The Annals of Applied Statistics
2020, Vol. 14, No. 2, 829–849
https://doi.org/10.1214/20-AOAS1329

EVIDENCE FACTORS IN A CASE-CONTROL STUDY WITH APPLICATION TO THE EFFECT OF FLEXIBLE SIGMOIDOSCOPY SCREENING ON COLORECTAL CANCER

BY BIKRAM KARMAKAR[1], CHYKE A. DOUBENI[2] AND DYLAN S. SMALL[3]

[1]*Department of Statistics, College of Liberal Arts and Sciences, University of Florida, bkarmakar@ufl.edu*
[2]*Center for Health Equity and Community Engagement Research, Mayo Clinic, Doubeni.Chyke@mayo.edu*
[3]*Department of Statistics, The Wharton School, University of Pennsylvania, dsmall@wharton.upenn.edu*

As in any observational study, in a case-control study a primary concern is potential unmeasured confounders. Bias, due to unmeasured confounders, can result in a false discovery of an apparent treatment effect when there is none. Replication of an observational study, which tries to provide multiple analyses of the data where the biases affecting each analysis are thought to be different, is one way to strengthen the evidence from an observational study. Evidence factors allow for internal replication by testing a hypothesis using multiple comparisons in a way that the comparisons yield independent evidence and differ in the sources of potential bias. We construct evidence factors in a case-control study in which there are two types of cases, "narrow" cases which are thought to be potentially more affected by the exposure and "marginal" cases which are thought to have more heterogeneous causes. We develop and study an inference procedure for using such evidence factors and apply it to a study of the effect of sigmoidoscopy screening on colorectal cancer.

1. Introduction.

1.1. *Distal and proximal colon cancer and sigmoidoscopy screening*. The U.S. Preventive Services Task Force (USPSTF) recommendations for colorectal cancer screening include flexible sigmoidoscopy every five years for men and women above 50 at average risk (Preventive Services Task Force et al. (2016)). Yet, only 58% of adults aged 50–75 were up to date with the screening recommendations (Joseph et al. (2016)). Is screening with sigmoidoscopy effective? Using a case-control study we aim to answer this question; more specifically, we study the effect of screening by flexible sigmoidoscopy as per USPSTF recommendations on reducing mortality from colorectal cancer.

In case-control studies patients with (cases) or without (controls) an outcome of interest are compared in terms of their exposure to treatment. Case-control studies are particularly useful for assessing treatment or exposure effects for rare outcomes. In a case-control study there is often a choice of how to define a case. In many settings there are two (or more) ways to define a case, one being more "narrow," in that it is more likely to be caused by the exposure, if that exposure in fact has an effect, and the other being "broad" in that it may have more heterogeneous causes. A case unit according to a narrow case definition is also a case unit in a broad case definition. A marginal case unit is not a case in a narrow case definition but is a case in broad case definition.

Sigmoidoscopy can evaluate the lower or distal one-third of the colon for lesions; if abnormal, then a full colon evaluation with a colonoscopy is typically done for confirming the presence of cancer or precancerous polyps. The distal colon is the lower one-third part of

Received May 2018; revised February 2020.
Key words and phrases. Case-control studies, colorectal cancer, evidence factors, observational study, replicability, sigmoidoscopy.

the colon on the left side of the body, consisting of the descending colon, the sigmoid colon and the rectum; the proximal colon is the higher two thirds of the colon. We consider broad cases to be all cases of colorectal cancer, and, following Doubeni et al. (2018) and Selby et al. (1992), we consider narrow cases to be cases where there are malignant polyps on the left side of the colon and rectum that are within the reach of the sigmoidoscope. We expect that sigmoidoscopy screening, if it is effective, would only directly reduce the risk of diagnosis or death from cancers in the distal colon (narrow cases) but would also indirectly find or prevent some colorectal cancers in the proximal colon because abnormal findings in the distal colon could trigger a colonoscopy. Is it possible to learn separate evidence about the treatment effect when we have two or more definitions for a case? Before answering this question in Section 1.3, we consider why one might want to construct separate evidence and what we mean by separate evidence.

1.2. *Evidence factors in an observational study.* Unlike in a randomized trial, in a case-control study, as in any observational study, treatment is not assigned to the subjects randomly. Therefore, a primary concern in a case-control study is the potential for unmeasured confounders. In an observational study, bias, due to unmeasured confounders, can result in a false discovery of an apparent treatment effect when there is none. In such a situation we should consider if it possible to replicate the study without repeating the bias (Cochran (1965), Section 4.1).

Consider the effect of exposure to radiation on leukemia incidence. Radiologists, who are occupationally exposed to radiation, have been found to have a high incidence of leukemia (Lewis (1963)). A replication of this observational study is a comparison of the leukemia risk in people living in Japan near epicenters of the atomic bomb drops at the end of World War II to people living further from them (Bizzozero, Johnson and Ciocco (1966)). Radiologists may have higher rates of leukemia because they are more likely to diagnose it, and people living near the atomic bomb might have higher rates of leukemia because living in an urban area may be a confounder for leukemia, but these are two different sources of potential bias. Concurring finding of higher rates of leukemia incidence in each exposed group relative to its control group strengthens the evidence for a causal effect since two sources of bias, rather than just one, would be needed to refute the evidence (Rosenbaum (2001)).

While the above two comparisons are from separate studies, in some studies there may be two comparisons we can make within the same study that have different sources of bias, offering an opportunity for internal replication. When these comparisons are statistically independent or "nearly" independent, the comparisons are called evidence factors (Rosenbaum (2010)). A general perspective on evidence factors in an observational study is provided in Karmakar, French and Small (2019), which we briefly review here, and the formal definition is given in Section 5. Suppose two analyses are performed to test for the null hypothesis; the first analysis requires a set of assumptions $\mathcal{A}_1$, and the second analysis requires a second set of assumptions $\mathcal{A}_2$. Let P_1 and P_2 be the corresponding p-values. Then, to be evidence factors, we require that under the null hypothesis, when both assumptions $\mathcal{A}_1$ and $\mathcal{A}_2$ hold, for $(p_1, p_2) \in [0, 1]^2$

$$\Pr(P_1 \le p_1, P_2 \le p_2) \le p_1 p_2. \tag{1.1}$$

The inequality in (1.1)—which would be an equality if P_1 and P_2 were independent—means that the joint distribution of the p-values under the null hypothesis is stochastically bigger than that of two independent p-values under the null hypothesis. So, treating them as independent when combining them would be conservative—this is the "near independence" we spoke of above. By asking for independence or near independence, we ensure that we are learning two separate pieces of evidence rather than essentially one piece which would be

the case if one uses two highly correlated tests, such as a t-test and a Wilcoxon rank sum test (Rosenbaum (2010, 2011)). We wish to avoid the mistake of the man who bought "several copies of the morning paper to assure himself that what it said was true" (Wittgenstein (1958), #265, quoted in Rosenbaum (2010)). If both analyses from the evidence factors are significant, both assumptions, $\mathcal{A}_1$ and $\mathcal{A}_2$, would have to be violated in order for there not to be evidence of a treatment effect.

An example of the use of evidence factors is discussed in Karmakar, Small and Rosenbaum (2020), which follows up on the question raised by Bazzano et al. (2003), does smoking increases homocysteine levels? Bazzano et al. (2003) looked at the association between homocysteine and cotinine, a biomarker for exposure to tobacco. Cotinine level is a personal measure of a dose for exposure to tobacco. An association between homocysteine and cotinine can be confounded by a physiological process that affects both homocysteine levels and the way the exposure is internalized into cotinine levels. Karmakar, Small and Rosenbaum (2020) pair smokers with nonsmokers on their age, gender, race and education levels. Two tests are considered. The first test is a Wilcoxon's signed-rank test of the differences in the homocysteine levels between the smoker and the nonsmoker in each pair. The second test is a cross-cut test statistic that looks at the association between differences in biomarker levels and differences in the homocysteine levels of the pairs. Pairs of test statistics that use the same data are typically dependent, but these two test statistics are independent when there is no effect of smoking and there is no effect of an increase in the cotinine biomarker on homocysteine levels. Further, a bias in who reports smoking does not affect the cross-cut test, and a confounding in the cotinine biomarker does not affect the signed-rank test. Because the two tests are independent when there is no treatment effect and affected by different biases, they are evidence factors. Their analysis found that the two factors concur in finding two independent pieces of information linking smoking with increased homocysteine. For other examples of evidence factors, see Zhang et al. (2011) and Zubizarreta et al. (2012).

Rosenbaum (2017) provides a general formulation for building evidence factors based on multiple treatment assignment mechanisms. Starting with a set of n units, Rosenbaum (2017) showed how to construct evidence factors using the knit product of two subgroups of the symmetric group of size n. This and other previous work have only considered constructing evidence factors based on different ways of assigning treatment.

In this paper we develop novel evidence factors for case-control studies that use different definitions of a case. To the best of our knowledge, ours is the first demonstration of using differences in outcomes to develop evidence factors. In previous presentations of evidence factors, evidence factors are constructed from a study design in which treatment assignment splits into multiple aspects that exhibit certain symmetries (Rosenbaum (2010, 2017)). A case-control study differs in this view. The retrospective measure of an exposure to the treatment does not split into multiple aspects. The implicit symmetries that create the evidence factors in a case-control study come from multiple case definitions. The following subsection elaborates on this point.

This paper further demonstrates the usefulness of evidence factors when there are overlapping, but not completely overlapping, potential sources of bias for the analyses. This differs from previous discussions of evidence factors in the literature where separate sources of bias would affect the factors. Our quantitative demonstration of how evidence factors can work with overlapping biases widens the applicability of evidence factors. Expansion of the scope of evidence factors to incorporate the design aspects of case-control studies and overlapping biases is crucial for our sigmoidoscopy study.

1.3. *Evidence factors in a case-control study with narrow and marginal cases.* In a case-control study with narrow and broad cases, we expect that if the exposure has an effect and

our theory that the narrow cases are more likely to be caused by the exposure than the more heterogeneous broad cases is correct and also there is no unmeasured confounding, then: (a) the exposure should have a larger association with narrow cases than marginal cases, that is, cases that are broad but not narrow and (b) the exposure should have an association with broad cases compared to controls. This is an elaborate theory of what a treatment effect, if there is an effect, is expected to look like. Elaborate theories, advocated by Sir Karl Popper and Sir Ronald Fisher, are an integral part of drawing causal conclusions from observational data (see Popper (1959), Cochran (1965), Section 5). For related discussion on considerations for deducing causality from observational data, see Hill (1965).

We compare the narrow cases to marginal cases to appraise association of pattern (a) in the elaborate theory and compare broad cases to controls to appraise association of pattern (b). To test for patterns (a) and (b), we would like to use nearly independent test statistics in the sense of (1.1). In other words, we would like to develop evidence factors associated with the patterns. These two comparisons could be biased differently. Continuing our discussion of Section 1.1, in the sigmoidoscopy study unmeasured variables, such as healthy lifestyle or greater compliance with medical treatment, could be associated with screening. Some of these variables may be more associated with whether a person dies from any colorectal cancer or not (broad case vs. control); some may be more associated with, among people who die from colorectal cancer, does the person die from a colorectal cancer on the distal colon or proximal colon (narrow case vs. marginal case)? If we find evidence for both patterns (a) and (b), this would require a skeptic to explain more types of bias than if we found one pattern alone; this point is developed formally in Section 6.

Using the notation in Section 3, we develop a method for building the evidence factors in Section 4 and Section 5 which proves that the test statistics developed are evidence factors. The data from the study is analyzed in Section 7, and in Section 8 a few other examples of case-control studies are discussed where multiple case definitions are used. Before developing our method, we discuss the data for the sigmoidoscopy study in Section 2.

2. Sigmoidoscopy and colorectal cancer. Based on the reasoning of Section 1, we consider the effectiveness of screening sigmoidoscopy in relation to mortality from distal and proximal colon cancer. In relation to sigmoidoscopy screening, distal cancer cases are narrow cases, and proximal cancer cases are marginal cases. Throughout the paper by sigmoidoscopy screening we mean specifically flexible sigmoidoscopy screening.

2.1. *SCOLAR data.* In a nested case-control study on members of Kaiser Permanente Northern California and Kaiser Permanente Southern California health-care systems, study subjects were selected who were 55–90 years old between 2006 and 2012. Details of the study design are given in Doubeni et al. (2018), Goodman et al. (2015). A selected case unit would be a man or a woman who was 55–90 years old on the date of death with colorectal adenocarcinoma as the underlying cause of death. Using cancer diagnosis data and tumor characteristics, 822 proximal and 886 distal cancer cases were identified. Each case patient was individually matched to controls on the reference date (which was the diagnosis date for each patient who died of colorectal cancer), gender, the duration of health plan prior to diagnosis and the health-care site. In this process 3635 controls were included.

Thus, in our design there are 822 narrow cases and 886 marginal cases. To facilitate the comparison of narrow cases to marginal cases, we pair matched narrow (distal cancer) cases to marginal (proximal cancer) cases using the `optmatch` package in R which uses methods of Hansen and Klopfer (2006). The matching algorithm used a weighted sum of rank based Mahalanobis distance and absolute distance of estimated logit propensity scores. It also near fine balanced on gender (Rosenbaum, Ross and Silber (2007)). By pair matching the narrow and marginal cases, we obtained 822 matched sets consisting of one narrow case, one

TABLE 1

Balance on the covariates in the matched sets. Distal cancer cases are those who have been diagnosed to have died from cancer on the left colon or rectum; proximal cancer cases are from right colon cancer. For each covariate the mean is calculated within a matched set, then averaged over sets

	Controls	Distal cancer cases	Proximal cancer cases
Number of years enrolled before reference date	12	12	12
% from Center 1	83	83	84
% of female	47	46	47

marginal cases and the controls associated with these cases and $886 - 822 = 64$ matched sets consisting of one marginal case and the controls associated with this case. Table 1 shows the covariate balance of the matched sets. Figure 1 further shows the distribution of the diagnosis year of the colorectal cancer patients. Gender, reference date and enrollment source are well balanced between the narrow cases, marginal cases and controls over the matched sets.

Although the match controls well for the above covariates, there could be unmeasured confounders. For example, lack of physical activity is a known risk factor of colorectal cancer incidence, and people who are less active also may be less likely to get screened (Eldridge et al. (2013)). Because we are not able to match on or adjust for physical activity in our analysis, the comparison of all colorectal cancer cases to controls may be biased. Family history of cancer screening is another likely unmeasured confounder in this analysis. The comparison of sigmoidoscopy screening in proximal vs. distal cancers may also be biased by unmeasured confounding. There are potential biological differences between proximal and distal colon cancers such that variables such as diet (e.g., use of the Mediterranean diet) may be differentially associated with proximal and distal colon cancer (Doubeni et al. (2012), Missiaglia et al. (2014)). Such diet choices may be associated with screening. If we find that sigmoidoscopy screening is associated with reduced morality from colorectal cancer when comparing all cases to controls and with reduced mortality from proximal vs. distal cancer cases when comparing proximal to distal cases, then, in order for these associations to arise

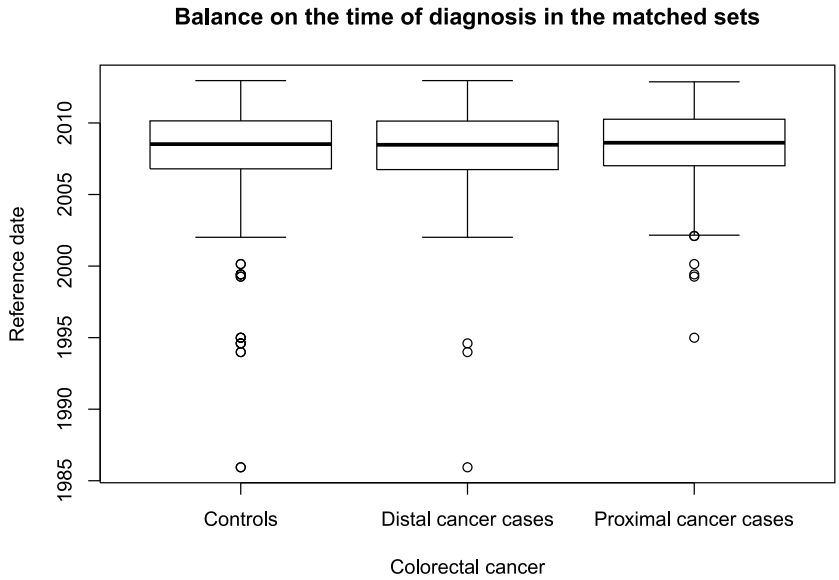

FIG. 1. *Reference date of the colorectal cancer cases and controls in the matched sets.*

purely from bias and not at all from a causal effect of sigmoidoscopy screening on reducing cancer, there would need to be unmeasured confounders in both comparisons rather than just one comparison. In Section 6 we show that, even if the unmeasured confounders for the two comparisons overlap but have different relative magnitudes, the evidence is strengthened by finding significant associations in both comparisons.

As suggested earlier, we shall assess the effect of sigmoidoscopy screening by comparing the prevalence of screening between all colorectal cancer cases and controls and also by comparing the prevalence between the distal cancer cases and proximal cancer cases. Results of this analysis will be discussed in Section 7. We first present the methodology.

3. Notation and review: Case-control studies. Let observational units be denoted by indices $l = 1, \ldots, L$. We use the binary variable Z_l to denote whether unit l was exposed to treatment ($Z_l = 1$) or spared from being exposed ($Z_l = 0$). Under the potential response model, suppose unit l, if exposed, would have response $\mathbf{r}_{Tl}$ and, if spared, exposure would have response $\mathbf{r}_{Cl}$. The observed response for unit l is $\mathbf{R}_l = Z_l\mathbf{r}_{Tl} + (1 - Z_l)\mathbf{r}_{Cl}$. Consequently, we cannot observe $\mathbf{r}_{Tl}$ and $\mathbf{r}_{Cl}$ simultaneously for one unit (Neyman (1923), Rubin (1974)). Now, let $\mathbf{x}_l$ denote the observed pretreatment covariates, that is, covariates recorded in the study that can potentially affect the treatment assignment and the response. The unobserved confounders are summarized by an unobserved number u_l for unit l scaled to be valued in $[0, 1]$ (Rosenbaum (1991)). Write $\mathcal{F} = \{(\mathbf{r}_{Tl}, \mathbf{r}_{Cl}, \mathbf{x}_l, u_l) : l = 1, \ldots, L\}$. The hypothesis we are interested in studying is Fisher's sharp null hypothesis of no treatment effect

$$H_0 : \mathbf{r}_{Tl} = \mathbf{r}_{Cl}, \quad l = 1, \ldots, L.$$

A case definition is a function $k(\cdot)$ which labels each unit as a case, or a control or neither based on the observed response. A case definition would identify a subset of the units as cases and a separate subset as controls.

For a given case definition, a test for the hypothesis H_0 can be carried out by matching as follows. Create S strata labeled $s = 1, \ldots, S$ where each stratum consists of a total of t_s units with some case units and the rest control units (say c_s) which are similar with respect to the observed covariates ($\mathbf{x}_l$'s). Now, let Y_s denote the total number of exposed case units in stratum s. A positive linear combination $T = \sum_{s=1}^{S} d_s Y_s$ can be taken as a test statistic for testing the hypothesis H_0. When all $d_s = 1$, the statistic T is exactly the total number of exposed cases which is the Mantel–Haenszel test statistic.

We assume that the treatment assignments for distinct units are independent. We consider the following model for treatment assignment:

$$\Pr(Z_l = 1 \mid \mathcal{F}) = \frac{\exp\{\theta(\mathbf{x}_l) + \gamma u_l\}}{1 + \exp\{\theta(\mathbf{x}_l) + \gamma u_l\}}, \tag{3.1}$$

where $\theta(\cdot)$ is an unknown function and $\gamma \geq 0$ is an unknown parameter. Since $0 \leq u_l \leq 1$, for two units, l and l' ($l \neq l'$), with the same observed covariates, $\mathbf{x}_l = \mathbf{x}_{l'}$, under this model their odds of exposure can vary at most by a factor of $\Gamma := \log(\gamma)$. Model (3.1) is equivalent to writing

$$\max_{1 \leq l, l' \leq L} \left\{ \frac{\Pr(Z_l = 1 \mid \mathcal{F}) / \Pr(Z_l = 0 \mid \mathcal{F})}{\Pr(Z_{l'} = 1 \mid \mathcal{F}) / \Pr(Z_{l'} = 0 \mid \mathcal{F})} : \mathbf{x}_l = \mathbf{x}_{l'} \right\} \leq \Gamma. \tag{3.2}$$

The fact that (3.1) implies (3.2) is obvious; the proof of the reverse implication constructs a set of u_l from the odds of exposure (Rosenbaum (2002), Section 4.4.4). The parameter Γ (≥ 1) is the hidden bias level. Thus, when $\Gamma = 1$, there is no unmeasured confounder, and there is no bias in treatment assignment after controlling for observed covariates. As Γ increases, this model allows more and more bias in treatment assignment. For example, when $\Gamma = 2$, due to

the presence of unmeasured confounders, it might be possible that, for individuals who are the same in their observed covariates, one has twice the odds of getting assigned treatment as the other.

Let e_s be the number of exposed units in stratum s. Then, under model (3.2) we can bound the tail probability of T under H_0 asymptotically,

$$\Pr\big(T \geq k \mid \{t_s\}, \{c_s\}, \{e_s\}, \mathcal{F}\big) \leq 1 - \Phi\bigg(\frac{k - \sum d_s(t_s - c_s)\bar{p}_s}{\sqrt{\sum d_s^2(t_s - c_s)\bar{p}_s(1-\bar{p}_s)}}\bigg), \tag{3.3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\bar{p}_s = \Gamma e_s/(\Gamma e_s + (t_s - e_s))$ (Small et al. (2013)). This tail bound is sharp, in that it is attained for a particular vector of unobserved confounders (Rosenbaum (1991), Rosenbaum (2002), Section 4.4.4).

Therefore, given a case-control study, after constructing a satisfactory stratum structure, when the hidden bias level is at most Γ, that is, (3.2) holds, (3.3) can be used to get an upper bound for the p-value of testing the hypothesis H_0. If this value is less than α, the significance level, then we have evidence to reject the null hypothesis as long as the hidden bias is at most Γ. A sensitivity analysis asks how much bias in the treatment assignment must be present so that the observed association can be explained just from bias under H_0.

4. Two case definitions and two comparisons. Following our discussion in Section 1.3, consider a design with availability of two case definitions, one narrow and one broad. A case unit according to a narrow case definition is also a case unit in a broad case definition. We label a unit as a marginal case unit if it is not a case in a narrow case definition but is a case in broad case definition. The study units which are noncases in broad case definition are, thus, also noncases in the narrow case definition and are labeled as controls. Matching argument similar to Section 3 can still be used with appropriate modifications.

4.1. *Matched strata for the comparisons.* Suppose the matching procedure creates S strata of all three types of units: narrow cases, marginal cases and controls where units in a stratum are similar in their observed covariates. Let a generic stratum labeled s have n_s narrow cases, m_s marginal cases, thus, a total of $b_s = n_s + m_s$ broad cases and c_s controls. In a cohort of L units, a narrow case definition might have a much smaller number of cases than a broad case definition. In such situations some of the stratum (s) may only have marginal cases and controls, resulting in $n_s = 0$, which is allowed in our notation. But each stratum must consist of at least two different labels of units. Let the letters n, m, b or c for denoting that the unit is a narrow case, a marginal case, a broad case or a control, respectively. For example, $Z_{n\{si\}}$ denotes the exposure (0 or 1) for the ith narrow case in the stratum s (s in $1, 2, \ldots, S$). The index i runs in $[n_s]$ (we use the notation $[k]$ to denote the set $\{1, \ldots, k\}$ if k is a positive integer or empty set $\{\}$ otherwise). Similarly, $\mathbf{x}_{c\{si\}}$ denotes the observed covariate for the ith control in stratum s. $\mathbf{R}_{m\{si\}}$, $\mathbf{r}_{Cn\{si\}}$, $u_{c\{si\}}$ etc. have similar meanings.

At this point we can quantify the evidence against H_0 by calculating the p-values from the two comparisons of narrow cases vs. marginal cases and broad cases vs. controls. We focus on the linear statistics of the number of exposed narrow cases and broad cases, respectively, for these two comparisons. Let $Y_{n\{s\}}$ and $Y_{b\{s\}}$ for stratum labeled s; denote the number of exposed narrow cases and the number of exposed broad cases. Notice that $Y_{n\{s\}} = \sum_{i\in[n_s]} Z_{n\{si\}}$ and $Y_{b\{s\}} = \sum_{i\in[b_s]} Z_{b\{si\}}$. Since broad cases encompass narrow cases, in fact,

$$Y_{b\{s\}} = \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}} = Y_{n\{s\}} + Y_{m\{s\}}.$$

Two test statistics for these two comparisons can be written as $T_{nm} = \sum_{s=1}^{S} d_{nm\{s\}} Y_{n\{s\}}$ and $T_{bc} = \sum_{s=1}^{S} d_{bc\{s\}} Y_{b\{s\}}$, where $d_{nm\{s\}}$ and $d_{bc\{s\}}$ are nonnegative constants given $\mathcal{F}$. Under assumption (3.2) about treatment assignment distribution, we can get bounds on the p-values for T_{nm} and T_{bc}. But there are a few subtleties here that are important to point out.

First, a p-value for T_{nm} should only be based on information from the narrow cases and marginal cases. In other words, the p-value P_{nm} is computed based on the tail distribution

$$\Pr\Big(T_{nm} \geq k | \{b_s\}, \{m_s\}, \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}, \mathcal{F}_b\Big), \tag{4.1}$$

where $\mathcal{F}_b$ is the subset of $\mathcal{F}$ restricted to the broad cases. In equation (3.3), t_s was used instead of b_s, c_s was used instead of m_s and the sum above replaces e_s. Similarly, the p-value P_{bc} is computed based on the tail distribution

$$\Pr\Big(T_{bc} \geq k | \{b_s + c_s\}, \{c_s\}, \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}} + \sum_{i\in[c_s]} Z_{c\{si\}}, \mathcal{F}\Big). \tag{4.2}$$

Thus, in technical terms P_{nm} and P_{bc} are measurable with respect to different sigma fields.

Second, in assumption (3.2) the sensitivity parameter Γ bounds the odds ratio of treatment assignment for all the units stratified on their observed covariates. But unmeasured confounders are likely to affect the two comparisons in different ways (see also Section 6). Therefore, while considering narrow versus marginal comparison, we should relax this assumption only to the broad cases since these are the only ones contributing to T_{nm}. Hence, we distinguish the effect of unmeasured covariates for the two comparisons by using two sensitivity parameters Γ_{nm} and Γ_{bc} for the narrow vs. marginal and broad vs. control comparisons, respectively. Then, Γ_{nm} measures the bias in treatment assignment among all the case units, and Γ_{bc} measures the bias in treatment assignment among all case and control units which are similar in their observed covariates.

Therefore, the comparison of narrow vs. marginal cases would compute the upper bound on the p-value for T_{nm} based on the tail distribution (4.1) for sensitivity parameter Γ_{nm}; the broad cases vs. controls comparison would compute the upper bound on the p-value for T_{bc} based on the tail distribution (4.2) for sensitivity parameter Γ_{bc}. We denote them by $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$, respectively, and, when $\Gamma_{nm} = \Gamma_{bc} = 1$, we simply write P_{nm} and P_{bc} for $P_{nm,1}$ and $P_{bc,1}$ respectively. Section 5 proves that $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$ are nearly independent.

4.2. *Two sensitivity parameters and their amplification.* In a sensitivity analysis the sensitivity parameters Γ_{nm} and Γ_{bc} would be used to get the max p-values $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$. How does a Γ_{nm} bias relate to the influence of the unmeasured confounding on the exposure to treatment of an unit and the influence of the unmeasured confounding on the narrow to marginal case status of the unit? The sensitivity analysis model (3.1) conditions on the information set $\mathcal{F}$ which includes the potential outcomes of the units. The maximum p-value calculated under this model is achieved when there is a near perfect relationship between the case definition and the unmeasured confounders. We discuss here that this model can be interpreted differently, "amplified," to be a model that limits the relationship between the case definition and the unmeasured confounders as well as the relationship between the exposure and the unmeasured confounders (Gastwirth, Krieger and Rosenbaum (1998), Rosenbaum and Silber (2009)).

Let the confounding variable in the broad cases to controls comparison be u_1 and the confounding variable in narrow to marginal comparison be u_2. Consider now the set $\mathcal{C} = \{(\mathbf{x}_l, u_{1l}, u_{2l}) : l = 1, \ldots, L\}$. As before, $0 \leq u_{1l} \leq 1$ and $0 \leq u_{2l} \leq 1$. Conditioning on the set $\mathcal{C}$ does not condition on the potential outcomes.

Consider two units $i1$ and $i2$ with the same observed covariates. We model the relationship between the unmeasured confounding and the treatment assignment with a parameter λ, for $z_{i1} + z_{i2} = 1$, as

$$\text{(4.3)}\quad \Pr(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} | \mathcal{C}, \mathbf{x}_{i1} = \mathbf{x}_{i2}, Z_{i1} + Z_{i2} = 1) = \frac{\exp\{\lambda(z_{i1}w_{i1} + z_{i2}w_{i2})\}}{\exp(\lambda w_{i1}) + \exp(\lambda w_{i2})},$$

where

$$\text{(4.4)}\qquad w_l = \xi_1 u_{1l} + \xi_2 u_{2l} \quad \text{for } l = 1, \ldots, L; \xi_1, \xi_2 \geq 0, \xi_1 + \xi_2 = 1.$$

If $\lambda = 0$, the probability is 1/2, and the confounders have no effect. A larger value of λ indicates a larger influence of the unmeasured confounders on the treatment assignment. Equation (4.4) in itself is not a new assumption. Any number w_l, taking value in $[0, 1]$, can be rewritten as $w_l = \xi_1 u_{1l} + \xi_2 u_{2l}$, for $\xi_1, \xi_2 \geq 0, \xi_1 + \xi_2 = 1$ and $0 \leq u_{1l}, u_{2l} \leq 1$, and vice versa. Hence, this model is similar in spirit to model (3.1) except that the principal conditioning now changes from $\mathcal{F}$ to $\mathcal{C}$.

Next, we model the relationship of the unmeasured confounding and the case status. Let us denote for unit l, when not exposed to the treatment, by the indicator variable k^b_{Cl}, whether the unit is a case, and by k^n_{Cl}, whether the unit is a narrow case. Thus, $k^b_{Cl} = 1$ if the lth unit is a case, either narrow or marginal, when not exposed to the treatment and $k^b_{Cl} = 0$ if the unit is a control when not exposed to the treatment. Similarly, $k^b_{Cl} = 1$ if the lth unit is a narrow case when not exposed to the treatment and $k^b_{Cl} = 0$ otherwise. It might be helpful to think of k^b_{Cl} and k^n_{Cl} as being determined by $\mathbf{r}_{Cl}$. For two units $i1$ and $i2$ with similar observed covariates, the following model relates the case label with the confounders:

$$\text{(4.5)}\qquad \frac{\Pr(k^b_{Ci1} = 1, k^b_{Ci2} = 0 \mid \mathcal{C}, \mathbf{x}_{i1} = \mathbf{x}_{i2})}{\Pr(k^b_{Ci1} = 0, k^b_{Ci2} = 1 \mid \mathcal{C}, \mathbf{x}_{i1} = \mathbf{x}_{i2})} = \exp\{\delta_{bc}(u_{1,i1} - u_{1,i2})\};$$

$$\text{(4.6)}\qquad \frac{\Pr(k^n_{Ci1} = 1, k^n_{Ci2} = 0 \mid \mathcal{C}, \mathbf{x}_{i1} = \mathbf{x}_{i2}, k^b_{Ci1} = k^b_{Ci2} = 1)}{\Pr(k^n_{Ci1} = 0, k^n_{Ci2} = 1 \mid \mathcal{C}, \mathbf{x}_{i1} = \mathbf{x}_{i2}, k^b_{Ci1} = k^b_{Ci2} = 1)} = \exp\{\delta_{nm}(u_{2,i1} - u_{2,i2})\}.$$

The level of bias from unmeasured confounding u_1 in being a broad case is δ_{bc}, and the level of bias from unmeasured confounding u_2 in being a narrow case over a marginal case is δ_{nm}—the larger the value of these parameters, the higher the influence of the unmeasured confounding.

How do λ, δ_{bc} and δ_{nm} relate to the sensitivity parameters Γ_{bc} and Γ_{nm}? Proposition 1 of Rosenbaum and Silber (2009) provides the correspondence. Let $\Lambda = \exp(\lambda)$, $\Delta_{bc} = \exp(\delta_{bc})$ and $\Delta_{nm} = \exp(\delta_{nm})$. Then, $\Gamma_{bc} = (\Delta_{bc}\Lambda + 1)/(\Delta_{bc} + \Lambda)$ and $\Gamma_{bc} = (\Delta_{nm}\Lambda + 1)/(\Delta_{nm} + \Lambda)$. These formulas allow one to interpret the result of a sensitivity analysis either using the sensitivity parameters Γ_{bc} and Γ_{nm} or, under model (4.3)–(4.6), using parameters λ, δ_{bc} and δ_{nm}. For example, $\Gamma_{nm} = 1.5$, $\Gamma_{bc} = 1.4$ corresponds to $\Lambda = 2$, $\Delta_{nm} = 5/3$ and $\Delta_{bc} = 2$. In words, a pair of bias levels of $\Gamma_{nm} = 1.5$ and $\Gamma_{bc} = 1.4$ is equivalent to an effect of unmeasured confounders that, for units that are similar in their observed covariates, doubles the chance an exposure, while also increasing the chance of being a case by 5/3-fold and increasing the chance of being a narrow case over a marginal case by twofold. Similarly, $\Gamma_{nm} = 3$, $\Gamma_{bc} = 2$ corresponds to $\Lambda = 5$, $\Delta_{nm} = 7$ and $\Delta_{bc} = 3$ and so on.

5. Evidence factors. This section aims to establish that the two comparisons discussed in Section 4.1 explore different aspects of the study design and give separate evidence and, thus, are evidence factors. The idea of evidence factors was first formalized by Rosenbaum (2010) and extended for studies with multiple treatment assignment mechanisms in Rosenbaum (2011), Rosenbaum (2017). As discussed in Section 1.2, Karmakar, French

and Small (2019) provide a general formulation of evidence factors in observational study designs. Readers interested in the results of the SCOLAR data analysis can skip this technical discussion and go to Section 5.1 and 7.

We start this section by stating the definition of evidence factors. To understand that equation (5.1) is a more general statement than (1.1) that was used to introduced evidence factors in Section 1.2, notice that replacing $X = (P_1, P_2)$, $D = [0, p_1] \times [0, p_2]$ and Y a uniform distribution on $[0, 1]^2$ recreate (1.1). The main result of this section, Theorem 5.1, says that, according to this definition, $(P_{nm,\Gamma_{nm}}, P_{bc,\Gamma_{bc}})$ form evidence factors.

DEFINITION 1. A set D is called a decreasing set if for any pair $(\mathbf{x}, \mathbf{y})$ with $\mathbf{x} \leq \mathbf{y}$, if $\mathbf{y} \in D$, then $\mathbf{x} \in D$. For two random vectors $\mathbf{X}$ and $\mathbf{Y}$ we say that $\mathbf{X}$ is stochastically larger than $\mathbf{Y}$ if

$$\Pr(\mathbf{X} \in D) \leq \Pr(\mathbf{Y} \in D) \tag{5.1}$$

for all nondecreasing sets D. If $\mathbf{X}$ is stochastically larger than $\mathbf{Y}$, we write $\mathbf{X} \succcurlyeq \mathbf{Y}$.

DEFINITION 2. For any pair of bias levels $(\Gamma_{nm}, \Gamma_{bc})$, $(P_{nm,\Gamma_{nm}}, P_{bc,\Gamma_{bc}})$ are evidence factors for testing H_0, if $(P_{nm,\Gamma_{nm}}, P_{bc,\Gamma_{bc}}) \succcurlyeq (U_1, U_2)$ under the bias levels Γ_{nm}, Γ_{bc} and under H_0 for two independent Unif[0, 1] random variables U_1 and U_2.

Now, we state the main theorem.

THEOREM 5.1. *Under H_0 and for bias levels Γ_{nm} and Γ_{bc}, we have $(P_{nm,\Gamma_{nm}}, P_{bc,\Gamma_{bc}}) \succcurlyeq (U_1, U_2)$ for two independent* Unif[0, 1] *random variables U_1 and U_2.*

The rest of the section is dedicated to proving this theorem using a few lemmas. The proof of all the lemmas are given in the Appendix. These lemmas clarify the functional relationships of $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$ on the exposure of the units' Z_ls. Since the Z_l's are the only random variables that determine the p-values or their upper bounds, the purpose of these lemmas in proving the theorem is to show that $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$ depend on different parts of the Z_ls. For a crude understanding of this, notice the term $Z_{c\{si\}}$ in the expression of $P_{bc,\Gamma_{bc}}$ in Lemma 5.2 which is missing from the corresponding expression of $P_{nm,\Gamma_{nm}}$—whether a control unit is exposed to the treatment does not affect the narrow vs. marginal cases analysis. Lemma 5.3 shows that, not only are $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$ stochasticaly larger than a uniform distribution on [0, 1], for they are larger than the true but unknown p-values, different conditional distributions of them are also stochastically larger than a uniform distribution on [0, 1]. Theorem 5.1 is about the joint distribution of $(P_{nm,\Gamma_{nm}}, P_{bc,\Gamma_{bc}})$. Thus, the facts about the marginal distributions of $P_{nm,\Gamma_{nm}}$, $P_{bc,\Gamma_{bc}}$ and their conditional distributions given certain events, along with a general lemma, Lemma 5.5, proves the theorem.

To slightly simplify our notation in what follows, for two random vectors X and Y we write $[X \mid Y]$ to denote the conditional distribution of X given Y. Since we are dealing with discrete spaces, $[X \mid Y]$ is a real valued measurable function of X and Y.

The following is one of the main lemmas needed to prove Theorem 5.1:

LEMMA 5.2. *There exists functions f_{nm} and f_{bc} on appropriate domains such that*

$$P_{nm,\Gamma_{nm}} = f_{nm}\left(\left\{Z_{n\{si\}}, i \in [n_s]; \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}} | s \in [S]\right\}\right)$$

and

$$P_{bc,\Gamma_{bc}} = f_{bc}\left(\left\{Z_{c\{si\}}, i \in [c_s]; \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}} | s \in [S]\right\}\right).$$

Following Definition 1, let us use the notation $X \succcurlyeq D$ for a random variable X and a probability distribution D to say that X is stochastically larger than D or $\Pr(X \le x) \le \Pr(Y \le x \mid Y \sim D)$ for all $x \in \mathbb{R}$.

LEMMA 5.3. *Under H_0, we have the following*:

(i) $[P_{nm,\Gamma_{nm}} \mid \{\sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}\}, \mathcal{F}_b, \{n_s\}] \succcurlyeq \text{Unif}[0,1]$.

(ii) $[P_{bc,\Gamma_{bc}} \mid \{\sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}} + \sum_{i\in[c_s]} Z_{c\{si\}}\}, \mathcal{F}, \{b_s + c_s\}] \succcurlyeq \text{Unif}[0,1]$.

(iii) $P_{nm,\Gamma_{nm}} \succcurlyeq \text{Unif}[0,1]$.

(iv) $P_{bc,\Gamma_{bc}} \succcurlyeq \text{Unif}[0,1]$.

The following lemma relies on the assumption of no interference in treatment assignment among the units, which is to say Z_l and $Z_{l'}$ are independently distributed for two distinct units l and l':

LEMMA 5.4. *Under H_0,*

$$\left[P_{nm,\Gamma_{nm}} \,\middle|\, \{Z_{c\{si\}}, i \in [c_s]\}; \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}\right] \succcurlyeq \text{Unif}[0,1].$$

LEMMA 5.5. *Suppose two random variables P_1 and P_2 satisfy*

C1 *random variable P_1 is a function of random quantity V_1,*

C2 $[P_2 \mid V_1] \succcurlyeq \text{Unif}[0,1]$,

then for $0 \le q \le 1$, $Pr(P_2 \le q \mid P_1) \le q$, that is, $[P_2 \mid P_1] \succcurlyeq \text{Unif}[0,1]$.

Now, we have all the necessary facts to prove Theorem 5.1.

PROOF OF THEOREM 5.1. In Lemma 5.5 take $P_1 = P_{bc,\Gamma_{bc}}$, $P_2 = P_{nm,\Gamma_{nm}}$ with $V_1 = \{\{Z_{c\{si\}}, i \in [c_s]\}; \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}\}$. Then, by Lemma 5.2 condition C1 is satisfied, and condition C2 is proved in Lemma 5.4. Thus, by Lemma 5.5 $[P_{nm,\Gamma_{nm}} \mid P_{bc,\Gamma_{bc}}] \succcurlyeq \text{Unif}[0,1]$.

Let U_1 and U_2 be two independent uniformly distributed random variables on $[0,1]$. We use the theory of Shaked and Shanthikumar ((2007), Section 6B), (U_1, U_2) being an independent pair is a conditionally increasing in sequence (CIS). Then, combining this with the facts that $P_{bc,\Gamma_{bc}} \succcurlyeq \text{Unif}[0,1]$ (by Lemma 5.3) and $[P_{nm,\Gamma_{nm}} \mid P_{bc,\Gamma_{bc}}] \succcurlyeq \text{Unif}[0,1]$, Theorem 6.B.4 of Shaked and Shanthikumar (2007) finally gives us

$$(P_{nm,\Gamma_{nm}}, P_{bc,\Gamma_{bc}}) \succcurlyeq (U_1, U_2).$$

Thus, the proof is complete. □

5.1. *Combining evidence.* In words, Theorem 5.1 says that the combined information from the two evidence factors, $P_{nm,\Gamma_{nm}}$ and $P_{bc,\Gamma_{bc}}$, carries as much evidence as two analyses from two independent studies. This allows us to combine these two pieces of evidence and provide a total evidence against the hypothesis under both the comparisons. Karmakar, French and Small (2019) discusses different methods for combining evidence. Any method of combining p-values that is monotone in both of the p-values can be used, for example, Fisher's combination method (Fisher (1932)), the mean of the normal transformation (Liptak (1958)) and the truncated product method of combining (Hsu, Small and Rosenbaum (2013), Zaykin et al. (2002)). Also see Becker (1994). These methods of combining p-values are

used when p-values are available from independent sources, for example, in meta-analysis. In an observational study, even when there are independent tests, combining them does not strengthen the evidence against the biases from unmeasured confounders if the analysis are affected by the same unmeasured confounding. The evidence factors are two analyses that are nearly independent and that do not share completely overlapping biases. Thus, combining the maximum p-values from the evidence factors strengthens the evidence in an observational study. The simulation section considers which combining method has largest power in sensitivity analysis for unmeasured confounding.

Fisher's method computes the joint evidence as the tail probability of χ^2_4 distribution over $-2\log(P_{nm,\Gamma_{nm}} \cdot P_{bc,\Gamma_{bc}})$. In the scenario of sensitivity analysis, since we only consider largest possible p-values for a given value of hidden bias level, the truncated product method, which weights the evidence by the strength of the evidence, is often preferred. For a given $\tilde{\alpha}$, the combined evidence using the truncated product method is given by $F_W\{\mathrm{Ev}(\Gamma_{nm}, \Gamma_{bc})\}$, where

$$\mathrm{Ev}(\Gamma_{nm}, \Gamma_{bc}) = \mathbb{1}_{P_{nm,\Gamma_{nm}} \le \tilde{\alpha}} \log(P_{nm,\Gamma_{nm}}) + \mathbb{1}_{P_{bc,\Gamma_{bc}} \le \tilde{\alpha}} \log(P_{bc,\Gamma_{bc}}) \quad \text{and}$$

$$F_W\{w\} = 2\tilde{\alpha}(1-\tilde{\alpha}) G_{\mathrm{Exp}(1)}\left\{-\log\left(\frac{w}{\tilde{\alpha}}\right)\right\} + \tilde{\alpha}^2 G_{\mathrm{Gamma}(2,1)}\left\{-\log\left(\frac{w}{\tilde{\alpha}^2}\right)\right\}. \tag{5.2}$$

In the above, $G_{\mathrm{Exp}(1)}$ is the survival function of a random variable with exponential distribution with rate 1, and $G_{\mathrm{Gamma}(2,1)}$ the survival function of a random variable with Gamma distribution with shape parameter 2 and rate 1. The advised choice of $\tilde{\alpha}$ is 0.20 (Hsu, Small and Rosenbaum (2013), Zaykin et al. (2002)).

We conducted a simulation study to compare the powers of Fisher's method and the truncated product method in the setting of our problem. The simulation scenario considered here is based on the case-control study structure. We are going to look at the favorable situation where there are no unmeasured confounders with treatment effect. Then, for varied treatment effect sizes we compare the power of the two combining methods for different values of $(\Gamma_{nm}, \Gamma_{bc})$.

We consider a population where the chance of exposure is $1/3$. Thus, for a unit l, $\Pr(Z_l = 1) = 1/3$. The treatment effect is denoted by β. We consider a univariate response and two types of response distributions in the population. The two types of distributions when spared exposure are a normal distribution with mean 0 and variance 1 and a t-distribution normalized to have variance 1. Therefore, if a unit l is exposed to treatment, then the response is a sample from $N(\beta, 1)$ (or $\beta + t_3/\sqrt{3}$), and if not exposed, then the response is a sample from $N(0, 1)$ (or $t_3/\sqrt{3}$). The case definition for each of the scenarios is taken such that if the treatment effect was 0.5, then 20% of the population would be broad cases. Thus, in the setting where the response is from normal distribution, the unit with response of more than the 0.8 quantile of the mixture distribution $1/3N(\beta, 1) + 2/3N(0, 1)$ would be labeled a broad case, and otherwise it would be labeled as a control. In our simulation we sample 2000 broad cases, and half of them with response above the median response of these broad cases are labeled as narrow cases. Then, we sample 2000 controls. In both comparisons of narrow cases vs. marginal cases and broad cases vs. controls, we consider paired stratum, that is, $n_s = m_s = 1, c_s = 2$.

Tables 2 and 3 report the simulated power for the two combining methods. The simulated power is based on 10,000 iterations with level of significance $\alpha = 0.05$. Except for very few situations in Table 2, the truncated product method has better simulated power than Fisher's combining method. The truncated product method seem to be less sensitive as we increase Γ_{nm} and Γ_{bc}. Fisher's method has slightly better simulated power in a few situations in the normal response model for moderate values of $(\Gamma_{nm}, \Gamma_{bc})$ when there is a large treatment effect ($\beta = 0.6$). After considering these simulation results, in our case-control study of the efficacy of screening sigmoidoscopy we use the truncated product method with $\tilde{\alpha} = 0.20$.

TABLE 2

Simulated power, in %, of a sensitivity analysis of combined evidence in a case-control study, where there is no unmeasured confounder and $\Pr(Z_l = 1) = 1/3$. *The response is simulated from* $N(\beta, 1)$ *if* $Z_l = 1$ *and* $N(0, 1)$ *if* $Z_l = 0$. *There are* 1000 *narrow cases and* 1000 *marginal cases with* 2000 *controls. Based on* 10,000 *iterations. Fisher = Fisher's combination method, tP = truncated product method with* $\tilde{\alpha} = 0.20$

		$\beta = 0$		$\beta = 0.2$		$\beta = 0.4$		$\beta = 0.6$	
Γ_{nm}	Γ_{bc}	Fisher	tP	Fisher	tP	Fisher	tP	Fisher	tP
1	1	5	5	100	100	100	100	100	100
	1.5	0.6	1	25	26	100	100	100	100
	2	0.6	1	18	22	87	86	100	100
	2.5	0.6	1	18	22	75	80	100	100
1.25	1.25	0	0	48	51	100	100	100	100
	2	0	0	0	0.1	15	15	100	100
	2.75	0	0	0	0.1	3	5	69	66
	3.5	0	0	0	0.1	3	5	69	66
1.5	1.5	0	0	0.2	0.3	99.2	99.4	100	100
	2.5	0	0	0	0	0	0	54	52
	3.5	0	0	0	0	0	0	1	2
1.75	1.75	0	0	0	0	51	58	100	100
	2	0	0	0	0	2	3	100	100
	3.25	0	0	0	0	0	0	0	0
2	2	0	0	0	0	2	3	100	100
	2.5	0	0	0	0	0	0	36	43
	3	0	0	0	0	0	0	0.1	0.2
	3.5	0	0	0	0	0	0	0	0
2.25	2.25	0	0	0	0	0	0	88	91
	2.5	0	0	0	0	0	0	35	42
	3	0	0	0	0	0	0	0.1	0.2

6. Evidence factors with differential effect of unmeasured confounders on the factors. The individual factors in an evidence factors analysis, if biased, are hoped to be biased by different mechanisms so that a critic would need to consider both sources of bias to explain the observed statistical significance. As discussed in Section 2.1, in the sigmoidoscopy study the bias in comparing all colorectal cancer cases to controls could be due to imbalance between the two groups in healthy lifestyle of the patients, family history and also, potentially, due to diet. The comparison of distal cancer cases to proximal cancer cases may be biased by diet, for example, Mediterranean diet. Hence, the main source of unmeasured confounding in the second analysis can, to some extent, also be a source of bias in the first analysis. The following discussion delineates the logic of evidence factors analysis for such a scenario in which the sources of bias overlap for the two evidence factors but are different in their relative size between the two evidence factors.

Recall that Section 4.2 provides the amplification of the sensitivity parameters Γ_{bc} and Γ_{nm} in terms of the λ, δ_{bc} and δ_{nm}. There, u_1 and u_2 are assumed to be two separate unmeasured confounds. The relation of the unmeasured confounding, u_1 and u_2, and the exposure to treatment is model by bias level λ. The relation of u_1 and the broad case status is modeled by the bias level δ_{bc}. Finally, the relation of u_2 and the broad case status is modeled by the bias level δ_{nm}. In the following we allow for u_1 and u_2 to be influenced by overlapping factors.

For individual l, let v_{1l} and v_{2l} be unmeasured numbers summarizing two sets of unmeasured variables so that $0 \leq v_{1l}, v_{2l} \leq 1$. We allow for both variables to bias each analysis but to have varying importance in their relationship with the outcomes. We formalize this as

TABLE 3

Simulated power of a sensitivity analysis of combined evidence in a case-control study, where there is no unmeasured confounder and $\Pr(Z_l = 1) = 1/3$. *The response is simulated from* $\beta + t_3/\sqrt{3}$ *if* $Z_l = 1$ *and* $t_3/\sqrt{3}$ *if* $Z_l = 0$. *There are* 1000 *narrow cases and* 1000 *marginal cases with* 2000 *controls. Based on* 10,000 *iterations. Fisher = Fisher's combination method, tP = truncated product method with* $\tilde{\alpha} = 0.20$

		$\beta = 0$		$\beta = 0.2$		$\beta = 0.4$		$\beta = 0.6$	
Γ_{nm}	Γ_{bc}	Fisher	tP	Fisher	tP	Fisher	tP	Fisher	tP
1	1	5	5	100	100	100	100	100	100
	1.5	1	1.5	0.5	0.5	100	100	100	100
	2	1	1.5	0	0.1	15	18	100	100
	2.5	1	1.5	0	0.1	0	0	98	98
1.25	1.25	0	0	47	54	100	100	100	100
	2	0	0	0	0	14	18	100	100
	2.75	0	0	0	0	0	0	71	77
	3.5	0	0	0	0	0	0	0.2	0.3
1.5	1.5	0	0	0.2	0.3	100	100	100	100
	2.5	0	0	0	0	0	0	98	98
	3.5	0	0	0	0	0	0	0.2	0.3
1.75	1.75	0	0	0	0	82	86	100	100
	2	0	0	0	0	14	18	100	100
	3.25	0	0	0	0	0	0	3	5
2	2	0	0	0	0	14	18	100	100
	2.5	0	0	0	0	0	0	98	98
	3	0	0	0	0	0	0	24	30
	3.5	0	0	0	0	0	0	0.2	0.3
2.25	2.25	0	0	0	0	0.2	0.4	100	100
	2.5	0	0	0	0	0	0	98	98
	3	0	0	0	0	0	0	24	30

follows. Let $u_{1l} = \psi_1 v_{1l} + \psi_2 v_{2l}$ where $\psi_1, \psi_2 \geq 0$, $\psi_1 + \psi_2 = 1$ and ψ_1 is larger than ψ_2. Also, let $u_{2l} = \tilde{\psi}_1 v_{1l} + \tilde{\psi}_2 v_{2l}$ where $\tilde{\psi}_1, \tilde{\psi}_2 \geq 0$, $\tilde{\psi}_1 + \tilde{\psi}_2 = 1$ and $\tilde{\psi}_2$ is larger than $\tilde{\psi}_1$. The fractions ψ_1, ψ_2, $\tilde{\psi}_1$ and $\tilde{\psi}_2$ are fixed numbers. The unmeasured confounders v_{1l} and v_{2l} relate to the broad case status and the narrow case status by models (4.5) and (4.6) via the variables u_{1l} and u_{2l}.

As for the relation between the unmeasured confounders v_{1l}, v_{2l} and the observed exposure to treatment, for two units $i1$ and $i2$ with the same observed covariates we write, for $z_{i1} + z_{i2} = 1$,

$$\Pr(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \mathcal{C}, \mathbf{x}_{i1} = \mathbf{x}_{i2}, Z_{i1} + Z_{i2} = 1) = \frac{\exp\{\lambda(z_{i1}\omega_{i1} + z_{i2}\omega_{i2})\}}{\exp(\lambda\omega_{i1}) + \exp(\lambda\omega_{i2})}, \tag{6.1}$$

where

$$\omega_l = \zeta_1 v_{1l} + \zeta_2 v_{2l} \quad \text{for } l = 1, \ldots, L;\ \zeta_1, \zeta_2 \geq 0,\ \zeta_1 + \zeta_2 = 1. \tag{6.2}$$

Now, consider the amplification of the sensitivity parameters Γ_{bc} and Γ_{nm} under the model specified by equations (6.1), (6.2) and (4.5) and (4.6) with $u_{1l} = \psi_1 v_{1l} + \psi_2 v_{2l}$ and $u_{2l} = \tilde{\psi}_1 v_{1l} + \tilde{\psi}_2 v_{2l}$. This can be communicated under three different scenarios depending on the source of bias under doubt—either bias from one of v_1 or v_2 or bias from both v_1 and v_2. Assume a value of λ in model (6.1)–(6.2). We find the parameters δ_{bc} and δ_{nm} from λ and Γ_{bc}, Γ_{nm}. Let $\Lambda = \exp(\lambda)$, $\Delta_{bc} = \exp(\delta_{bc})$ and $\Delta_{nm} = \exp(\delta_{nm})$. Then, (i) if only v_1 is the bias in question, that is, we put the restriction $v_{2,l} = v_{2,l'}$, then $\Delta_{bc} = \{(\Lambda\Gamma_{bc} - 1)/$

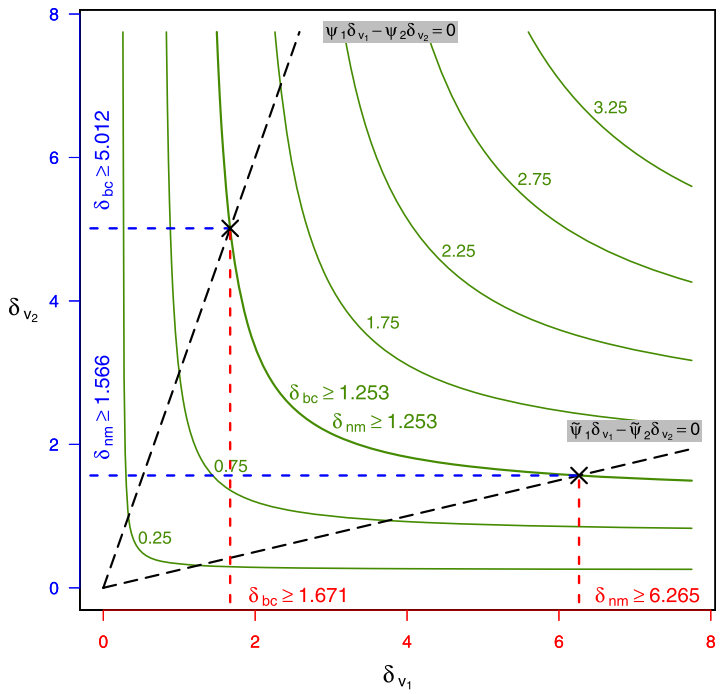

FIG. 2. *Level of bias from unmeasured confounding plotted under three speculations—bias only from v_1, plotted on the x-axis and in "red"; bias only from v_2, plotted on the y-axis and in "blue"; and biases from both v_1 and v_2, plotted in "green" contours. The contours are of the function $f(\delta_{v_1}, \delta_{v_2}) = (1/\delta_{v_1} + 1/\delta_{v_2})^{-1}$. Here, $\psi_1 = 3/4$, $\psi_2 = 1/4$, $\tilde{\psi}_1 = 1/5$ and $\tilde{\psi}_2 = 4/5$. The bias levels δ_{bc} and δ_{nm} change with the speculation, and the required bias level is minimized when biases from both v_1 and v_2 are assumed.*

$(\Lambda - \Gamma_{bc})\}^{1/\psi_1}$ and $\Delta_{nm} = \{(\Lambda\Gamma_{nm} - 1)/(\Lambda - \Gamma_{nm})\}^{1/\tilde{\psi}_1}$. This correspondence holds with $|v_{1,i1} - v_{1,i2}| = 1$. (ii) If only v_2 is the bias in question, that is, we put the restriction $v_{1,l} = v_{1,l'}$, then $\Delta_{bc} = \{(\Lambda\Gamma_{bc} - 1)/(\Lambda - \Gamma_{bc})\}^{1/\psi_2}$, $\Delta_{nm} = \{(\Lambda\Gamma_{nm} - 1)/(\Lambda - \Gamma_{nm})\}^{1/\tilde{\psi}_2}$ and $|v_{2,i1} - v_{2,i2}| = 1$. (iii) Finally, if both the confounders v_1 and v_2 are in question, then $\Delta_{bc} = (\Lambda\Gamma_{bc} - 1)/(\Lambda - \Gamma_{bc})$ and $\Delta_{bc} = (\Lambda\Gamma_{nm} - 1)/(\Lambda - \Gamma_{nm})$. This correspondence holds with $|v_{1,i1} - v_{1,i2}| = 1$ and $|v_{2,i1} - v_{2,i2}| = 1$. A closer look at these formulas immediately shows that bias parameters $\delta_{bc} = \log(\Delta_{bc})$ and $\delta_{nm} = \log(\Delta_{nm})$ change wildly across the scenarios.

Guided by the above calculations, Figure 2 provides an illustration of the influence of unmeasured confounders on the broad case status, δ_{bc}, and on the narrow case status to a marginal case status, δ_{nm}. In this illustration we assume $\psi_1 = 3/4$, so that, in determining a broad case status, the magnitude of unmeasured confounding from v_1 over v_2 has the ratio 3 : 1. Whereas, in determining a narrow case status to a marginal case status, the magnitude of unmeasured confounding from v_1 over v_2 has the ratio 1:4, that is, $\tilde{\psi}_1 = 1/5$. The plot considers three critics, showed in three colors, with different positions on their beliefs in the source of bias from unmeasured confounding. The first critic assumes bias only from v_1, the second critic assumes bias only from v_2 and, finally, the third critic assumes biases from both v_1 and v_2. The x-axis on the plot (in red) shows the amount of bias the first critic would have to assume; the y-axis on the plot (in blue) shows the amount of bias the second critic would have to assume, and, finally, the green curves show the amount of bias the third critic would have to assume. For example, the plot highlights the situation where the critics want to explain the sensitivity of the comparisons at level $\Gamma_{bc} = 2$ and $\Gamma_{nm} = 2$, and all of them speculate $\Lambda = 4$. The first critic would have to assume biases at the amounts of $\delta_{bc} \geq 1.671$ and $\delta_{nm} \geq 6.265$. The second critic would have to assume biases at the amounts of

$\delta_{bc} \geq 1.566$ and $\delta_{nm} \geq 5.012$. The third critic, however, can assume bias levels of $\delta_{bc} \geq 1.253$ and $\delta_{nm} \geq 1.253$. Hence, unless a skeptic of the study assumes unmeasured confounding from both sources of bias mechanisms she would be forced to consider a larger influence of unmeasured confounding in one case definition over the other.

Thus, when the factors overlap but do not completely overlap in their sources of bias, evidence factors will be useful in narrowing the range of explanations for how an observed association could not be causal.

7. Results: Efficacy of screening sigmoidoscopy. In our study of mortality from colorectal cancer and screening sigmoidoscopy, the two evidence factors analyses are summarized in Table 4. The count for screening sigmoidoscopy represent the number of individuals who had a screening procedure in 10 years before the reference date. The raw odds ratio, without controlling for any covariates, of screening sigmoidoscopy between proximal and distal cancer cases is 0.63 (95% CI, 0.55 to 0.72) and that between all colorectal cancer cases and controls is 0.64 (95% CI, 0.50 to 0.81). To control for important covariates, we utilize the matched sets we constructed in Section 2.1. Using this matched sets design, the p-value for efficacy of screening sigmoidoscopy for the distal colorectal cancer cases vs. the proximal colorectal cancer cases is 2.3×10^{-5}, with the corresponding odds ratio 0.60 (95% CI, 0.46 to 0.76). The p-value for all cases (distal and proximal) vs. the matched controls is 5.0×10^{-11}, with odds ratio 0.62 (95% CI, 0.54 to 0.72) (this result is similar to previously reported odds ratios; see Atkin et al. (2010) and Segnan et al. (2011)).

We further conduct a sensitivity analysis to assess whether possible covariates, which were not controlled for in our study, may have been the reason behind the observed association above. Being consistent with the notation of Section 4, we consider two sensitivity parameters Γ_{nm} and Γ_{bc} for the two comparisons. A value of 1 for a sensitivity parameter would say that there is no bias from unmeasured confounding in the respective comparison, and the higher the value is of the parameter, the bigger is the bias. Figure 3 shows the bias levels where the combined evidence for a beneficial effect of screening sigmoidoscopy is sensitive. The p-value upper bounds for each bias level of the two evidence factors are combined using the truncated product method with $\tilde{\alpha} = 0.20$. As can be seen in this plot, only a substantial amount of bias in both comparisons could explain the observed association in the data if, in fact, the null hypothesis is true. For example, with a maximum bias of $\Gamma_{nm} = 1.4$ in the comparison of distal cancer cases to proximal cancer cases, the combined evidence is sensitive only when the bias in the second comparison of all colorectal cases to the controls is larger than $\Gamma_{bc} = 1.45$. The overall evidence remains insensitive for $\Gamma_{nm} = 2$ when $\Gamma_{bc} \leq 1.35$. Thus, the overall evidence for the efficacy of the procedure is strengthened compared to evidence from an analysis that only looks at the screening rates between all colorectal cancer cases and controls. The maximum p-values are calculated using the "mh" function in `R` package `sensitivity2x2xk`.

TABLE 4

Screening sigmoidoscopy and colorectal cancer summary data. Numbers in the parentheses show the 95% *confidence intervals*

	Distal cancer cases	Proximal cancer cases	All colorectal cancer cases	Controls
No screening sigmoidoscopy	678	662	1340	2538
Screening sigmoidoscopy	144	224	368	1097
Odds ratio from matched sets	0.60 (0.46 to 0.76)		0.62 (0.54 to 0.72)	
p-value from matched sets	2.3×10^{-5}		5.0×10^{-11}	

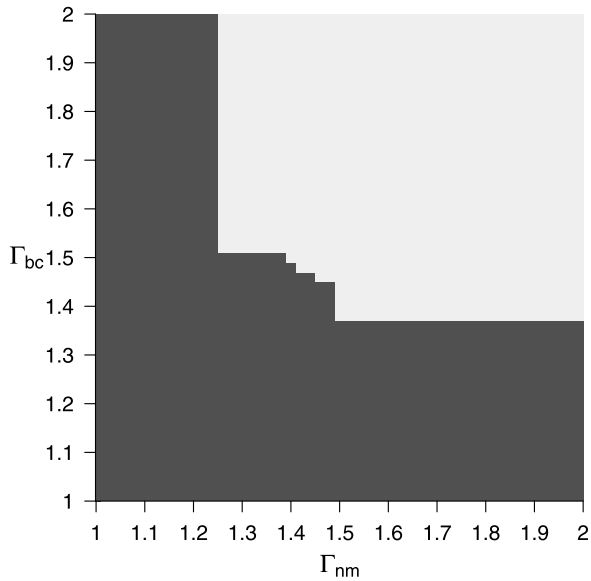

FIG. 3. *Sensitivity analysis of the efficacy of screening sigmoidoscopy in reducing mortality from colorectal cancer. The darker gray color represents the bias levels where the combined evidence for a beneficial effect of screening sigmoidoscopy is sensitive.*

To better understand which part of the evidence is contributing to our inferences about the effect of sigmoidoscopy screening, we can use closed testing (Marcus, Peritz and Gabriel (1976)) as in Karmakar, French and Small (2019). When both biases are small, suppose $\Gamma_{nm} = \Gamma_{bc} = 1.1$, by the closed testing procedure, the joint evidence is insensitive, and both evidence factors are also insensitive with $P_{nm,1.1} = 1.36 \times 10^{-7}$ and $P_{bc,1.3} = 0.0005$. The closed testing procedure also says that, when $\Gamma_{nm} = 1.5$ and $\Gamma_{bc} = 1.4$, the comparison of proximal to distal cancer cases is sensitive with maximum possible p-value of $P_{nm,1.5} = 0.21$, but there is evidence from the comparison of all colorectal cancer cases to the controls which is insensitive with a maximum possible p-value of $P_{bc,1.4} = 0.034$. Recall from the discussion of Section 4.2 that the pair of bias levels $\Gamma_{nm} = 1.5$ and $\Gamma_{bc} = 1.4$ is equivalent to an effect of unmeasured confounders that doubles the chance of a sigmoidoscopy screening for a case relative to a control, while also increasing the chance of death from colorectal cancer by 5/3-fold and increasing the chance of death from a proximal colorectal cancer over a distal colorectal cancer by twofold. On the other hand, if the effect of unmeasured confounders is smaller on being a proximal cancer case so that it increases the chance of death from proximal colorectal cancer over a distal cancer only by 5/3-fold but increases the chance of death by any colorectal cancer by twofold, the joint evidence is sensitive to such unmeasured confounders. The closed testing procedure for two or for many evidence factors and plots similar to Figure 3 can be produced by the `R` package `evidenceFactors` available from `CRAN` (R Core Team (2020)).

8. Discussion. In this paper we have developed evidence factors in a case-control study in which there is a narrow and a broad case definition. These evidence factors are formed by two sets of comparisons, the first one comparing narrow cases to marginal cases and the second one comparing all cases to controls. Use of these evidence factors in a case-control study can provide better insight into the study especially in a discussion and analysis of possible bias in the study.

In the sigmoidoscopy study considered in this paper, the elaborate theory (Section 1.1 and 1.3) suggested that, if there is an efficacy of sigmoidoscopy screening in reducing mortality

from colorectal cancer, the benefit should be larger for the proximal cancer cases compared to distal cancer cases and for any colorectal cancer case over controls. Following this theory, the evidence factors were thus useful in assessing the hypothesis of no benefit of sigmoidoscopy screening. While the standard discussion of evidence factors analyses emphasizes that the biases affecting the different factors are different (Section 1.2), for the sigmoidoscopy study it was more likely the biases overlap but not completely. For case-control studies, this paper also shows that the evidence factors analyses also strengthens the evidence for a causal effect when the biases from unmeasured confounders affecting the different analyses may overlap.

The technical results of Section 5 can be extended to more complex designs, for example, to designs with more than two types of cases (see Keogh and Cox (2014)) using more complex notation. But these technical results are only a part of what makes an evidence factor useful for a case-control study. It is also equally important that the factors are coherent with the elaborate theory of a causal effect of an exposure; for two case definitions other examples, where an evidence factors analysis may be considered, are discussed in the final subsection. Lastly, it would also be important to establish that under overlapping biases, which is likely more prominent when there are multiple types of cases, the multiple analyses considered still strengthens the evidence against a large number of plausible patterns of biases. Regarding this point, for the arguments of Figure 2 in Section 6 to work, one has to think of appropriate extensions of the models in equations (4.5) and (4.6). Such extensions are not readily available in the literature. We leave these developments as a potential future research direction.

Our study paired narrow cases to marginal cases on the observed covariates and included their controls in the matched sets and, then, put the remaining marginal cases in matched sets with their controls. Other matching methods could be used, for example, full matching (Hansen (2004)) and variable ratio matching (Ming and Rosenbaum (2000), Pimentel, Yoon and Keele (2015)).

8.1. *Other examples with multiple case definitions.* In certain diseases, like cancer in the body of the uterus, atherosclerosis, hypertension and mental illness, multiple case definitions are considered or often necessary (Acheson (1979), Cole (1979), Cohen et al. (2005)). Some other specific studies where multiple case definitions have been considered are discussed here. These studies illustrate various ways to design a broad case vs. narrow case distinction in case-control studies. In a study to assess whether statin causes peripheral neuropathy, Gaist et al. (2002) classify the neuropathy cases as definite and nondefinite cases of idiopathic peripheral neuropathy based on the intensity of the symptom and the quality of the clinical information. In the terminology of the present paper, the definite cases would be the narrow cases where the association, if present, would be stronger compared to the marginal cases, that is, the nondefinite cases. Small et al. (2013) use an illustrative case-control study for physical abuse by parents in childhood and tendency for more anger in adulthood. In this study the cases were split in two definitions based on whether or not anger score was on a higher range. Here, a case on a higher quantile of anger score could be defined as a narrow case. As a final example, in an effort to understand association between genetic traits and cerebral malaria, Small et al. (2017) consider cerebral malaria cases with and without retinopathy. The World Health Organization (WHO) defines a child as having cerebral malaria when the child is in a coma (cannot localize a painful stimulus), has malaria parasites in his or her blood and has no other known cause of the coma. This definition is not specific as hospitals in malaria-endemic areas often lack diagnostic facilities to identify nonmalarial causes of coma and many children in malaria endemic areas have nonsymptomatic malaria infections. There are characteristic retinal abnormalities (retinopathy) that increase the specificity of a cerebral malaria diagnosis (Taylor et al. (2004)). Cerebral malaria cases with such retinal abnormalities could be considered as narrow cases and those without the retinal abnormalities could be considered as marginal cases.

APPENDIX: PROOF OF THE LEMMAS

PROOF OF LEMMA 5.2. First we note that T_{nm} is a function of $Y_{n\{s\}}$ which are, simply, linear functions of $Z_{n\{si\}}$. Given the strata, from equation (4.1) we have that the maximum p-value of the narrow vs. marginal comparison, $P_{nm,\Gamma_{nm}}$, is computed based on the conditional distributions $\{[Z_{n\{si\}} \mid \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}]\}$. Combining these facts, we get the first result that marginally $P_{nm,\Gamma_{nm}}$ is a function of $\{Z_{n\{si\}}\}$ and $\sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}$.

Next, we note that T_{bc} is a function of $\sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}$. Now, by looking at equation (4.2), $P_{bc,\Gamma_{bc}}$ is computed based on the family of conditional distributions $\{[\sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}} \mid \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}} + \sum_{i\in[c_s]} Z_{c\{si\}}]\}$. Consequently, $P_{bc,\Gamma_{bc}}$ is determined by the number of exposed cases $\{\sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}\}$ and the total number of exposed individuals $\{\sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}} + \sum_{i\in[c_s]} Z_{c\{si\}}\}$. But it is enough to know whether each control is exposed or not, that is, $Z_{c\{si\}}$, to know the number of exposed cases when we have the information on total number of exposed units. Hence, the result is proved. □

PROOF OF LEMMA 5.3. For parts (i) and (ii) note that p-values or their upper bounds are valid p-values, thus, are stochastically larger than Unif[0, 1]. Parts (iii) and (iv) follows from (i) and (ii) simply by marginalizing since marginalization preserves stochastic ordering. □

PROOF OF LEMMA 5.4. Note that, since conditional on $\sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}$ the random variables $Z_{n\{si\}}$ and $Z_{c\{si\}}$ are independently distributed, by Lemma 5.2 the conditional distribution in the statement of the lemma is same as $[P_{nm,\Gamma_{nm}} \mid \sum_{i\in[n_s]} Z_{n\{si\}} + \sum_{i\in[m_s]} Z_{m\{si\}}]$. Now, the result follows from part (i) of Lemma 5.3. □

PROOF OF LEMMA 5.5. We can write for any $0 \le p, q, \le 1$, the conditional probability as

$$\begin{aligned}\Pr(P_2 \le q \mid P_1 \le p) &\overset{\text{by C1}}{=} \Pr\big(P_2 \le q \mid \{V_1 : P_1 \le p\}\big)\\ &= \mathrm{E}\big[Pr(P_2 \le q \mid V_1) \mid \{V_1 : P_1 \le p\}\big]\\ &\overset{\text{by C2}}{\le} \mathrm{E}\big[q \mid \{V_1 : P_1 \le p\}\big] = q.\end{aligned}$$

The second equality above follows from the tower property of conditional expectation. The lemma then follows. □

Software. An `R` package `evidenceFactors`, available from `CRAN` (R Core Team (2020)), contains code for reproducing the simulation results of Section 5.1, and code used for analyzing the sigmoidoscopy study.

Acknowledgments. The authors thank Dr. Noel Weiss for helpful discussion that structured the paper.

Grant support. This study was supported by an award (number R01CA213645 and number U01CA151736) from the National Cancer Institute of the National Institute of Health. The views expressed here are those of the authors only and do not represent any official position of the National Cancer Institute or National Institutes of Health.

Disclaimer. Dr. Doubeni is a member of the U.S. Preventive Services Task Force (USPSTF). This article does not necessarily represent the views and policies of the USPSTF.

REFERENCES

ACHESON, E. D. (1979). Comment on "The evolving case-control study." *J. Chronic Dis*. **32** 28–29.

ATKIN, W. S., EDWARDS, R., KARLJ-HANS, I. et al. (2010). Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: A multicentre randomized controlled trial. *Lancet* **375** 1624–1633.

BAZZANO, L. A., HE, J., MUNTNER, P., VUPPUTURI, S. and WHELTON, P. K. (2003). Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the United States. *Ann. Intern. Med*. **138** 891–897.

BECKER, B. J. (1994). Combining significance levels. In *The Handbook of Research Synthesis* 215–230. Russell Sage Foundation, Thousand Oaks, CA.

BIZZOZERO, O. J., JOHNSON, K. G. and CIOCCO, A. (1966). Radiation related leukemia in Hiroshima and Nagasaki, 1946–1964. *N. Engl. J. Med*. **274** 1095–1101.

COCHRAN, W. G. (1965). The planning of observational studies of human populations. *J. Roy. Statist. Soc. Ser. A* **128** 134–155. Reprinted in *Readings in Economic Statistics and Econometrics* (A. Zellner, ed.), Little Brown, Boston, MA, pp. 11–36 (1968).

COHEN, J. C., KISS, R. S., PERTSEMLIDIS, A., KOTOWSKI, I. K., GRAHAM, R., KIM GARCIA, C. and HOBBS, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutation in PCSK9. *Nat. Genet*. **37** 161–165.

COLE, P. (1979). The evolving case-control study. *J. Chronic Dis*. **32** 15–27. https://doi.org/10.1016/0021-9681(79)90006-7

DOUBENI, C. A., MAJOR, J. M., LAIYEMO, A. O., SCHOOTMAN, M., ZAUBER, A. G., HOLLENBECK, A. R., SINHA, R. and ALLISON, J. (2012). Contribution of behavioral risk factors and obesity to socioeconomic differences in colorectal cancer incidence. *J. Natl. Cancer Inst*. **104** 1353–1362.

DOUBENI, C. A., CORLEY, D. A., QUINN, V. P., JENSEN, C. D., ZAUBER, A. G., GOODMAN, M., JOHNSON, J. R., MEHTA, S. J., BECERRA, T. A. et al. (2018). Effectiveness of screening colonoscopy in reducing the risk of death from right and left colon cancer: A large community-based study. *Gut* **67** 291–298. https://doi.org/10.1136/gutjnl-2016-312712

ELDRIDGE, R. C., DOUBENI, C. A., FLETCHER, R. H., ZAUBER, A. G., CORLEY, D. A., DORIA-ROSE, V. P. and GOODMAN, M. (2013). Uncontrolled confounding in studies of screening effectiveness: An example of colonoscopy. *J. Med. Screen*. **20** 198–207. https://doi.org/10.1177/0969141313508282

FISHER, R. A. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

GAIST, D., JEPPESEN, U., ANDERSEN, M., GARCÍA RODRÍGUEZ, A. L., HALLAS, J. and SINDRUP, H. S. (2002). Statins and risk of polyneuropathy: A case-control study. *Neurology* **58** 1333–1337.

GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85** 907–920.

GOODMAN, M., FLETCHER, R. H., DORIA-ROSE, V. P., JENSEN, C. D., ZEBROWSKI, A. M., BECERRA, T. A., QUINN, V. P., ZAUBER, A. G., CORLEY, D. A. et al. (2015). Observational methods to assess the effectiveness of screening colonoscopy in reducing right colon cancer mortality risk: SCOLAR. *J. Comp. Eff. Res*. **4** 541–551.

HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc*. **99** 609–618. MR2086387 https://doi.org/10.1198/016214504000000647

HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist*. **15** 609–627. MR2280151 https://doi.org/10.1198/106186006X137047

HILL, A. B. (1965). The environment and disease: Association or causation? *Proc. R. Soc. Med*. **58** 295–300.

HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Amer. Statist. Assoc*. **108** 135–148. MR3174608 https://doi.org/10.1080/01621459.2012.742018

JOSEPH, D. A., MEESTER, R. G. S., ZAUBER, A. G., MANNINEN, D. L., WINGES, L., DONG, F. B., PEAKER, B. and VAN BALLEGOOIJEN, M. (2016). Colorectal cancer screening: Estimated future colonoscopy need and current volume and capacity. *Cancer* **122** 2479–2486.

KARMAKAR, B., FRENCH, B. and SMALL, D. S. (2019). Integrating the evidence from evidence factors in observational studies. *Biometrika* **106** 353–367. MR3949308 https://doi.org/10.1093/biomet/asz003

KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2020). Using evidence factors to clarify exposure biomarkers. *Am. J. Epidemiol*. To appear. https://doi.org/10.1093/aje/kwz263

KEOGH, R. H. and COX, D. R. (2014). *Case-Control Studies*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **4**. Cambridge Univ. Press, Cambridge. MR3443808 https://doi.org/10.1017/CBO9781139094757

LEWIS, E. B. (1963). Leukemia, multiple myeloma, and aplastic anemia in American radiologists. *Science* **142** 1492–1494.

LIPTAK, T. (1958). On the combination of independent tests. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl*. **3** 171–197.

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056 https://doi.org/10.1093/biomet/63.3.655

MING, K. and ROSENBAUM, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56** 118–124.

MISSIAGLIA, E., JACOBS, B., D'ARIO, G., NARZO, A. F. D., SONESON, C., BUDINSKA, E., POPOVICI, V., VECCHIONE, L., GERSTER, S. et al. (2014). Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. *Ann. Oncol.* **25** 1995–2001. https://doi.org/10.1093/annonc/mdu275

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Ann. Agric. Sci.* **10** 1–51 (in Polish). [Reprinted in English with discussion by T. Speed and D. B. Rubin in *Statist. Sci.* **5** (1990) 463–480.] MR1092986

PIMENTEL, S. D., YOON, F. and KEELE, L. (2015). Variable-ratio matching with fine balance in a study of the Peer Health Exchange. *Stat. Med.* **34** 4070–4082. MR3431322 https://doi.org/10.1002/sim.6593

POPPER, K. R. (1959). *The Logic of Scientific Discovery*. Hutchinson and Co., Ltd., London. MR0107593

U. S. PREVENTIVE SERVICES TASK FORCE, BIBBINS-DOMINGO, K., GROSSMAN, D. C. et al. (2016). Screening for colorectal cancer: US preventive services task force recommendation statement. *J. Am. Med. Assoc.* **315** 2564–2575.

ROSENBAUM, P. R. (1991). Sensitivity analysis for matched case-control studies. *Biometrics* **47** 87–100. MR1108691 https://doi.org/10.2307/2532498

ROSENBAUM, P. R. (2001). Replicating effects and biases. *Amer. Statist.* **55** 223–227. MR1963397 https://doi.org/10.1198/000313001317098220

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

ROSENBAUM, P. R. (2010). Evidence factors in observational studies. *Biometrika* **97** 333–345. MR2650742 https://doi.org/10.1093/biomet/asq019

ROSENBAUM, P. R. (2011). Some approximate evidence factors in observational studies. *J. Amer. Statist. Assoc.* **106** 285–295. MR2816721 https://doi.org/10.1198/jasa.2011.tm10422

ROSENBAUM, P. R. (2017). The general structure of evidence factors in observational studies. *Statist. Sci.* **32** 514–530. MR3730520 https://doi.org/10.1214/17-STS621

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534 https://doi.org/10.1198/016214506000001059

ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. MR2750570 https://doi.org/10.1198/jasa.2009.tm08470

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

SEGNAN, N., ARMAROLI, P., BONELLI, L. et al. (2011). Once-only sigmoidoscopy in colorectal cancer screening: Follow-up findings of the Italian Randomized Control Trial—SCORE. *J. Natl. Cancer Inst.* **103** 1310–1322.

SELBY, J. V., FRIEDMAN, G. D., QUESENBERRY, C. P. and WEISS, N. (1992). A case-control study of screening sigmoidoscopy and mortality from colorectal cancer. *N. Engl. J. Med.* **326** 653–657.

SHAKED, M. and SHANTHIKUMAR, J. G. (2007). *Stochastic Orders. Springer Series in Statistics*. Springer, New York. MR2265633 https://doi.org/10.1007/978-0-387-34675-5

SMALL, D. S., CHENG, J., HALLORAN, M. E. and ROSENBAUM, P. R. (2013). Case definition and design sensitivity. *J. Amer. Statist. Assoc.* **108** 1457–1468. MR3174721 https://doi.org/10.1080/01621459.2013.820660

SMALL, D. S., TAYLOR, T. E., POSTELS, D. G., BEARE, N. A., CHENG, J., MACCORMICK, I. J. and SEYDEL, K. B. (2017). Evidence from a natural experiment that malaria parasitemia is pathogenic in retinopathy-negative cerebral malaria. *eLife* **6**. https://doi.org/10.7554/eLife.23699

TAYLOR, T. E., FU, W. J., CARR, R. A., WHITTEN, R. O., MUELLER, J. S., FOSIKO, N. G., LEWALLEN, S., LIOMBA, N. G., MOLYNEUX, M. E. et al. (2004). Differentiating the pathologies of cerebral malaria by postmortem parasite counts. *Nat. Med.* **10** 143–145.

WITTGENSTEIN, L. (1958). *Philosophical Investigations*, 2nd ed. The Macmillan Co., New York. MR0078292

ZAYKIN, D., ZHIVOTOVSKY, L. A., WESTFALL, P. and WEIR, B. (2002). Truncated product method for combining p-values. *Genet. Epidemiol.* **22** 170–185.

ZHANG, K., SMALL, D. S., LORCH, S., SRINIVAS, S. and ROSENBAUM, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *J. Amer. Statist. Assoc.* **106** 511–524. MR2847966 https://doi.org/10.1198/jasa.2011.ap10604

ZUBIZARRETA, J. R., NEUMAN, M., SILBER, J. H. and ROSENBAUM, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study of alternative forms of anesthesia. *J. Amer. Statist. Assoc.* **107** 901–915. MR3010879 https://doi.org/10.1080/01621459.2012.682533

R CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available at http://www.R-project.org.