The subtype-free average causal effect for heterogeneous disease etiology

A. Sasson 1, M. Wang 2,3,4, S. Ogino 3,5,6, D. Nevo 1,*

1Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel; 2Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, United States; 3Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, United States; 4Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA 02115, United States; 5Broad Institute of MIT and Harvard, Cambridge, MA 02142, United States; 6Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women’s Hospital Harvard Medical School, Boston, MA 02115, United States

*Corresponding author: D. Nevo, Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel (danielnevo@tauex.tau.ac.il).

ABSTRACT

Studies have shown that the effect an exposure may have on a disease can vary for different subtypes of the same disease. However, existing approaches to estimate and compare these effects largely overlook causality. In this paper, we study the effect smoking may have on having colorectal cancer subtypes defined by a trait known as microsatellite instability (MSI). We use principal stratification to propose an alternative causal estimand, the Subtype-Free Average Causal Effect (SF-ACE). The SF-ACE is the causal effect of the exposure among those who would be free from other disease subtypes under any exposure level. We study non-parametric identification of the SF-ACE and discuss different monotonicity assumptions, which are more nuanced than in the standard setting. As is often the case with principal stratum effects, the assumptions underlying the identification of the SF-ACE from the data are untestable and can be too strong. Therefore, we also develop sensitivity analysis methods that relax these assumptions. We present 3 different estimators, including a doubly robust estimator, for the SF-ACE. We implement our methodology for data from 2 large cohorts to study the heterogeneity in the causal effect of smoking on colorectal cancer with respect to MSI subtypes.

KEYWORDS: competing risks; molecular pathological epidemiology; principal stratification; survivor average causal effect.

1 INTRODUCTION

In recent decades, it has become clear that many diseases that share general clinical characteristics evolve through a range of heterogeneous molecular pathologic processes, which may be affected differently by the same exposure. For this reason, classifying a disease into subtypes according to criteria based on molecular characteristics can improve our etiologic understanding of the disease (Ogino et al., 2016).

In this paper, we consider the causal effect smoking may have on having colorectal cancer (CRC). One well-established CRC subtype classification is based on microsatellite instability (MSI). The 2 mutually exclusive subtypes are MSI-high (~15% of the cases) and non-MSI-high (~85%). Smoking has been found to be a strong risk factor for the MSI-high subtype, while weaker evidence was found that smoking is a risk factor for the non-MSI-high subtype (Carr et al., 2018; Amitay et al., 2020).

To study etiologic heterogeneity, researchers often use a multinomial regression model in which being healthy or having each of the disease subtypes forms the possible values of the outcome. However, the multinomial regression parameters do not correspond to well-defined causal effects, because the multinomial regression parameters are equivalent to the parameters obtained from a series of logistic regressions, each comparing 1 disease subtype to the healthy controls, resulting in selection bias (Nevo et al., 2021).

In this paper, we use data from 2 large cohorts to study the effect of smoking on the 2 CRC subtypes. We introduce an alternative estimand, inspired by the Survivor Average Causal Effect (SACE) (Rubin, 2006), commonly used to address truncation by death, where the outcome of interest is undefined for individuals who died before their outcome could be measured. The SACE is the average causal effect of a treatment among individuals who would have survived under both treatment/exposure levels.

Our proposed approach is based on the same idea as the SACE. First, we define principal strata with respect to the potential outcomes of other disease subtypes. Then, we define the Subtype-Free Average Causal Effect (SF-ACE) to be the average causal effect of the exposure on 1 MSI disease subtype among the individuals who would have been free of the other MSI disease subtype under either exposure level (smoking status).

A common assumption in truncation-by-death problems is Monotonicity, namely that exposure cannot hurt survival (Zhang and Rubin, 2003). In our setting, the definition of monotonicity is more nuanced. First, the assumption needs to be made (or not) for each subtype separately. Second, the disease could occur under both exposure statuses, but the subtype may differ,
a phenomenon we term exposure-induced subtype switching. Unlike the SACE, the SF-ACE is non-parametrically identifiable under certain monotonicity assumptions. Under weaker monotonicity assumptions, the SF-ACE is no longer identifiable. We develop bounds as well as a sensitivity analysis, the latter based on the unidentifiable exposure-induced subtype switching probabilities.

We consider 3 SF-ACE estimators. The first 2, a standardization-based estimator and an inverse probability of treatment weighting (IPTW) estimator, may rely on parametric modeling assumptions. We therefore present a doubly robust (DR) estimator that will be consistent if at least 1 of 2 models, but not necessarily both, is correctly specified.

The rest of the paper is organized as follows. Section 2 describes the epidemiologic problem and the available data. Section 3 presents the notations and the causal framework. Section 4 reviews existing approaches for etiologic disease heterogeneity and discusses available causal estimands. Section 5 presents a detailed study of the proposed causal estimand, the SF-ACE, including its interpretation, assumptions needed for its identification, bounds, sensitivity analyses, and estimation. Section 6 summarizes the results of our simulation studies. Section 7 presents various analyses of the data to study the causal effect of smoking on the 2 CRC subtypes. Section 8 offers conclusions. Proofs of the theoretical results are given in Web Appendix A. The R package TheFACE, available from the Comprehensive R Archive Network (CRAN), implements our methodology (Sasson, 2022b). Reproducibility materials are available from GitHub (Sasson, 2022a).

2 DATA AND PROBLEM DESCRIPTION

We use data from 2 large US cohorts. The Nurses Health Study (NHS) was established in 1976 and consists of female nurses aged 30–55 at the beginning of follow-up. The Health Professionals Follow-up Study (HPFS) was established in 1986 and consists of male health professionals aged 40–75 at the beginning of follow-up. Participants answered biennial questionnaires about lifestyle and health-related information. CRC diagnoses were reported by study participants. Formalin-fixed paraffin-embedded tumor tissue specimens were retrieved from hospitals across the United States, and a CRC diagnosis was confirmed by the study pathologist. Further details can be found in Ugai et al. (2022) and references therein.

We set age 60 as the baseline and defined exposure as ever versus never smoking by that age, a common approach to reduce measurement error. The outcome was CRC by age 70, subtyped by MSI status, with age 70 chosen to reduce the impact of death as a competing event.

We excluded individuals older than 60 at enrollment or those who died, had CRC, or were lost to follow-up before age 60, as well as those with missing baseline smoking data (Ugai et al., 2022), leaving 114,947 participants. We further excluded 5132 individuals (4.4%) who died before age 70 without a CRC diagnosis, resulting in a final sample of 109,815. Of these, 17,854 were CRC-free and under 70 at the time of the last questionnaire (94.8% were older than 65). We treated them as CRC-free at age 70 and assessed the impact on our results by comparing alternative analyses.

In our sample, the number of exposed (ever smoking) was 58,432 (53.3%), and the number of CRC cases by age 70 was 961 (0.9%). Of these, 358 had non-MSI-high CRC (37.3%) and 61 had MSI-high CRC (6.3%), and the subtype was missing for the remaining 542 (56.4%), as is often the case in such studies (Liu et al., 2018; Nevo et al., 2018). To have a clear presentation of the issues at the end, we develop our theory and methodology under the assumption that subtypes are known for all disease cases. We address the issue of missing subtypes in our analysis in Section 7 using inverse probability weighting (Liu et al., 2018).

The NHS and HPFS cohorts include a vast number of covariates collected through the biennial questionnaires. Table C.8 provides information on key baseline confounders we adjusted for in our analyses.

3 PRELIMINARIES

Using the potential outcomes framework, let $Y_i(a)$ denote the potential disease status had the exposure of individual $i = 1, \ldots, n$ been set to $A_i = a$ for $a = 0, 1$. In our study, $A_i = 0$ indicates no smoking by age 60, while $A_i = 1$ indicates ever smoking by age 60. If individual $i$ would have been diagnosed with disease subtype $k = 1, 2$ under exposure value $A_i = a$, then $Y_i(a)$ is $Y_i$ and if they would be disease-free under $A_i = a$, then $Y_i(a) = 0$. Let also $Y_i^{(k)}(a) = \mathbb{1}[Y_i(a) = k]$ indicate whether individual $i$ is diagnosed with disease type $k$ under $A_i = a$. We focus on the common scenario of 2 CRC subtypes (Carr et al., 2018; Wang et al., 2021). Some disease heterogeneity studies involve more than 2 subtypes. Web Appendix A extends the framework to $K \geq 2$ subtypes. Because the subtypes are mutually exclusive, $\sum_{k=1}^{2} Y_i^{(k)}(a) = 1$. In our study, subtype 1 is non-MSI-high CRC, and subtype 2 is MSI-high CRC. So, for example, $Y_i^{(1)}(0) = 1$ if individual $i$ would have been diagnosed with non-MSI-high CRC by age 70 had they not been smoking by age 60.

We assume there is no interference between individuals and that there are no multiple versions of the exposure leading to different potential outcomes. Under these assumptions, known collectively as the stable unit treatment value assumption (SUTVA), the observed disease status for individual $i$, hereafter denoted by $Y_i$, observed exposure value $A_i$, equals the appropriate potential outcome for that individual. That is, $Y_i = Y_i(A_i)$ (and $Y_i^{(k)} = Y_i^{(k)}(A_i)$). The latter is also known as the consistency assumption.

We assume weak ignorability, namely that rich enough data were collected as part of the NHS and HPFS cohorts such that for a vector of measured covariates $X$, we have $A \perp (a) \mid X$ for $a = 0, 1$. We also assume positivity, meaning that the exposure assignment probability $\pr(A = a \mid X = x) = 1$ for $a = 0, 1$ and all $x$.

In summary, the observed data consist of $n$ i.i.d. copies sampled from the distribution of $(X, A, Y^{(1)}, Y^{(2)})$, which can be equivalently represented as $(X, A, Y)$. 
4 REVIEW OF EXISTING APPROACHES

A principled approach to causal inference typically starts with a definition of causal effects of interest, followed by the assumptions needed for their identification from observed data, and only then statistical models are discussed. However, as the existing approaches employed in practice for studying etiologic heterogeneity are model-based and are not grounded in the causal inference paradigm, we start with reviewing these models (Section 4.1). Then, we discuss relevant causal effects arising in the presence of competing events (Section 4.2).

4.1 Statistical models for disease heterogeneity

Let \( \pi_k(a, x) = \Pr(Y = k|A = a, X = x) \), \( k = 0, 1, 2 \), be the probability of subtype \( k \) disease (or being CRC-free for \( k = 0 \)) among those with exposure \( A = a \) and confounders \( X = x \). Let also \( \pi_0(a, x) = (\pi_0(a, x), \pi_1(a, x), \pi_2(a, x)) \). A commonly used approach for studying etiologic heterogeneity is built upon the multinomial regression model:

\[
\pi_k(a, x) = \frac{\exp(a_k + \beta_k a + \gamma_k^T x)}{1 + \sum_{j=1}^{2} \exp(a_j + \beta_j a + \gamma_j^T x)},
\]

(1)

for \( k = 1, 2 \), and \( \pi_0(a, x) = 1 - \sum_{k=0}^{2} \pi_k(a, x) \). Under this model, \( \exp(\beta_k) \) is the odds ratio (OR) between the exposure \( A \) and the observed subtype-specific outcome \( Y(k) \) compared with not having the disease, that is, \( \exp(\beta_k) = \gamma_k/(1 - \gamma_k) \), and approximates the risk ratio (RR) for rare diseases. Model (1) was also extended for studies within which subtypes were not a priori known and may be defined using multiple, possibly mismeasured, biomarkers (Nevo et al., 2016). See Wang et al. (2016) for a survey of related methods.

Researchers are often interested in whether there is heterogeneity in the effects of \( A \) across disease subtypes. When the null hypothesis \( H_0: \beta_1 = \beta_2 \) is rejected, researchers conclude that the exposure \( A \) contributes to the risk of at least one subtype differently from its contribution to the risk of other subtypes.

Sun et al. (2017) proposed an alternative approach that postulates 2 logistic regressions, each comparing a subtype \( k \) disease to the alternative of being free of that subtype. Formally,

\[
\pi_k(a, x) = \frac{\exp(\alpha_k + \tilde{\beta}_k a + \tilde{\gamma}_k^T x)}{1 + \exp(\alpha_k + \tilde{\beta}_k a + \tilde{\gamma}_k^T x)},
\]

(2)

for \( k = 1, 2 \). Under this model, \( \exp(\tilde{\beta}_k) \) is the OR between the exposure \( A \) and the observed indicator of subtype-specific outcome \( Y(k) \); that is, \( \exp(\tilde{\beta}_k) = \gamma_k/(1 - \gamma_k) \). Because fitting these models separately may result in \( \sum_{k=0}^{2} \pi_k \neq 1 \), Sun et al. (2017) proposed constrained Bayesian estimation, restricting the posterior distribution to parameter values such that \( \sum_{k=0}^{2} \pi_k = 1 \). Comparing models (1) and (2), it has been recently argued (Sun et al., 2017) that variation in \( \beta_k \) across the subtypes (across \( k \)) is not a satisfactory measure of etiologic heterogeneity because the parameter space of a multinomial regression model requires a risk factor for 1 subtype to be a risk factor for the other subtypes, and that \( \tilde{\beta}_k \) may be preferred, although this issue was also debated (Begg et al., 2018; Sun et al., 2018).

4.2 Causal effects for competing risks

To fix ideas, consider the effect of \( A \) on \( Y(1) \). A naive approach that does not correspond to a causal effect compares the subtype proportion of cases at each exposure level only among individuals who are free of other disease subtypes. When confounders \( X \) are adjusted for, the statistical estimand on the difference scale is

\[
\mathbb{E}_X \left[ \mathbb{E}[Y(1)|A = 1, X, Y(2) = 0] \right] - \mathbb{E}_X \left[ \mathbb{E}[Y(1)|A = 0, X, Y(2) = 0] \right].
\]

(3)

and can be analogously defined on the RR or OR scales. Because model (1) is equivalent to 2 separate logistic regression models, each comparing 1 disease subtype to the disease-free individuals (Wang et al., 2016), then \( \exp(\beta_k) \) is the OR analog of (3) and approximates the RR. However, (3) does not correspond to a causal effect because of selection bias due to common causes of the different subtypes (Nevo et al., 2021). Importantly, because MSI-high and non-MSI-high are subtypes of the same disease, CRC, they are likely to have unknown and/or unmeasured common causes that can be independent of the exposure.

Young et al. (2020) considered causal estimands in the presence of competing events. The total effect, \( TE(1) = \mathbb{E}[Y(1)|A = 1] - \mathbb{E}[Y(1)|A = 0] \), contrasts the population-level subtype 1 risk under exposure versus non-exposure. In the presence of measured confounders, this effect is identified under SUTVA, weak ignorability, and positivity by

\[
TE(1) = \mathbb{E}_X \left[ \mathbb{E}[Y(1)|A = 1, X] \right] - \mathbb{E}_X \left[ \mathbb{E}[Y(1)|A = 0, X] \right].
\]

(4)

Unlike (3), the definition of \( TE(1) \) does not condition on the absence of other subtypes. Under model (2), \( \exp(\tilde{\beta}_k) \) is the total effect for subtype \( k \) on the OR/RR scales.

However, despite \( TE(1) \) being a valid causal effect, it does not suffice to describe exposure effects in the presence of etiologic disease heterogeneity. The total effect of \( A \) on \( Y(1) \) includes the effect caused by changes in \( Y(2) \), that is, \( TE(1) \) is sensitive to the phenomenon of exposure-induced subtype switching. For example, assume some individuals would have been diagnosed with subtype 1 under \( A = 0 \) and would have been diagnosed with subtype 2 under \( A = 1 \), while the rest of the population will be free of subtype 1 for all \( A \) values. The total effect on subtype 1 will be negative, and it would appear as if the exposure protects from subtype 1, even though the “protected” individuals would still have the disease, just of another subtype.

The direct effect compares subtype 1 rates under 2 hypothetical joint interventions. The first sets the exposure to \( A = 1 \) and eliminates the possibility of subtype 2 disease, and the second sets the exposure to \( A = 0 \) and also eliminates subtype 2 risk.

This estimand is irrelevant for our study because an intervention eliminating the risk of a particular CRC subtype is not currently available and is not expected to be in the near future.

Separable effects (Stensrud et al., 2020) rely on a conceptual separation of the exposure into the components affecting each disease subtype. Such separation requires profound understanding of the mechanisms leading to each CRC subtype.

To summarize this section, all 3 families of estimands (total, direct, and separable effects) are well defined mathematically. However, both the direct and separable effects correspond to interventions that are not currently conceivable. While the
total effect corresponds to standard interventions, it may not provide comprehensive information as it is susceptible to exposure-induced subtype switching.

5 THE SUBTYPE-FREE AVERAGE CAUSAL EFFECT

We propose a new estimand, inspired by the SACE (Rubin, 2006). We define the SF-ACE on the difference scale by

$$SF-ACE_D^{(1)} = \frac{E[Y(1) - Y(0)|Y(2) = 0]}{E[Y(0)|Y(2) = 0]} = 0,$$

and analogously define

$$SF-ACE_{RR}^{(1)} = \frac{E[Y(1)|Y(0) = Y(2) = 0]}{E[Y(0)|Y(2) = 0]} = 0$$

to be the SF-ACE on the RR scale. The SF-ACE is the average causal effect among the individuals that would have been free of disease subtype 2 under either exposure level (i.e., regardless of smoking status). Let SF-ACE is and SF-ACE RR be the analogous average causal effects of A on Y (2) among the individuals that would have been free of disease subtype 1 under either exposure level.

For heterogeneity of the causal effects across disease subtypes, one can use, for example, \( \theta_0 = SF-ACE_D^{(1)} - SF-ACE_D^{(2)} \) and test the null hypothesis of \( H_0: \theta_0 = 0 \) versus the alternative of \( H_1: \theta_0 \neq 0 \).

A caveat of defining smoking history as a binary exposure is the potential violation of the assumption of no multiple versions of the treatment leading to different potential outcomes, as \( Y(k) (1) \) may depend on the frequency and amount of smoking by age 60 for ever-smokers. This choice to treat smoking as binary rather than continuous is in line with previous studies (Carr et al., 2018; Amitay et al., 2020), and is motivated by the potential substantial measurement error of smoking habits as captured in biannual questionnaires.

Nevertheless, we present an alternative. With a slight abuse of notation, let \( Y_i(\epsilon) (c) \) be CRC subtype k status if individual \( i \) had smoked \( C = \epsilon \) cumulative pack-years by age 60. Let \( Y(\epsilon) = 1[\exists c \in C : Y(\epsilon) (c) = 1] \) be the indicator that a person would develop CRC subtype k under at least 1 smoking level \( c \in C \), where \( C \) is the domain of C. We define the Subtype-Free Population Attributable Fraction (SF-PAF) by

$$SF-PAF^{(1)} = \frac{E[Y(1) - Y(0)|Y(2) = 0]}{E[Y(1)|Y(2) = 0]},$$

and analogously for \( SF-PAF^{(2)} \). PAF measures are common in cancer epidemiology, specifically for quantifying the impact of smoking on cancer incidence (Park et al., 2014; Poole, 2015; Mansournia and Altman, 2018). \( SF-PAF^{(1)} \) is the proportion of CRC subtype 1 cases attributable to smoking among those that would have been free of CRC subtype 2 regardless of their smoking level. Modifications needed to the assumptions, as well as the generalization of the results presented later in the paper, are provided in Web Appendix A.

5.1 Interpretation of the SF-ACE

As explained in Section 4.2, competing events present a challenge in defining causal estimands that shed light on the scientific question of interest. The SF-ACE presents another source of information. A positive SF-ACE means that among those free of MSI-high CRC regardless of smoking status by age 60, more non-MSI-high CRC cases are expected if the entire subpopulation had smoked compared to if they had not.

Therefore, the SF-ACE offers a way to circumvent non-zero causal effects due to exposure-induced subtype switching. The SF-ACE can also help assess heterogeneity in the exposure effect across different subtypes, which is a common goal; see Ogino et al. (2016) for a review. Informally, such studies explore whether, and to what extent, the causal effect of an exposure differs across subtypes. The goal of these studies is ultimately to better understand the disease formation process. However, as reviewed in Section 4, the methods used in practice do not correspond to formal causal effects, and existing causal estimands do not provide answers to these questions. The SF-ACE brings us 1 step closer to understanding the heterogeneity of the causal effect of an exposure across subtypes.

Nevertheless, the above appealing advantages do not come without a price. One disadvantage is that the SF-ACE is defined in a latent subset of the population. This is a general problem with principal stratification estimands. A related issue is that when contrasting SF-ACE and SF-ACE, we contrast effects defined in 2 overlapping, but not identical, sub-populations. The SF-ACE is defined within the sub-population \( Y(1) (0) = Y(2) (1) = 0 \), and the SF-ACE is defined within the sub-population \( Y(1) (0) = Y(1) (1) = 0 \).

Thus, \( \theta_0 \) might be non-zero due to the different populations, and as a result, the exact source of heterogeneity is not revealed. Instead, a non-zero \( \theta_0 \) presents evidence of heterogeneity and calls for further investigation. In rare disease scenarios, this problem is less concerning because the 2 sub-populations \( Y(1) (0) = Y(2) (1) = 0 \) and \( Y(1) (0) = Y(1) (1) = 0 \) have substantial overlap; those who would be CRC-free under any exposure value belong to both groups and comprise the vast majority of the population.

Finally, the SF-ACE can be viewed as a special case of the SACE with 2 important properties. First, unlike most of the SACE literature, the outcome is binary. Second, we study 2 SACEs simultaneously, where each subtype takes on the role of the main event or the truncating event.

5.2 Identifiability

The SF-ACE is not identifiable from the data under the standard assumptions of SUTVA, weak ignorability, and positivity. One commonly made assumption when targeting the SACE is monotonicity (Zhang and Rubin, 2003), stating that treatment cannot hurt survival. In our study, the definition and plausibility of monotonicity-like assumptions are more nuanced, and, as it turns out, result in new insights regarding principal causes and their identification. We start with the definition of Subtype Monotonicity (S-Monotonicity).
Assumption 1 Subtype Monotonicity (S-Monotonicity): $Y^{(k)}_i(0) \leq Y^{(k)}_i(1)$.

S-Monotonicity states that an individual who would have been diagnosed with subtype $k$ under no exposure (never smoking) would have also been diagnosed with subtype $k$ under exposure (ever smoking). Therefore, an individual who would have been free of subtype $k$ under exposure would have also been free of subtype $k$ under no exposure. In other words, S-Monotonicity asserts that the exposure cannot prevent the subtype $k$-specific disease. The analogous assumption, when smoking is coded as ordinal or continuous with a domain $C$, is that if $Y^{(k)}_i(c) = 1$ then $Y^{(k)}_i(c') = 1$ for all $c' > c$ (Web Appendix A).

Importantly, S-Monotonicity is a subtype-specific assumption. It can hold for both subtypes, for 1 subtype, or for none of them. When assumed for subtype $k$, S-Monotonicity constrains the signs of SF-ACE $(k)$ and $TE^{(k)}$, while leaving the signs of the effects for the other subtype undetermined. S-Monotonicity for subtype $k$ precludes 1 direction of exposure-induced subtype switching, but switching in the other direction is possible. Interestingly, under S-Monotonicity for both subtypes, the SF-ACE $^{(k)}_{RR}$ equals the total effect $TE^{(k)}_{RR}$ and is identifiable from the data under SUTVA, weak ignorability, and positivity. Web Appendix A provides the technical details. The following proposition establishes identification of SF-ACE $^{(1)}_D$.

Proposition 1 Under SUTVA, weak ignorability, positivity, and S-Monotonicity for subtypes $k = 1, 2$, the SF-ACE $^{(1)}_D$ is identifiable from the observed data by

$$SF-ACE^{(1)}_D = \frac{E[\pi_i(1,X) Y^{(1)}_i] - E[\pi_i(0,X) Y^{(1)}_i]}{E[\pi_i(1,X)] - E[\pi_i(0,X)]},$$

using standardization, or via IPTW by

$$SF-ACE^{(1)}_D = \frac{E[\frac{1}{n} \sum_{j=1}^{n} I(\pi_{ij}^{(1)}(X)) Y^{(1)}_{ij}] - E[\frac{1}{n} \sum_{j=1}^{n} I(\pi_{ij}^{(1)}(X)) Y^{(2)}_{ij}]}{1 - E[\frac{1}{n} \sum_{j=1}^{n} I(\pi_{ij}^{(1)}(X))]}.$$  

Analogous expressions for $SF-ACE^{(2)}_D$, $\theta_D$, and for causal contrasts on the RR scale are given in Web Appendix A. Proposition 1 reveals that under S-Monotonicity for both subtypes, the SF-ACE $^{(1)}_D$ will be close to the $TE^{(1)}_{DR}$ when the prevalence of the other subtype under exposure is close to 0; we further study this issue in simulations in Section 6.

Identification of the SF-ACE is based upon identifying terms of the form $Pr[Y^{(k)}(a) = 1]$. Identification based on augmented-IPTW (Lunceford and Davidian, 2004; Bang and Robins, 2005) is also possible and is omitted from Proposition 1 only for brevity. We discuss augmented-IPTW estimators in Section 5.6.

The identification approach for the SF-ACE leverages the competing event nature of the subtypes, specifically that $Y^{(k)}$ are binary and mutually exclusive. Consequently, identification is possible without requiring assumptions beyond monotonicity, though point identification requires monotonicity to hold for both subtypes. For SACE identification, monotonicity of the survival outcome, or other assumptions about survival status under both treatment values, is typically assumed alongside an assumption relating the potential survival outcomes to the continuous outcome, sometimes called principal ignorability (see, eg, Hayden et al., 2005; Ding and Lu, 2017; Zehavi and Nevo, 2023). Assuming monotonicity for the survival outcome (ie, for 1 subtype only) is insufficient for identification (Zhang and Rubin, 2003). As a special case of the SACE, the SF-ACE $(1)$ can be identified under S-Monotonicity for subtype 2 if a principal ignorability assumption is added. However, this assumption is unlikely to hold because unmeasured common causes of both subtypes are likely (Nevo et al., 2021).

5.3 Relaxing S-Monotonicity

Assuming S-Monotonicity for both subtypes might be too restrictive, even if we believe the exposure cannot protect from the outcome, as S-Monotonicity disallows subtype switching (for that particular subtype). Therefore, we consider a weaker assumption, termed Disease Monotonicity (D-Monotonicity), that states that an individual who would have been diagnosed with subtype $k$ under no exposure would have also been diagnosed with the disease under exposure, but not necessarily with the same subtype.

Assumption 2 Disease Monotonicity (D-Monotonicity):

$$Y^{(k)}_i(0) \leq \max\{Y^{(1)}_i(1), Y^{(2)}_i(1)\}.$$  

D-Monotonicity keeps the premise that the exposure cannot protect from the disease while being less restrictive than S-Monotonicity.

For each disease subtype, we may assume S-Monotonicity, D-Monotonicity, or neither. With 2 disease subtypes, we have 8 different combinations of possible assumptions. Under each combination, certain potential outcome profiles $\{Y^{(1)}_i(0), Y^{(1)}_i(1), Y^{(2)}_i(0), Y^{(2)}_i(1)\}$ are assumed to be absent from the population. For example, under S-Monotonicity for subtype 1, there are no individuals with $\{Y^{(1)}_i(0), Y^{(1)}_i(1), Y^{(2)}_i(0), Y^{(2)}_i(1)\} = \{1, 0, 0, 1\}$.

Table 1 presents the 9 possible potential outcome profiles and specifies whether each profile exists under different combinations of assumptions. It highlights the explicit implications of different monotonicity assumptions on potential outcomes. For example, the last row shows individuals diagnosed with subtype 2 when unexposed, and with subtype 1 when exposed. This profile is impossible under S-Monotonicity for disease subtype 2. Table A.1 extends Table 1 when either S-Monotonicity or D-Monotonicity is assumed for 1 subtype only. The observed data do not provide enough information to classify each individual into a specific profile. Table A.2 specifies the possible profiles for each participant according to their observed $(A, Y^{(1)}, Y^{(2)})$ under different combinations of assumptions.

As previously shown, under S-Monotonicity for both subtypes, $SF-ACE^{(1)}_D$ and $SF-ACE^{(2)}_D$ are identifiable from the observed data. If S-Monotonicity is replaced with D-Monotonicity for at least 1 subtype, the causal effects are no longer identifiable, and $SF-ACE^{(k)}_{RR}$ is not equal to the $TE^{(k)}_{RR}$. An alternative to making strong assumptions is either calculating large-sample bounds (partial identification) under weaker assumptions or performing...
a sensitivity analysis based on unidentifiable sensitivity parameters (Zhang and Rubin, 2003; Ding and Lu, 2017; Nevo and Gorfine, 2022). We now develop these 2 approaches.

5.4 Large-sample bounds

We present bounds for effects on the difference scale under standardization. Bounds for $\text{SF-ACE}_{D}^{(k)}$ and $\text{SF-PAF}$, as well as those derived using (augmented) IPTW, are provided in Web Appendix A. We begin with bounds obtained without any monotonicity assumptions.

**Proposition 2** Under SUTVA, weak ignorability, and positivity, $\text{SF-ACE}_{D}^{(1)}$ is partially identifiable from the observed data by $L_{D,\text{stand}}^{(1)} \leq \text{SF-ACE}_{D}^{(1)} \leq U_{D,\text{stand}}^{(1)}$ where

$$L_{D,\text{stand}}^{(1)} = \frac{\max \left\{ \hat{E}[Y_{1}(1)] - \hat{E}[Y(0)] \right\}}{1 - \max \left\{ \hat{E}[Y(1)] \right\}}$$

$$U_{D,\text{stand}}^{(1)} = \frac{\min \left\{ \hat{E}[Y_{1}(1)] - \hat{E}[Y(0)] \right\}}{1 - \min \left\{ \hat{E}[Y(1)] \right\}}$$

The bounds for $\text{SF-ACE}_{D}^{(2)}$ are analogous.

More informative bounds can be constructed by incorporating additional assumptions. We present bounds under D-Monotonicity for subtype 1 and S-Monotonicity for subtype 2, as this combination of assumptions is more plausible in our study. The S-Monotonicity assumption may be too restrictive for the non-MSI-high subtype (subtype 1), given the limited information about the smoking effect strength on non-MSI-high CRC, while smoking’s effect on MSI-high CRC is known to be strong (Amitay et al., 2020). Therefore, we allow exposure-induced subtype switching from the non-MSI-high subtype.

**Proposition 3** Under SUTVA, weak ignorability, positivity, D-Monotonicity for subtype 1, and S-Monotonicity for subtype 2, $\text{SF-ACE}_{D}^{(k)}$ are partially identifiable from the observed data by $L_{D,\text{stand}}^{(k)} \leq \text{SF-ACE}_{D}^{(k)} \leq U_{D,\text{stand}}^{(k)}$ where

$$L_{D,\text{stand}}^{(k)} = \frac{\max \left\{ \hat{E}[Y_{1}(1)] - \hat{E}[Y_{0}(1)] \right\}}{1 - \max \left\{ \hat{E}[Y_{0}(1)] \right\}}$$

$$U_{D,\text{stand}}^{(k)} = \frac{\min \left\{ \hat{E}[Y_{1}(1)] - \hat{E}[Y_{0}(1)] \right\}}{1 - \min \left\{ \hat{E}[Y_{0}(1)] \right\}}$$

Comparing $L_{D,\text{stand}}^{(k)}$ and $U_{D,\text{stand}}^{(k)}$ to $L_{D,\text{stand}}^{(1)}$ and $U_{D,\text{stand}}^{(1)}$, the denominator is point identified and need not to be bounded, and so is the first term in the numerator, while the second term is bounded, possibly more sharply.

5.5 Sensitivity analysis

We present a sensitivity analysis, commonly used when studying causal effects under untestable assumptions. Unlike the bounds, the sensitivity analysis maps sensitivity parameter values to the estimand of interest. Our approach is motivated by Ding and Lu (2017) and Zehavi and Nevo (2023), who express the causal estimand in terms of the observed data and an unidentifiable
but meaningful sensitivity parameter. Let $\lambda_1 = \Pr[Y^{(1)}(1) = 1|Y^{(1)}(0) = 1]$ be the proportion of individuals who would have been diagnosed with disease subtype 2 when exposed, out of the individuals who would have been diagnosed with disease subtype 1 when unexposed. That is, $\lambda_1$ is the probability of exposure-induced subtype switching from subtype 1 to subtype 2. Note that $\lambda_1 = 0$ under S-Monotonicity for subtype 1. Let also $\lambda_2 = \Pr[Y^{(1)}(1) = 1|Y^{(2)}(0) = 1]$ be the subtype-switching probability from subtype 2 to subtype 1. While $\lambda_1$ and $\lambda_2$ are not identifiable from the observed data, they can be bounded. Specifically, because $\lambda_1 \leq \min \left\{1, \frac{\Pr[Y^{(1)}(1) = 1]}{\Pr[Y^{(1)}(0) = 1]} \right\}$, a data-driven upper bound for $\lambda_1$ is $\min \left\{1, \frac{\sum_{i \in [1]} \Pr[Y^{(1)}(1)]}{\sum_{i \in [1]} \Pr[Y^{(1)}(0)]} \right\}$. Similarly, $\lambda_2 \leq \min \left\{1, \frac{\sum_{i \in [1]} \Pr[Y^{(2)}(1)]}{\sum_{i \in [1]} \Pr[Y^{(2)}(0)]} \right\}$.

The following proposition demonstrates that the subtype-switching probabilities play a key role in the identification of the SF-ACE.

**Proposition 4** Under SUTVA, weak ignorability, positivity, and D-Monotonicity for both subtypes, the SF-ACE$^{(1)}$ is identifiable from the data by

$$SF-ACE^{(1)} = \frac{\sum_{i \in [1]} \sum_{j \in [1]} \Pr[Y^{(1)}(1), Y^{(2)}(1)]}{\sum_{i \in [1]} \sum_{j \in [1]} \Pr[Y^{(1)}(0), Y^{(2)}(0)]}$$

**The identification of SF-ACE$^{(2)}$ is analogous.**

If one further assumes S-Monotonicity for subtype $k$, then identification formulas for the SF-ACEs are obtained by setting the appropriate $\lambda_k$ to 0 in (8) and (9). A second remark is that, technically, results (8) and (9) require D-Monotonicity for subtype 2 only, and similarly, the identification for subtype 2 requires D-Monotonicity for subtype 1 only. Web Appendix A presents results for RR scale, and under other assumption combinations.

As a sensitivity analysis, researchers may set $\lambda_1$ or $\lambda_2$ to 0 and estimate the SF-ACE$^{(k)}$ as a function of $\lambda_1$ and/or $\lambda_2$. Since $\lambda_k$ involves a cross-world quantity, empirical information about its value is, and will always be, unavailable to researchers. This limitation of the sensitivity analysis approach may prompt researchers to prefer the large-sample bounds, which do not have this issue. Nevertheless, this analysis can reveal the sensitivity of the results to varying levels of exposure-induced subtype switching.

### 5.6 Estimation

Recall that for each individual $i$, the observed data are $(A_i, X_i, Y_i)$ or equivalently $(A_i, X_i, Y^{(1)}(1), Y^{(2)}(1))$. Estimation of SF-ACE$^{(k)}$ from the data, or the bounds and sensitivity analyses, requires estimation of $\pi_k(a, x)$, $k = 1, 2$, $a = 0, 1$ for the standardization-based formulas, or $\epsilon(x)$ for the IPTW. Expectations over $X$ or $(X, A, Y^{(k)})$ are estimated by taking sample analogs.

Non-parametric estimation of $\pi_k(a, x)$ or $\epsilon(x)$ when the vector of measured confounders $X$ is of high dimension or includes many continuous variables could be a difficult task that will yield an estimator with a too large variance to be able to derive conclusions from the data. Therefore, parametric estimation might be attractive, and also aligns with the models used in practice and that were presented in Section 4.1.

Starting with standardization, let $\pi(a, x; \eta_i) = (\pi_0(a, x; \eta_i), \pi_1(a, x; \eta_i), \pi_2(a, x; \eta_i))$ be the parametric model for $\pi(a, x)$ with parameter vector $\eta_i$. This could be, for example, the regression models (1) or (2). Upon estimating $\hat{\eta}_i$, the estimated probabilities $\pi(a, X_i; \hat{\eta}_i)$ are calculated in the entire sample, for $a = 0, 1$, before plugged in the relevant identification formulas. For example, the proposed SF-ACE$^{(1)}$ estimator is

$$SF-ACE^{(1)} = \frac{\sum_{i \in [1]} \pi_1(1, X_i; \hat{\eta}_i) - \sum_{i \in [1]} \pi_1(0, X_i; \hat{\eta}_i)}{n - \sum_{i \in [1]} \pi_1(1, X_i; \hat{\eta}_i)}$$

Turning to IPTW, let $e(x; \eta_2)$ be a parametric model for $e(x)$, with parameter vector $\eta_2$. Then, after estimating $\hat{\eta}_2$, the SF-ACE$^{(1)}$ estimator is obtained by

$$SF-ACE^{(1)} = \frac{\sum_{i \in [1]} \frac{A_i Y^{(1)}(1)}{e(X_i; \eta_2)} - \sum_{i \in [1]} \frac{(1-A_i) Y^{(1)}(1)}{e(X_i; \eta_2)}}{n - \sum_{i \in [1]} \frac{A_i Y^{(1)}(1)}{e(X_i; \eta_2)}}$$

and similar estimators for SF-ACE$^{(2)}$ and $\theta_D$, or for effects on other scales, can be obtained.

The maximum likelihood estimators $\hat{\eta}_1$, or $\hat{\eta}_2$, are consistent and asymptotically normal. By the continuous mapping theorem and the delta method, the SF-ACE estimators are also consistent and asymptotically normal (Web Appendix A). In practice, the bootstrap can be used to estimate standard errors (SEs) and construct confidence intervals (CIs).

The proposed estimators are consistent only if their respective models are correctly specified. Therefore, we extend our estimation approach to DR estimation (Robins et al., 1994; Lunceford and Davidian, 2004). We construct an estimator that will be consistent whenever at least 1 of the 2 models, $\pi(a, x; \eta_1)$ or $e(x; \eta_2)$, but not necessarily both, is correctly specified. DR estimators are frequently used to reduce dependency on model choice. Furthermore, a DR estimator constructed as an augmented-IPTW estimator is also more efficient than IPTW estimators and has desirable theoretical properties (Robins et al., 1994).

Let $\hat{\eta}_1$ and $\hat{\eta}_2$ be the estimators obtained for $\eta_1$ and $\eta_2$. As can be seen from the proof of Proposition 1 (Web Appendix A), under S-Monotonicity, the SF-ACEs and $\theta$ are functions of $Pr[Y^{(k)}(a) = 1]$, $a = 0, 1$, $k = 1, 2$. For example, SF-ACE$^{(1)}$ is equal to $\frac{Pr[Y^{(1)}(1) = 1] - Pr[Y^{(1)}(0) = 1]}{Pr[Y^{(1)}(1) = 0]}$. Under weak ignorability and positivity, a DR augmented-IPTW estimator for $Pr[Y^{(1)}(1) = 1]$ is

$$\frac{1}{n} \sum_{i \in [1]} \left[ \frac{A_i Y^{(1)}(1)}{e(X_i; \eta_2)} - \frac{(1-A_i) Y^{(1)}(1)}{e(X_i; \eta_2)} \right]$$

**Biometrics, 2025, Vol. 81, No. 1 • 7**
missing subtype (Liu et al., 2018). We first fitted a model for the probability of observing the subtypes (Table C.9). The model included additional variables available for all CRC cases (eg, tumor stage and size). Then, for each CRC case with a known subtype, we calculated the weight as the reciprocal of the probability of observing the subtype. The weights were truncated at the 99th percentile.

We estimated the TE of smoking on each disease subtype by standardization (4), using a multinomial regression model (1) (Table C.10). Point estimates indicated smoking increases the risk of both CRC-MSI subtypes (Table 2). The effect on the non-MSI-high subtype was larger on the difference scale but smaller on the RR scale, in agreement with the literature, as MSI-high is rarer—leading to a higher RR but fewer excess cases due to smoking.

Turning to the SF-ACEs, we first considered an analysis under S-Monotonicity for both subtypes, using multinomial regression (1) for the outcome model and logistic regression for the exposure model (Table C.11). The DR estimator produced larger estimated effects and had the lowest SEs (Table C.12). On the difference scale, the effects were 24% (161.3 versus 129.9 cases for non-MSI-high) and 14% (37.2 versus 32.7 cases for MSI-high) larger under the DR estimator compared with standardization. In the United States alone, there are ~4.25 million 60-years-olds (U.S. Census Bureau, 2023), meaning the additional 31.4 and 5.5 cases per 100 000 according to the DR estimator (compared with standardization) correspond to ~1330 and 230 more cases, annually.

The only significant smoking SF-ACE effect was for the non-MSI-high subtype, which estimated additional 161.3 cases per 100 000 people (95% CI: 8.6-314.0), had the entire principal stratum free of MSI-high-subtype were smokers (compared with had they all been non-smokers). The RR estimate was 1.26 (95% CI: 1.03-1.58). The estimated effect on MSI-high subtype was smaller on the difference scale (37.2 additional MSI-high cases per 100 000 people, 95% CI: −9.6 to 84.0) and larger on the RR scale (1.48, 95% CI: 0.64-2.32) than the corresponding effects on the non-MSI-high subtype. The null hypothesis of $H_0 : \theta_D = 0$ was not rejected ($P = .148$). The estimated SF-PAF was 15.7% (95% CI: [4.2%-27.5%]) and 23.6% (95% CI: [0.0%-52.4%]) for non-MSI-high and MSI-high subtypes, respectively.

As discussed in Section 5.4, the S-Monotonicity assumption may be too restrictive for the non-MSI-high subtype. We therefore also present bounds (Table 3) and the results of the sensitivity analysis. Without monotonicity assumptions, the bounds are only informative for the effect on the non-MSI-high subtype and on the difference and RR scales. When S-Monotonicity is assumed for the non-MSI-high subtype and D-Monotonicity for the MSI-subtype, the bounds for effect on the non-MSI-high subtype are more tight, reflecting an effect that might be larger than the effect estimated under S-Monotonicity for both subtypes.

For the sensitivity analysis (Proposition 4), we fixed $\lambda_2 = 0$ and varied the values of $\lambda_1$ between 0 and 0.95% CIs and $P$-values were calculated using the bootstrap with 200 repetitions.

To minimize the impact of potential selection bias due to missing subtype status, we used inverse probability weighting for
TABLE 2 Estimated smoking total and SF-ACE effects on the 2 CRC subtypes under S-Monotonicity for both subtypes

| Subtype          | Effect       | Estimate | SE | 95% CI      |
|------------------|--------------|----------|----|-------------|
| non-MSI-high     | $T_E D$      | 161.1    | 78.2 | [7.8, 314.4] |
|                  | $SF-ACE D$   | 161.3    | 77.9 | [8.6, 314.0] |
|                  | $SF-ACE RR$  | 1.26     | 0.14 | [0.98, 1.54] |
|                  | SF-PAF       | 15.7%    | 6.4% | [4.2%, 27.5%] |
|                  | $T_E D$      | 36.9     | 24.3 | [−10.7, 84.5] |
|                  | $SF-ACE D$   | 37.2     | 23.9 | [−9.6, 84.0] |
|                  | $SF-ACE RR$  | 1.48     | 0.43 | [0.64, 2.32]  |
|                  | SF-PAF       | 23.6%    | 14.8%| [0.0%, 52.4%] |

DR estimators are presented. Effects on the difference scale are presented per 100,000 people. Under these assumptions, the SF-ACE D is the same as the $TE D$, $SE$: estimated standard error; 95% CI: 95% confidence interval estimated via percentile bootstrap.

Abbreviations: CRC, colorectal cancer; MSI, microsatellite instability.

TABLE 3 Bounds for SF-ACE on the difference scale, risk ratio scale, and the PAF for the effect of smoking on the 2 CRC subtypes under no assumptions and under S-Monotonicity for MSI-high and D-Monotonicity for non-MSI-high

| Subtype          | Effect       | Without Monotonicity assumptions | Bounds Non-MSI-high D-Monotonicity | MSI-high S-Monotonicity |
|------------------|--------------|----------------------------------|-----------------------------------|-------------------------|
| non-MSI-high     | $SF-ACE D$   | [84.3, 275.4]                    | [161.3, 275.2]                    |                         |
|                  | $SF-ACE RR$  | [1.14, 1.54]                     | [1.26, 1.54]                     |                         |
|                  | SF-PAF       | [0%, 100%]                       | [0%, 100%]                       |                         |
| MSI-high         | $SF-ACE D$   | [−77.5, 115.4]                   | [−77.5, 115.4]                   |                         |
|                  | $SF-ACE RR$  | [0.00, ∞]                        | [0.00, 1.48]                     |                         |
|                  | SF-PAF       | [0%, 100%]                       | [0%, 23.6%]                      |                         |

Effects on the difference scale are presented per 100,000 people.

Abbreviations: CRC, colorectal cancer; MSI, microsatellite instability; PAF, population attributable fraction.

FIGURE 1 Sensitivity analysis for causal effects under S-Monotonicity for MSI-high (subtype 2) and D-Monotonicity for non-MSI-high (subtype 1) colorectal cancer (CRC) subtypes. The figure presents doubly robust (DR) estimates as a function of $\lambda_1$. The shadows are 95% Wald-type confidence intervals (CIs) are connected continuously for clarity of presentation.

There were 17,854 study participants for whom the last available data were before age 70 (Section 2). We treated these people as CRC-free at age 70. To assess the impact of this decision, we repeated our analyses while removing these people. In an additional analysis, we again removed these people but used inverse probability of weighting to take them into account. The results did not change materially, although the estimated effect sizes and the SEs were generally larger (Tables C.14 and C.15).

To summarize, our analysis provided some evidence for smoking effect on both CRC-MSI subtypes and evidence for heterogeneity under slight 1-sided subtype switching.

8 DISCUSSION

Etiologic heterogeneity is central to our understanding of how diseases evolve. Existing approaches often overlook causal estimands and identifiability assumptions. Recent progress in causal inference methodology for competing events offers causal estimands that either provide limited insight into etiologic heterogeneity or rely on assumptions that are implausible for CRC subtypes. In this paper, we propose the SF-ACEs as causal estimands to study the effect of smoking on CRC subtypes defined by MSI status, complementing existing methods. This approach provides insights into heterogeneity in disease risk, which can inform prevention strategies and guide interventions. For instance, if an exposure impacts the risk of subtype only, and the prognosis of this subtype is worse, public health experts may prioritize interventions targeting that exposure for individuals at higher risk. Furthermore, it has been argued that focusing on non-MSI-high CRC when unexposed, increased, the estimated SF-ACE for non-MSI-high CRC increased and the estimated SF-ACE for MSI-high CRC decreased (Figure 1). Notably, for $\lambda_1 \geq 0.04$, $H_0 : \theta_D = 0$ was rejected. This result serves as evidence for heterogeneity in the SF-ACE as long as we believe subtype switching is possible and not extremely rare. Results on the RR scale and illustration of sensitivity analyses under other assumption combinations are provided in Web Appendix C.
subtype risks, rather than disease as a whole, provides stronger causal evidence and can lead to a deeper understanding of disease mechanisms (Ogino et al., 2016). This understanding can ultimately guide advancements in precision medicine and prevention strategies.

The SF-ACE analysis of the effect of smoking on CRC MSI subtypes treated smoking as a binary exposure. While this simplification reduces the impact of measurement error, it challenges the validity of both SUTVA and monotonicity assumptions. To address these issues, we also defined and studied the SF-PAF, which does not explore the full “dose-response” relationship of smoking but is widely used in practice.

An additional challenge with our approach using multinomial regression with cohort data is the issue of right censoring. To assess its impact, we considered a number of analyses. The censoring problem would have been more pronounced if follow-up was longer than 10 years. Therefore, considering time-to-event outcomes is an important topic of future research.

ACKNOWLEDGMENTS

We thank two anonymous reviewers, the Associate Editor, and the Co-Editor for many helpful comments and suggestions that improved the paper.

SUPPLEMENTARY MATERIALS

Supplementary material is available at Biometrics online.

Web Appendices, Tables, and Figures referenced in Sections 1–8, are available with this paper at the Biometrics website on Oxford Academic. The R package TheSFACE, available from CRAN on https://cran.r-project.org/web/packages/TheSFACE/index.html, implements our methodology (Sasson, 2022b). Reproducibility materials are available from GitHub https://github.com/amitSasson/SFACE.Reproduce (Sasson, 2022a). Code posted online on Oxford Academic includes simulations code, a synthetic dataset, and data analysis code for reproducibility of results in the paper.

FUNDING

This research was supported by the Israel Science Foundation (ISF grant No. 827/21); U.S. National Institutes of Health (NIH) grants R35 CA197735, R01 CA151993, UM1 CA186107, P01 CA87969, P01 CA55075, UM1 CA167552, U01 CA16755; and Cancer Research UK Grand Challenge Award (UK C10674/A27140).

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The data that support the findings in this paper are not publicly available due to privacy or ethical restrictions. A synthetic dataset and its analysis are available from GitHub https://github.com/amitSasson/SFACE_Reproduce (Sasson, 2022a).

REFERENCES

Amity, E. L., Carr, P. R., Jansen, L., Roth, W., Alwers, E., Herpel, E. et al. (2020). Smoking, alcohol consumption and colorectal cancer risk by molecular pathological subtypes and pathways. British Journal of Cancer, 122, 1604–1610.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61, 962–973.

Begg, C. B., Seshan, V. E. and Zabor, E. C. (2018). Re: “A multinomial regression approach to model outcome heterogeneity”. American Journal of Epidemiology, 187, 1129–1130.

Carr, P., Alwers, E., Bienert, S., Weberpals, J., Klooß, M., Brenner, H. et al. (2018). Lifestyle factors and risk of sporadic colorectal cancer by microsatellite instability status: a systematic review and meta-analyses. Annals of Oncology, 29, 825–834.

Ding, P. and Lu, J. (2017). Principal stratification analysis using principal scores. Journal of the Royal Statistical Society Series B: Statistical Methodology, 79, 757–777.

Hayden, D., Pailer, D. K. and Schoenfeld, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. Biometrics, 61, 305–310.

Liu, L., Nevo, D., Nishihara, R., Cao, Y., Song, M., Twombly, T. S. et al. (2018). Utility of inverse probability weighting in molecular pathological epidemiology. European Journal of Epidemiology, 33, 381–392.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine, 23, 2937–2960.

Mansournia, M. A. and Altman, D. G. (2018). Population attributable fraction. BMJ, 360, k757.

Nevo, D. and Gorfine, M. (2022). Causal inference for semi-competing risks data. Biostatistics, 23, 1115–1132.

Nevo, D., Nishihara, R., Ogino, S. and Wang, M. (2018). The competing risks Cox model with auxiliary case covariates under weaker missing-at-random condition of failure. Lifetime Data Analysis, 24, 425–442.

Nevo, D., Ogino, S. and Wang, M. (2021). Reflection on modern methods: causal inference considerations for heterogeneous disease etiology. International Journal of Epidemiology, 50, 1030–1037.

Nevo, D., Zucker, D. M., Tamimi, R. M. and Wang, M. (2016). Accounting for measurement error in biomarker data and misclassification of subtypes in the analysis of tumor data. Statistics in Medicine, 35, 5686–5700.

Ogino, S., Nishihara, R., VanderWeele, T. J., Wang, M., Nishi, A., Lochhead, P. et al. (2016). The role of molecular pathological epidemiology in the study of neoplastic and non-neoplastic diseases in the era of precision medicine. Epidemiology, 27, 602.

Park, S., Jee, S. H., Shin, H.-R., Park, E. H., Shin, A., Jung, K.-W. et al. (2014). Attributable fraction of tobacco smoking on cancer using population-based nationwide cancer incidence and mortality data in Korea. BMC Cancer, 14, 1–12.

Poole, C. (2015). A history of the population attributable fraction and related measures. Annals of Epidemiology, 25, 147–154.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89, 846–866.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with “censoring” due to death. Statistical Science, 21, 299–309.

Sasson, A. (2022a). Reproducibility materials for “the subtype-free average causal effect for heterogeneous disease etiology”. https://github.com/amitSasson/SFACE_Reproduce, Accessed: January 26, 2025.

Sasson, A. (2022b). The Subtype Free Average Causal Effect. R package version 1.0.0. https://cran.r-project.org/web/packages/TheSFACE/index.html. [Accessed 1 June 2022].

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M. and Hernán, M. A. (2020). Separable effects for causal inference in the presence of competing events. Journal of the American Statistical Association, 117, 175–183.
Sun, B., VanderWeele, T. and Tchetgen Tchetgen, E. J. (2017). A multinomial regression approach to model outcome heterogeneity. American Journal of Epidemiology, 186, 1097–1103.

Sun, B., VanderWeele, T. and Tchetgen Tchetgen, E. J. (2018). The authors reply. American Journal of Epidemiology, 187, 1130–1131.

Ugai, T., Väyrynen, J. P., Haruki, K., Akimoto, N., Lau, M. C., Zhong, R. et al. (2022). Smoking and incidence of colorectal cancer subclassified by tumor-associated macrophage infiltrates. Journal of the National Cancer Institute, 114, 68–77.

U.S. Census Bureau, (2023). Annual estimates of the resident population by single year of age and sex: April 1, 2020 to July 1, 2023. Accessed September 15, 2024.

Wang, M., Spiegelman, D., Kuchiba, A., Lochhead, P., Kim, S., Chan, A. T. et al. (2016). Statistical methods for studying disease subtype heterogeneity. Statistics in Medicine, 35, 782–800.

Wang, X., Amitay, E., Harrison, T. A., Banbury, B. L., Berndt, S. I., Brenner, H. et al. (2021). Association between smoking and molecular subtypes of colorectal cancer. JNCI Cancer Spectrum, 5, pkab056.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J. and Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. Statistics in Medicine, 39, 1199–1236.

Zehavi, T. and Nevo, D. (2023). Matching methods for truncation by death problems. Journal of the Royal Statistical Society Series A: Statistics in Society, 186, 659–681.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. Journal of Educational and Behavioral Statistics, 28, 353–368.