp(\hat{\theta}_k)}{1+\exp(\hat{\theta}_k)}\right] \begin{bmatrix} 1 \\ \boldsymbol{M}_k \\ \boldsymbol{C}_k \end{bmatrix},$$

where $\hat{\theta}_k^*$, $\hat{\theta}_k^{**}$ and $\hat{\theta}_k$ are defined in Section 2.1 and $\hat{\boldsymbol{\eta}}^* = \left(\hat{\beta_0}^*, \hat{\beta}^*, \hat{\boldsymbol{\alpha}}^*, \hat{\zeta}^*\right)$ are estimates obtained from SEER 2018 data set, $\hat{\boldsymbol{\eta}}^{**} = (\hat{\beta_0}^{**}, \hat{\boldsymbol{\alpha}}^{**}, \hat{\boldsymbol{\gamma}}^{**})$ are estimates obtained from external source (only $\hat{\boldsymbol{\gamma}}^{**}$ is reported) and $\hat{\boldsymbol{\eta}} = (\hat{\beta}_0, \hat{\beta}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\alpha}}, \hat{\zeta})$ are our estimates of the parameters in the full model (2.3). By using a Taylor expansion around the true values $(\boldsymbol{\eta}, \beta_0^{**}, \boldsymbol{\alpha}^{**})$ and $(\boldsymbol{\eta}^*, \boldsymbol{\gamma}^{**})$ in the sums (2.6) and (2.7), we can demonstrate that

$$\begin{aligned} o_p\left(\sqrt{N}\right) = & \sum_{k=1}^{N} w_k \boldsymbol{S}\left(E_k, \boldsymbol{M}_k, \boldsymbol{C}_k; \boldsymbol{\eta}, \boldsymbol{\eta}^*, \boldsymbol{\eta}^{**}\right) + \\ & + \sum_{k=1}^{N} w_k \nabla_{\boldsymbol{\eta}, \beta_0^{**}, \boldsymbol{\alpha}^{**}} \boldsymbol{S}\left(E_k, \boldsymbol{M}_k, \boldsymbol{C}_k; \boldsymbol{\eta}, \boldsymbol{\eta}^*, \boldsymbol{\eta}^{**}\right) \left(\begin{bmatrix} \hat{\boldsymbol{\eta}} \\ \hat{\beta}_0^{**} \\ \hat{\boldsymbol{\alpha}}^{**} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\eta} \\ \beta_0^{**} \\ \boldsymbol{\alpha}^{**} \end{bmatrix}\right) + \\ & + \sum_{k=1}^{N} w_k \nabla_{\boldsymbol{\eta}^*, \boldsymbol{\gamma}^{**}} \boldsymbol{S}\left(E_k, \boldsymbol{M}_k, \boldsymbol{C}_k; \boldsymbol{\eta}, \boldsymbol{\eta}^*, \boldsymbol{\eta}^{**}\right) \left(\begin{bmatrix} \hat{\boldsymbol{\eta}}^* \\ \hat{\boldsymbol{\gamma}}^{**} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{\gamma}^{**} \end{bmatrix}\right), \end{aligned}$$

where ∇ denotes the gradient operator that yields the matrix of partial derivatives. By applying Theorem 1 of Yuan and Jennrich[50] we obtain the following asymptotic result:

$$\lim_{N\to\infty} \sqrt{N}\left[(\hat{\boldsymbol{\eta}}, \hat{\beta}_0^{**}, \hat{\boldsymbol{\alpha}}^{**}) - (\boldsymbol{\eta}, \beta_0^{**}, \boldsymbol{\alpha}^{**})\right] \xrightarrow[d]{} N\left[\boldsymbol{0}, \mathbb{J}^{-1} B\left(\boldsymbol{\eta}, \boldsymbol{\eta}^*, \boldsymbol{\eta}^{**}\right) \mathbb{J}^{-1'} + \mathbb{J}^{-1}\, \Omega \Sigma_{\boldsymbol{\eta}^*, \boldsymbol{\gamma}^{**}} \Omega' \, \mathbb{J}^{-1'}\right],$$

where $\mathbb{J} = \mathbb{E}\left\{\nabla_{\boldsymbol{\eta},\beta_0^{**},\boldsymbol{\alpha}^{**}}\boldsymbol{S}\left(E_k,\boldsymbol{M}_k,\boldsymbol{C}_k;\boldsymbol{\eta},\boldsymbol{\eta}^*,\boldsymbol{\eta}^{**}\right)\right\}$, $\Omega = \mathbb{E}\left\{\nabla_{\boldsymbol{\eta}^*,\boldsymbol{\gamma}^{**}}\boldsymbol{S}\left(E_k,\boldsymbol{M}_{\boldsymbol{k}},\boldsymbol{C}_{\boldsymbol{k}};\boldsymbol{\eta},\boldsymbol{\eta}^*,\boldsymbol{\eta}^{**}\right)\right\}$, $B\left(\boldsymbol{\eta},\boldsymbol{\eta}^*,\boldsymbol{\eta}^{**}\right) = Cov\left\{\boldsymbol{S}\left(E_k,\boldsymbol{M}_{\boldsymbol{k}},\boldsymbol{C}_{\boldsymbol{k}};\boldsymbol{\eta},\boldsymbol{\eta}^*,\boldsymbol{\eta}^{**}\right)\right\}$, and $\Sigma_{\boldsymbol{\eta}^*,\boldsymbol{\gamma}^{**}} = NCov(\hat{\boldsymbol{\eta}}^*,\hat{\boldsymbol{\gamma}}^{**})$ is based on the reported covariance matrix of the estimates from SEER and the external study. Estimation of $\mathbb{J},\Omega$, and $B\left(\boldsymbol{\eta},\boldsymbol{\eta}^*,\boldsymbol{\eta}^{**}\right)$ was based on the NHIS survey, accommodating the complex survey sample design. We used the R package survey[31] to calculate the expected values of the Jacobians and the covariances of the score vectors. As an example, we provide the following R code that was used to estimate $B\left(\boldsymbol{\eta},\boldsymbol{\eta}^*,\boldsymbol{\eta}^{**}\right)$:

```
ScoreVector = calculateScoreVectorasAbove(Data1,Data2,DataNHIS)
DataNHISplusScoreVetor = cbind(DataNHIS, ScoreVector)
nhissvy=svydesign(id=~PPSU,strata=~STRAT,nest=TRUE,weights=~TFA_SA,data=DataNHISplusScoreVetor)
subgrp <- subset(nhissvy,(RACERPI2<3 & SEX<2 & HISPAN_I>8))
MeanScoreVector = svymean(~cbind(S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11,S12,S13,S14),design=subgrp)
VarianceScoreVector = vcov(MeanScoreVector)
```

The estimation of the expectations $\mathbb{J}$ and Ω follows similar steps, using the function *svymean*[31].

B PARAMETRIC BOOTSTRAP

We used a parametric bootstrap to estimate standard errors and 95% confidence intervals of the causal effect estimates in (2.12). In each parametric bootstrap replication, we generated a vector of $(\boldsymbol{\eta},p_1,\dots,p_{L-1})$ from their asymptotic distribution that we derive here. The derivation of the asymptotic covariance matrix of the estimate of $\left(\boldsymbol{\eta},p_1,\dots,p_{L-1}\right)$ follows similar steps as outlined in Appendix A. First, we extended the sums in (2.6) and (2.7) by adding the following set of estimating equations for $\hat{\boldsymbol{p}}_{-L}$, the estimate of the empirical distribution defined by the vector $\boldsymbol{p}_{-L} = \left(p_1,\dots,p_{L-1}\right)$

$$\sum_{k=1}^{N} w_k \boldsymbol{S}\left(E_k,\boldsymbol{M}_k,\boldsymbol{C}_k,\hat{\boldsymbol{p}}_{-L}\right) = \sum_{k=1}^{N} w_k \begin{bmatrix} \frac{I\left(E_k=e_1,\boldsymbol{M}_k=\boldsymbol{m}_1,\boldsymbol{C}_k=\boldsymbol{c}_1\right)}{\hat{p}_1} - \frac{I\left(E_k=e_L,\boldsymbol{M}_k=\boldsymbol{m}_L,\boldsymbol{C}_k=\boldsymbol{c}_L\right)}{1-\sum_{l=1}^{L-1}\hat{p}_l} \\ \vdots \\ \frac{I\left(E_k=e_{L-1},\boldsymbol{M}_k=\boldsymbol{m}_{L-1},\boldsymbol{C}_k=\boldsymbol{c}_{L-1}\right)}{\hat{p}_{L-1}} - \frac{I\left(E_k=e_L,\boldsymbol{M}_k=\boldsymbol{m}_L,\boldsymbol{C}_k=\boldsymbol{c}_L\right)}{1-\sum_{l=1}^{L-1}\hat{p}_l} \end{bmatrix} = \boldsymbol{0}$$

Second, we extend the vector $\boldsymbol{S}\left(E_k,\boldsymbol{M}_k,\boldsymbol{C}_k;\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\eta}}^*,\hat{\boldsymbol{\eta}}^{**}\right)$ by adding $\boldsymbol{S}\left(E_k,\boldsymbol{M}_k,\boldsymbol{C}_k;\hat{\boldsymbol{p}_{-L}}\right)$ to it. Lastly, the asymptotic distribution of $\left(\boldsymbol{\eta},p_1,\dots,p_{L-1}\right)$ can be easily derived by applying the same steps to

$$\boldsymbol{S}\left(E_k,\boldsymbol{M}_k,\boldsymbol{C}_k;\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\eta}}^*,\hat{\boldsymbol{\eta}}^{**},\hat{\boldsymbol{p}}_{-L}\right) = \left[\boldsymbol{S}\left(E_k,\boldsymbol{M}_k,\boldsymbol{C}_k,;\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\eta}}^*\right),\boldsymbol{S}\left(E_k,\boldsymbol{M}_k,\boldsymbol{C}_k;\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\eta}}^{**}\right),\boldsymbol{S}\left(E_k,\boldsymbol{M}_k,\boldsymbol{C}_k;\hat{\boldsymbol{p}}_{-L}\right)\right],$$

as in Appendix A.

C BOOTSTRAP THAT ACCOMMODATES SURVEY DATA

We also proposed a bootstrap approach that can accommodate the complex survey sampling used in the NHIS to estimate standard errors and 95% confidence intervals of the causal effect estimates in (2.12). Here, in contrast to the parametric bootstrap, we do not estimate the joint distribution of $(\hat{\alpha},\hat{\beta},\hat{\gamma},\hat{\mathbf{p}})$, but instead directly estimate the variance of the estimated NDE and NIE, based on the Rao-Wu-Yue-Beaumont method[30]. At each bootstrap replication, instead of re-sampling a vector $(\boldsymbol{E},\boldsymbol{M},\boldsymbol{C})$ with replacement, we generate new sampling weights w_b for entire NHIS data set and simulate new values of $\hat{\boldsymbol{\eta}}_b^*$ and $\hat{\boldsymbol{\gamma}}_b^{**}$ from the asymptotic distributions $N(\hat{\boldsymbol{\eta}}^*,\Sigma_{\boldsymbol{\eta}^*})$ and $N(\hat{\boldsymbol{\eta}}^{**},\Sigma_{\boldsymbol{\gamma}^{**}})$. We next estimate $(\boldsymbol{\eta},\mathbf{p})$ and both the NIE and NDE based on these new bootstrap weights and marginal estimates. Lastly, we estimate the variances and confidence intervals of the NDE and NIE using the sample variance and corresponding quantiles of these bootstrap replications. We used the R package *svrep*[51] to generate boostrap weights for the entire NHIS data set as in Rao-Wu-Yue-Beaumont[30], using sampled clusters with replacement for each stratum. We provide example R code below to estimate $(\hat{\boldsymbol{\eta}}_b,\hat{\mathbf{p}}_b)$ for a single boostrap sample. The estimation of the causal effects is conducted using output *hatetapB*.

```
subsetForAnalysis = indexOfMalesEligibleForAnalysis(DataNHIS)
etastarB = rmvnorm(1,hateta_star,Sigma_eta)
gammastarB = rmvnorm(1,hatganna_star,Sigma_gamma)
```

```
nhissvy=svydesign(id=~PPSU,strata=~STRAT,nest=TRUE,weights=~TFA_SA,data=DataNHIS)
bootstrap = as_bootstrap_design(nhissvy,type=’Rao-Wu-Yue-Beaumont’,replicates=1)
RWScale = bootstrap$repweights
UpdatedWeightB = DataNHIS$WTFA_SA*RWScale
hatetapB = estimateParameters(etastarB,gammastarB,UpdatedWeightB,DataNHIS,subsetForAnalysis)
```

References

1. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.. *Journal of Personality and Social Psychology* 1986; 51(6): 1173.

2. MacKinnon D, Dwyer J. Estimating mediated effects in prevention studies. *Evaluation Review* 1993; 17(2): 144–158.

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

4. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* 2011; 25(1): 51–71.

5. VanderWeele TJ. A unification of mediation and interaction: a four-way decomposition. *Epidemiology* 2014; 25(5): 749.

6. Steen J, Loeys T, Moerkerke B, Vansteelandt S. Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology* 2017; 186(2): 184–193.

7. Daniels MJ, Roy J, Kim C, Hogan J, Perri M. Bayesian Inference for the Causal Effect of Mediation. *Biometrics* 2012; 68(4): 1028-1036. doi: 10.1111/j.1541-0420.2012.01781.x

8. Derkach A, Pfeiffer R, Chen T, Sampson J. High dimensional mediation analysis with latent variables. *Biometrics* 2019; 75(3): 745-756. doi: 10.1111/biom.13053

9. Huang YT. Variance component tests of multivariate mediation effects under composite null hypotheses. *Biometrics* 2019; 75(4): 1191–1204.

10. Sampson J, Boca S, Moore S, Heller R, Wren J. FWER and FDR control when testing multiple mediators. *Bioinformatics* 2018; 1: 7.

11. Valeri L, Lin X, VanderWeele T. Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. *Statistics in Medicine* 2014; 33(28): 4875–4890.

12. Bind MA, Vanderweele T, Coull B, Schwartz J. Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics* 2016; 17(1): 122–134.

13. Sohn MB, Li H. Compositional mediation analysis for microbiome studies. *The Annals of Applied Statistics* 2019; 13(1): 661–681.

14. Bodelon C, Oh H, Derkach A, et al. Polygenic risk score for the prediction of breast cancer is related to lesser terminal duct lobular unit involution of the breast. *NPJ Breast Cancer* 2020; 6(1): 1–6.

15. Jin Z, Du X, Xu Y, et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 2020; 582(7811): 289–293.

16. Choudhury P, Wilcox AN, Brook MN, et al. Comparative validation of breast cancer risk prediction models and projections for future risk stratification. *JNCI: Journal of the National Cancer Institute* 2020; 112(3): 278–285.

17. Evans K, Sun B, Robins J, Tchetgen EJT. Doubly robust regression analysis for data fusion. *Statistica Sinica* 2021; 31(3): 1285–1307.

18. Miao W, Li W, Hu W, Wang R, Geng Z. Invited commentary: estimation and bounds under data fusion. *American Journal of Epidemiology* 2021; 191(4): 674-678. doi: 10.1093/aje/kwab194

19. Shi X, Pan Z, Miao W. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics* 2023; 15(1): e1581.

20. Derkach A, Sampson J, Pfeiffer R. Integrating incomplete data for mediation analysis. *Statistica Sinica* 2022: Epub ahead of print.

21. Doubeni CA, Selby K, Gupta S. Framework and strategies to eliminate disparities in colorectal cancer screening outcomes. *Annual review of medicine* 2021; 72: 383–398.

22. Grubbs SS, Polite BN, Carney Jr J, et al. Eliminating racial disparities in colorectal cancer in the real world: it took a village. *Journal of Clinical Oncology* 2013; 31(16): 1928.

23. Laiyemo AO, Doubeni C, Pinsky PF, et al. Race and colorectal cancer disparities: health-care utilization vs different cancer susceptibilities. *Journal of the National Cancer Institute* 2010; 102(8): 538–546.

24. Freedman AN, Slattery ML, Ballard-Barbash R, et al. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *Journal of Clinical Oncology* 2009; 27(5): 686.

25. McCullough ML, Zoltick ES, Weinstein SJ, et al. Circulating vitamin D and colorectal cancer risk: an international pooling project of 17 cohorts. *JNCI: Journal of the National Cancer Institute* 2019; 111(2): 158–169.

26. Wang L, He X, Ugai T, et al. Risk factors and incidence of colorectal cancer according to major molecular subtypes. *JNCI Cancer Spectrum* 2021; 5(1): pkaa089.

27. Chatterjee N, Chen Y, Maas P, Carroll R. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* 2016; 111(513): 107-117. doi: 10.1080/01621459.2015.1123157

28. White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; 50(1): 1-25.

29. Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass.)* 2017; 28(2): 258.

30. Rao J, Wu C, Yue K. Some recent work on resampling methods for complex surveys. *Survey Methodology* 1992; 18(2): 209–217.

31. Lumley T. Analysis of complex survey samples. *Journal of Statistical Software* 2004; 9: 1–19.

32. Derkach A, Moore SC, Boca SM, Sampson JN. Group testing in mediation analysis. *Statistics in Medicine* 2020; 39(18): 2423–2436.

33. Brenner H, Chang-Claude J, Jansen L, Knebel P, Stock C, Hoffmeister M. Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy. *Gastroenterology* 2014; 146(3): 709–717.

34. Wang X, O'Connell K, Jeon J, et al. Combined effect of modifiable and non-modifiable risk factors for colorectal cancer risk in a pooled analysis of 11 population-based studies. *BMJ Open Gastroenterology* 2019; 6(1): e000339.

35. Guirguis-Blake JM, Evans CV, Perdue LA, Bean SI, Senger CA. Aspirin use to prevent cardiovascular disease and colorectal cancer: updated evidence report and systematic review for the US Preventive Services Task Force. *Jama* 2022; 327(16): 1585–1597.

36. Doubeni CA, Corley DA, Quinn VP, et al. Effectiveness of screening colonoscopy in reducing the risk of death from right and left colon cancer: a large community-based study. *Gut* 2018; 67(2): 291–298.

37. Brenner H, Stock C, Hoffmeister M. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. *Bmj* 2014; 348.

38. Chen Y, Chen H. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2000; 62(3): 449-460. doi: 10.1111/1467-9868.00243

39. Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; 47(260): 663–685.

40. Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 2009; 96(3): 723–734.

41. Breslow N, Holubkov R. Maximum Likelihood Estimation of Logistic Regression Parameters under Two-phase, Outcome-dependent Sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1997; 59(2): 447-461. doi: 10.1111/1467-9868.00078

42. Lawless JF, Kalbfleisch JD, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1999; 61(2): 413-438. doi: 10.1111/1467-9868.00185

43. Lin DY, Zeng D. Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies. *Journal of the American Statistical Association* 2006; 101(473): 89-104. doi: 10.1198/016214505000000808

44. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. *Biometrika* 1997; 84(1): 57-71. doi: 10.1093/biomet/84.1.57

45. VanderWeele TJ. Invited commentary: counterfactuals in social epidemiology—thinking outside of "the box". *American Journal of Epidemiology* 2020; 189(3): 175–178.

46. VanderWeele TJ, Robinson WR. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* 2014; 25(4): 473.

47. Jackson JW, Williams DR, VanderWeele TJ. Disparities at the intersection of marginalized groups. *Social Psychiatry and Psychiatric Epidemiology* 2016; 51: 1349–1359.

48. Jackson JW, VanderWeele TJ. Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology* 2018; 29(6): 825.

49. Valeri L, Chen JT, Garcia-Albeniz X, Krieger N, VanderWeele TJ, Coull BA. The role of stage at diagnosis in colorectal cancer black–white survival disparities: a counterfactual causal inference approach. *Cancer Epidemiology, Biomarkers & Prevention* 2016; 25(1): 83–89.

50. Yuan KH, Jennrich RI. Estimating equations with nuisance parameters: Theory and applications. *Annals of the Institute of Statistical Mathematics* 2000; 52(2): 343–350.

51. Schneider B. svrep: Tools for Creating, Updating, and Analyzing Survey Replicate Weights. 2023. R package version 0.6.0.