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 5 years for men and women above 50 at average risk (U.S. Preventive Services Task Force, 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 setting, 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 of interest if that exposure in fact has an effect and the other being “broad” in that it may have more heterogeneous causes.

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 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-third 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 of treatment effect, if there is concern that some unknown bias is the reason for the assessment of significant treatment effect, we should consider if it possible to replicate the study without repeating the bias ( Cochran, 1965, Section 4.1).

For example, 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 of the effect of radiation that does not replicate the bias is a comparison of the leukemia risk in people living in Japan near where the atomic bombs were dropped in Hiroshima and Nagasaki at the end of World War II to people living further from the bombs (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. The finding of higher rates of leukemia incidence in radiologists and in people living near where the atomic bombs were dropped strengthens the evidence for a causal effect since the skeptic would have assume two sources of bias rather than just one (Rosenbaum, 2001).

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 (in the sense of (1.1)), 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 (2017), which we briefly review here. Suppose two analyses are performed to test for the null hypothesis; the first analysis requires a set of assumptions \( A_1 \) and the second analysis requires a second set of assumptions \( A_2 \). Let \( P_1 \) and \( P_2 \) be the corresponding p-values from these two analyses. Then, to be evidence factors, we require the independence of the analyses in the following sense – under the null hypothesis, when both assumptions \( A_1 \) and \( A_2 \) hold for \((p_1, p_2) \in [0, 1]^2\)

\[
\Pr(P_1 \leq p_1, P_2 \leq p_2) \leq p_1 p_2.
\] (1.1)

If the \( P_1 \) and \( P_2 \) were independent, then (1.1) would be an equality. The inequality in (1.1) 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 that treating the two p-values as independent when using methods such as Fisher’s combination test (Fisher, 1932) to combine the two tests would be conservative – this is the “near independence” of the p-values 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 of evidence 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). With the near independence from (1.1), if both analyses are significant, then we have two separate pieces of evidence against the null hypothesis of no treatment effect which are robust to violations of one of the two sets of assumptions \( A_1 \) or \( A_2 \). Both assumptions would have to be violated in order for there not to be evidence of a treatment effect. We give
a formal definition of evidence factors in Section 5.

An example of the use of evidence factors discussed in Rosenbaum (2010, 2015) that builds on a study of Silber, et al. (2007) is concerned with the effect of intensive chemotherapy for women with ovarian cancer. Chemotherapy for ovarian cancer can be administered by a medical oncologist (MO) or gynecological oncologist (GO); MOs tend to give higher doses of chemotherapy. In the same study, Rosenbaum (2010, 2015) considers the comparison of outcomes in pairs of women, one treated by an MO and one by a GO, who are matched for clinical stage and tumor grade, by a Wilcoxon signed rank test and also considers the comparison of the difference in outcomes between the MO patient and GO patient among pairs who have a higher difference in chemotherapy dose between the MO patient and GO patient vs. pairs with a lower difference in chemotherapy dose, by Kendall’s test. These tests are statistically independent under the null hypothesis of no treatment effect and also are likely to be biased by different factors – the comparison within pairs (the Wilcoxon signed rank test) might be biased by aspects of the way in which patients are referred to particular physicians while the comparison among pairs (Kendall’s test) might be biased by considerations that the physicians can see in their patients. Rosenbaum (2010) found that both comparisons provide evidence that patients receiving higher amounts of chemotherapy have more weeks of toxicity after treatment. In contrast to a finding from just one comparison, the supportive findings from the two comparisons strengthens the evidence for more chemotherapy causally increasing toxicity because, in order for a skeptic to claim that more chemotherapy does not increase toxicity, the skeptic would have to assert two sources of bias rather than just one. Other examples of the use of evidence factors include Zhang et al. (2011) and Zubizarreta et al. (2012).

Rosenbaum (2017) provides a mathematical formulation for building evidence factors based on multiple treatment assignment mechanisms. Starting with a set of \( n \) study units Rosenbaum (2017) showed how to construct evidence factors using the knit product of two subgroups of the symmetric group of size \( n \). In this construction the assumptions made in the two analyses state that the treatment is randomized in the corresponding subgroups.

Previous work has 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, this is the first demonstration of using differences in outcomes to develop evidence factors.
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, i.e. 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).

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). These two comparisons could be biased differently. 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) and 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 pattern (a) and pattern (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.

To test for patterns (a) and (b), we would like to use nearly independent test statistics in the sense of (1.1) so as to distinguish new evidence from the same evidence repeated twice. In other words, we would like to develop evidence factors associated with patterns (a) and (b). We develop a method for doing this in Section 4 and Section 5 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 and provide notation in Section 3.

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. Specifically, we consider the comparison of distal cancer cases to proximal cancer cases and also the
comparison of all cases (any colorectal cancer) to controls. In relation to sigmoidoscopy screening, distal cancer cases are narrow cases and proximal cancer cases are marginal cases. We will show later that these two comparisons constitute separate evidence (evidence factors). 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 Goodman et al. (2015); Doubeni et al. (2018). 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 fine balanced on gender. By pair matching the narrow and marginal cases, we obtained 822 matched sets consisting of one narrow case, one 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 gender, reference date and enrollment source, 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 as-

<table>
<thead>
<tr>
<th>Table 1</th>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of years</td>
<td>% from Center 1</td>
</tr>
<tr>
<td>Controls</td>
<td>12</td>
</tr>
<tr>
<td>Distal cancer cases</td>
<td>12</td>
</tr>
<tr>
<td>Proximal cancer cases</td>
<td>12</td>
</tr>
</tbody>
</table>

Fig 1. Reference date of the colorectal cancer cases and controls in the matched sets.
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 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 \( r_{Tl} \) and if spared exposure would have response \( r_{Cl} \). Thus the observed response for unit \( l \) is \( R_l = Z_l r_{Tl} + (1 - Z_l) r_{Cl} \). Consequently, we cannot observe \( r_{Tl} \) and \( r_{Cl} \) simultaneously for one unit (Neyman, 1923; Rubin, 1974). Now let \( x_l \) denote the observed pre-treatment covariates, i.e., 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). We gather population parameters to write the set \( \mathcal{F} = \{ (r_{Tl}, r_{Cl}, x_l, u_l) : l = 1, \ldots, L \} \). The hypothesis we are interested in studying is Fisher’s sharp null hypothesis of no treatment effect

\[
H_0 : r_{Tl} = r_{Cl}, \quad l = 1, \ldots, L.
\]

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 \( (x_i)'s \). 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

$$\Pr(Z_l = 1 \mid \mathcal{F}) = \frac{\exp\{\lambda(x_l) + \gamma u_l\}}{1 + \exp\{\lambda(x_l) + \gamma u_l\}},$$

where $\lambda(\cdot)$ 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, $x_l = x_{l'}$, under this model, their odds of exposure can vary at most by a factor of $\Gamma := \log(\text{exposure odds})$. Model (3.1) is equivalent to writing

$$\max_{1 \leq l, l' \leq L} \left\{ \frac{\Pr(Z_l = 1 \mid \mathcal{F})/\Pr(Z_l = 0 \mid \mathcal{F})}{\Pr(Z_{l'} = 1 \mid \mathcal{F})/\Pr(Z_{l'} = 0 \mid \mathcal{F})} : x_l = x_{l'} \right\} \leq \Gamma.$$

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 $\Gamma(\geq 1)$ 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 $\Gamma$ 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 individual has twice the odds of getting assigned treatment as the other or vice versa.

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

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

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, 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 $\Gamma$, i.e., (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 $\alpha$, the significance level, then we have evidence to reject the null hypothesis as long as the hidden bias is at most $\Gamma$. 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 \( H_0 \). Thus a sensitivity analysis looks for \( \Gamma \) values where the upper bound of (3.3) is larger than \( \alpha \).

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 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 non-cases in broad case definition are thus also non-cases in the narrow case definition, and are labeled as controls.

4.1. Matched strata for the comparisons. In Section 3 we have discussed a matching procedure for a comparison between cases and controls to test for \( H_0 \) when there is one standard case definition. Now as we have two case definitions, a similar matching argument can still be used with appropriate modifications. We discuss the details here.

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 \( i \)th 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 \( x_c\{si\} \) denotes the observed covariate for the \(i\)-th control in stratum \( s \). \( r_{m\{si\}}, r_{Cn\{si\}}, u_{c\{si\}}\) 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 versus marginal cases and broad cases versus 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
EVIDENCE FACTORS IN A CASE-CONTROL STUDY

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 non-negative constants given \( F \). 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

\[ \Pr \left( T_{nm} \geq k \mid \{ b_s \}, \{ m_s \}, \sum_{i \in [n_s]} Z_{n(si)}, \sum_{i \in [m_s]} Z_{m(si)}; F_b \right), \]

where \( F_b \) is the subset of \( F \) 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

\[ \Pr \left( T_{bc} \geq k \mid \{ b_s + c_s \}, \{ c_s \}, \sum_{i \in [n_s]} Z_{n(si)} + \sum_{i \in [m_s]} Z_{m(si)} + \sum_{i \in [c_s]} Z_{c(si)}; F \right). \]

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 \( \Gamma \) bounds the odds ratio of treatment assignment for all the units stratified on their observed covariates. But as noted in the previous subsection, 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 \( \Gamma_{nm} \) and \( \Gamma_{bc} \) for the narrow versus marginal and broad versus control comparisons respectively. Then \( \Gamma_{nm} \) measures the bias in treatment assignment among all the case units which are similar in their observed covariates and \( \Gamma_{bc} \) measures the bias in treatment assignment among all case and control units which are similar in their observed covariates.

Therefore, the comparison of narrow versus marginal cases would compute the upper bound on the p-value for \( T_{nm} \) based on the tail distribution (4.1) for sensitivity parameter \( \Gamma_{nm} \) and the broad cases versus controls comparison would compute the upper bound on the p-value for \( T_{bc} \) based on the tail distribution (4.2) for sensitivity parameter \( \Gamma_{bc} \).
distribution (4.2) for sensitivity parameter \( \Gamma_{bc} \). 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.

4.2. Two sensitivity parameters and their amplification. In a sensitivity analysis the sensitivity parameters \( \Gamma_{nm} \) and \( \Gamma_{bc} \) would be used to get the max p-values \( P_{nm, \Gamma_{nm}} \) and \( P_{bc, \Gamma_{bc}} \). How does an \( \Gamma_{nm} \) 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 \( \mathcal{C} = \{(x_l, u_{1l}, u_{2l}) : l = 1, \ldots, L\} \). As before, \( 0 \leq u_{1l} \leq 1 \) and \( 0 \leq u_{2l} \leq 1 \). A principal stratification 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 \( \lambda \), for \( z_{i1} + z_{i2} = 1 \), as

\[
\Pr(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \mathcal{C}, x_{i1} = x_{i2}, Z_{i1} + Z_{i2} = 1) = \frac{\exp\{\lambda(z_{i1}w_{i1} + z_{i2}w_{i2})\}}{\exp(\lambda w_{i1}) + \exp(\lambda w_{i2})},
\]

(4.4) where \( w_l = \xi_1 u_{1l} + \xi_2 u_{2l} \), 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 \( \lambda \) 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 $k_{C,l}^b$ whether the unit is a case and by $k_{C,l}^n$ whether the unit is a narrow case. Thus, $k_{C,l}^b = 1$ if the $l$th unit is a case, either narrow or marginal, when not exposed to the treatment and $k_{C,l}^b = 0$ if the unit is a control when not exposed to the treatment. Similarly, $k_{C,l}^n = 1$ if the $l$th unit is a narrow case when not exposed to the treatment and $k_{C,l}^n = 0$ otherwise. It might be helpful to think of $k_{C,l}^b$ and $k_{C,l}^n$ as being determined by $r_{C,l}$. For two units $i1$ and $i2$ with similar observed covariates, the following model relates the case label with the confounders:

\[
\frac{\Pr(k_{C,l}^b = 1, k_{C,l}^b = 0 \mid C, x_{i1} = x_{i2})}{\Pr(k_{C,l}^b = 0, k_{C,l}^b = 1 \mid C, x_{i1} = x_{i2})} = \exp\{\delta_{bc}(u_{1,i1} - u_{1,i2})\};
\]

\[
\frac{\Pr(k_{C,l}^n = 1, k_{C,l}^n = 0 \mid C, x_{i1} = x_{i2}, k_{C,l}^b = 1)}{\Pr(k_{C,l}^n = 0, k_{C,l}^n = 1 \mid C, x_{i1} = x_{i2}, k_{C,l}^b = 1)} = \exp\{\delta_{nm}(u_{2,i1} - u_{2,i2})\}.
\]

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

How do $\lambda$, $\delta_{bc}$ and $\delta_{nm}$ relate to the sensitivity parameters $\Gamma_{bc}$ and $\Gamma_{nm}$? 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 $\Gamma_{bc} = (\Delta_{bc} + 1)/(\Delta_{bc} + \Lambda)$ and $\Gamma_{bc} = (\Delta_{nm} + 1)/(\Delta_{nm} + \Lambda)$. These formulas allow one to interpret the result of a sensitivity analysis either using the sensitivity parameters $\Gamma_{bc}$ and $\Gamma_{nm}$ or, under model (4.3)–(4.6), using parameters $\lambda$, $\delta_{bc}$ and $\delta_{nm}$. For example, $\Gamma_{nm} = 2$, $\Gamma_{bc} = 2.5$ corresponds to $\Lambda = 4$, $\Delta_{nm} = 3.5$ and $\Delta_{bc} = 8$; $\Gamma_{nm} = 3$, $\Gamma_{bc} = 2$ 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, 2017). As discussed in Section 1.2, Karmakar, French, and Small (2017) provide a general formulation of evidence factors in observational study designs. We start this section by stating this definition. Readers interested in the results of the SCOLAR data analysis can skip this technical discussion and go directly to Section 5.1 and 7.
DEFINITION 1. A set $D$ is called a decreasing set if for any pair $(x, y)$ with $x \leq y$, if $y \in D$ then $x \in D$. For two random vectors $X$ and $Y$ we say that $X$ is stochastically larger than $Y$ if

$$\Pr(X \in D) \leq \Pr(Y \in D),$$

for all nondecreasing sets $D$. If $X$ is stochastically larger than $Y$ we write $X \succeq Y$.

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, $(P_{nm, \Gamma_{nm}}, P_{bc, \Gamma_{bc}}) \succeq (U_1, U_2)$ under the bias levels $\Gamma_{nm}$, $\Gamma_{bc}$ and under $H_0$, for two independent $\text{Unif}[0,1]$ random variables $U_1$ and $U_2$.

The main result of this section is the following theorem which says that according to the above definition, $(P_{nm, \Gamma_{nm}}, P_{bc, \Gamma_{bc}})$ form evidence factors.

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

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

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

and

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

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

**Lemma 5.3.** Under $H_0$, we have the following
Under says that the combined
\begin{itemize}
\item[(i)] \(\left[ P_{nm, \Gamma_{nm}} \mid \{ \sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}} \}, \mathcal{F}_b, \{ n_s \} \right] \Unif[0,1]. \)
\item[(ii)] \(\left[ P_{bc, \Gamma_{bc}} \mid \{ \sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}} + \sum_{i \in [c_s]} Z_{c\{si\}} \}, \mathcal{F}, \{ b_s + c_s \} \right] \Unif[0,1]. \)
\item[(iii)] \(P_{nm, \Gamma_{nm}} \Unif[0,1]. \)
\item[(iv)] \(P_{bc, \Gamma_{bc}} \Unif[0,1]. \)
\end{itemize}

The following lemma relies on the assumption of no interference in treat-
ment assignment among the units, which is to say \(Z_l\) and \(Z_{l'}\) are indepen-
dently distributed for two distinct units \(l\) and \(l'\).

**Lemma 5.4.** Under \(H_0\)
\[
\left[ P_{nm, \Gamma_{nm}} \mid \{ Z_{c\{si\}}, i \in [c_s] \} ; \sum_{i \in [n_s]} Z_{n\{si\}} + \sum_{i \in [m_s]} Z_{m\{si\}} \right] \Unif[0,1].
\]

**Lemma 5.5.** Suppose two random variables \(P_1\) and \(P_2\) satisfy
\begin{itemize}
\item[(C1)] random variable \(P_1\) is a function of random quantity \(V_1\),
\item[(C2)] \(P_2 | V_1 \Unif[0,1], \)
\end{itemize}
then for \(0 \leq q \leq 1\), \(Pr(P_2 \leq q | P_1) \leq q\), i.e., \([P_2 | P_1] \Unif[0,1]. \)

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, \([P_{nm, \Gamma_{nm}} \mid P_{bc, \Gamma_{bc}}] \Unif[0,1]. \)

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) §6B. (\(U_1, U_2\) being an independent pair is a conditionally increasing in sequence (CIS). Then combining this with the facts that \(P_{bc, \Gamma_{bc}} \Unif[0,1]\) (by
Lemma 5.3) and \([P_{nm, \Gamma_{nm}} \mid P_{bc, \Gamma_{bc}}] \Unif[0,1]\), Theorem 6.B.4 of Shaked and Shanthikumar (2007) finally gives us
\[
(P_{nm, \Gamma_{nm}}, P_{bc, \Gamma_{bc}}) \Unif(U_1, U_2).
\]

Thus the proof is complete. \(\square\)

5.1. **Combining evidence.** In words, Theorem 5.1 says that the combined
information from the two evidence \(P_{nm, \Gamma_{nm}}\) and \(P_{bc, \Gamma_{bc}}\) carries as much evi-
dence as two independent evidence. This allows us to combine these two
pieces of evidence and provide a total evidence against the hypothesis under
both the comparisons. Karmakar et al. (2017) discusses different methods for combining evidence. Any method of combining p-values that is monotone in both of the p-values can be used, e.g. Fisher’s combination method (Fisher, 1932), the mean of the normal transformation (Liptak, 1958), the truncated product method of combining (Hsu, Small, and Rosenbaum, 2013; Zaykin et al., 2002). Also see Becker (1994).

Fisher’s method computes the joint evidence as the tail probability of $\chi^2_4$ 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 $\bar{\alpha}$, the combined evidence using the truncated product method is given by $F_W(E_{\bar{\alpha}, \Gamma_{nm}, \Gamma_{bc}})$, where

\begin{equation}
E_{\bar{\alpha}, \Gamma_{nm}, \Gamma_{bc}} = \mathbb{1}_{P_{nm, \Gamma_{nm}} \leq \bar{\alpha}} \log(P_{nm, \Gamma_{nm}}) + \mathbb{1}_{P_{bc, \Gamma_{bc}} \leq \bar{\alpha}} \log(P_{bc, \Gamma_{bc}}), \quad \text{and}
\end{equation}

\begin{equation}
F_W(w) = 2\bar{\alpha}(1 - \bar{\alpha})[1 - F_{\text{Exp}}(1)(-\log(\frac{w}{\bar{\alpha}}))] + \bar{\alpha}^2[1 - F_{\text{Gamma}}(2,1)(-\log(\frac{w}{\bar{\alpha}^2}))].
\end{equation}

The advised choice of $\bar{\alpha}$ is 0.20 (Hsu et al., 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 $\beta$. 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 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. In our simulation we sample 2,000 broad cases and largest half of them are labeled as narrow cases. Then we sample 2,000 controls. In both
Table 2
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 \( N(\beta, 1) \) if \( Z_l = 1 \) and \( N(0, 1) \) if \( Z_l = 0 \). There are 1,000 narrow cases and 1,000 marginal cases with 2,000 controls. Based on 10,000 iterations. Fisher = Fisher’s combination method, \( TP = \) truncated product method with \( \alpha = 0.20 \).

<table>
<thead>
<tr>
<th>( \Gamma_{nm} )</th>
<th>( \Gamma_{bc} )</th>
<th>( \beta = 0 )</th>
<th>( \beta = 0.2 )</th>
<th>( \beta = 0.4 )</th>
<th>( \beta = 0.6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Fisher tP</td>
<td>Fisher tP</td>
<td>Fisher tP</td>
<td>Fisher tP</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.0491 0.0463</td>
<td>0.9999 1</td>
<td>1 1 1 1</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0.0061 0.0102</td>
<td>0.2525 0.257</td>
<td>1 1 1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0061 0.0102</td>
<td>0.1758 0.2236</td>
<td>0.8671 0.8564</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>0.0061 0.0102</td>
<td>0.1758 0.2236</td>
<td>0.7478 0.7971</td>
<td>0.9993 0.9991</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>1.25</td>
<td>0 0 0.4807 0.507</td>
<td>1 1 1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0 0 0.0005 0.0011</td>
<td>0.1552 0.1544</td>
<td>0.9999 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.75</td>
<td>0 0 0.0005 0.0011</td>
<td>0.0318 0.0459</td>
<td>0.6865 0.6638</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>0 0 0.0005 0.0011</td>
<td>0.0318 0.0459</td>
<td>0.6865 0.6638</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1.5</td>
<td>0 0 0.0023 0.0033</td>
<td>0.9921 0.9936</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>0 0 0 0</td>
<td>0.0001 0.0001</td>
<td>0.5444 0.5243</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>0 0 0 0</td>
<td>0.0001 0.0001</td>
<td>0.0135 0.0205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.75</td>
<td>1.75</td>
<td>0 0 0 0.5115 0.5784</td>
<td>1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0 0 0 0</td>
<td>0.0217 0.0342</td>
<td>0.9983 0.9989</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.25</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0.0001 0.0002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0 0 0 0</td>
<td>0.0215 0.0342</td>
<td>0.9983 0.9989</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0.3569 0.4188</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0.001 0.0022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.25</td>
<td>2.25</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0.8795 0.9117</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0.3542 0.4187</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0 0 0 0</td>
<td>0 0</td>
<td>0.0009 0.0022</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

comparisons of narrow cases versus marginal cases and broad cases versus controls we consider paired stratum, i.e., \( n_s = m_s = 1, c_s = 2 \).

Table 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 \( \Gamma_{nm} \) and \( \Gamma_{bc} \). 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 \( \alpha = 0.20 \).

6. Evidence factors with differential effect of unmeasured confounders on the factors. The individual factors in an evidence factors
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 1,000 narrow cases and 1,000 marginal cases with 2,000 controls. Based on 10,000 iterations. Fisher = Fisher’s combination method, $tP = \text{truncated product method with } \tilde{\alpha} = 0.20$.

<table>
<thead>
<tr>
<th>$\Gamma_{nm}$</th>
<th>$\Gamma_{bc}$</th>
<th>$\beta = 0$</th>
<th>$\beta = 0.2$</th>
<th>$\beta = 0.4$</th>
<th>$\beta = 0.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Fisher $tP$</td>
<td>Fisher $tP$</td>
<td>Fisher $tP$</td>
<td>Fisher $tP$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.0522 0.0625</td>
<td>0.9994 0.9996</td>
<td>1 1</td>
<td>1 1</td>
</tr>
<tr>
<td>1.5</td>
<td>2</td>
<td>0.0105 0.0152</td>
<td>0.0049 0.0051</td>
<td>0.9995 0.9997</td>
<td>1 1</td>
</tr>
<tr>
<td>2</td>
<td>0.0105 0.0152</td>
<td>0.0004 0.0008</td>
<td>0.1475 0.1774</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>0.0105 0.0152</td>
<td>0.0004 0.0008</td>
<td>0.0002 0.0002</td>
<td>0.9779 0.9845</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>1.25</td>
<td>0 0</td>
<td>0.4687 0.5363</td>
<td>1 1</td>
<td>1 1</td>
</tr>
<tr>
<td>2</td>
<td>0 0</td>
<td>0 0</td>
<td>0.1367 0.1766</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>2.75</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0.7112 0.7684</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0.0021 0.0035</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>1.5</td>
<td>0 0</td>
<td>0.0017 0.003</td>
<td>0.9995 0.9997</td>
<td>1 1</td>
</tr>
<tr>
<td>2.5</td>
<td>0 0</td>
<td>0 0</td>
<td>0.0001 0.0001</td>
<td>0.9768 0.9844</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0.0021 0.0035</td>
<td></td>
</tr>
<tr>
<td>1.75</td>
<td>1.75</td>
<td>0 0</td>
<td>0 0</td>
<td>0.8185 0.8612</td>
<td>1 1</td>
</tr>
<tr>
<td>2</td>
<td>0 0</td>
<td>0 0</td>
<td>0.1364 0.1766</td>
<td>1 1</td>
<td></td>
</tr>
<tr>
<td>3.25</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0.0343 0.0472</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0 0</td>
<td>0 0</td>
<td>0.1364 0.1766</td>
<td>1 1</td>
</tr>
<tr>
<td>2.5</td>
<td>0 0</td>
<td>0 0</td>
<td>0.0001 0.0001</td>
<td>0.9768 0.9844</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0.2408 0.3008</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0.0021 0.0035</td>
<td></td>
</tr>
<tr>
<td>2.25</td>
<td>2.25</td>
<td>0 0</td>
<td>0 0</td>
<td>0.0024 0.0042</td>
<td>0.9998 0.9999</td>
</tr>
<tr>
<td>2.5</td>
<td>0 0</td>
<td>0 0</td>
<td>0.0001 0.0001</td>
<td>0.9768 0.9844</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0.2408 0.3008</td>
<td></td>
</tr>
</tbody>
</table>
confounding, \( u_1 \) and \( u_2 \), and the exposure to treatment is model by bias level \( \lambda \). The relation of \( u_1 \) and the broad case status is modeled by the bias level \( \delta_{bc} \). Finally, the relation of \( u_2 \) and the broad case status is modeled by the bias level \( \delta_{nm} \). In the following we allow for \( u_1 \) and \( u_2 \) to be influenced by overlapping factors.

For individual \( i \), let \( v_{1i} \) and \( v_{2i} \) be unmeasured numbers summarizing two sets of unmeasured variables so that \( 0 \leq v_{1i}, v_{2i} \leq 1 \). We allow for both variables to bias each analysis but to have varying importance in their relationship with the outcomes. We formalize this as follows. Let \( u_{1i} = \psi_1 v_{1i} + \psi_2 v_{2i} \) where \( \psi_1, \psi_2 \geq 0 \), \( \psi_1 + \psi_2 = 1 \) and \( \psi_1 \) is larger than \( \psi_2 \). Also, let \( u_{2i} = \tilde{\psi}_1 v_{1i} + \tilde{\psi}_2 v_{2i} \) 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 \( \psi_1, \tilde{\psi}_1, \psi_2, \tilde{\psi}_2 \) are fixed numbers. The unmeasured confounders \( v_{1i} \) and \( v_{2i} \) relates to the broad case status and the narrow case status by models (4.5) and (4.6) via the variables \( u_{1i} \) and \( u_{2i} \).

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

\[
(6.1) \quad \Pr(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} | C, x_{i1} = x_{i2}, Z_{i1} + Z_{i2} = 1) = \frac{\exp\{\lambda(z_{i1}\omega_{i1} + z_{i2}\omega_{i2})\}}{\exp(\lambda\omega_{i1}) + \exp(\lambda\omega_{i2})},
\]

\[
(6.2) \quad \text{where } \omega_l = \zeta_1 v_{1l} + \zeta_2 v_{2l}, \text{ for } l = 1, \ldots, L; \zeta_1, \zeta_2 \geq 0, \zeta_1 + \zeta_2 = 1.
\]

Now consider the amplification of the sensitivity parameters \( \Gamma_{bc} \) and \( \Gamma_{nm} \) under the model specified by equations (6.1), (6.2) and (4.5), (4.6) with \( u_{1i} = \psi_1 v_{1i} + \psi_2 v_{2i} \) and \( u_{2i} = \tilde{\psi}_1 v_{1i} + \tilde{\psi}_2 v_{2i} \). 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 \( \lambda \) in model (6.1)–(6.2). We find the parameters \( \delta_{bc} \) and \( \delta_{nm} \) from \( \lambda \) and \( \Gamma_{bc}, \Gamma_{nm} \). 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, i.e. we put the restriction \( v_{2i} = v_{2i}' \), then \( \Delta_{bc} = \{(\Lambda \Gamma_{bc} - 1) / (\Lambda - \Gamma_{bc})\}^{1/\psi_1} \) and \( \Delta_{nm} = \{(\Lambda \Gamma_{nm} - 1) / (\Lambda - \Gamma_{nm})\}^{1/\psi_2} \).

This correspondence holds with \( |v_{1,i1} - v_{1,i2}| = 1 \). (ii) If only \( v_2 \) is the bias in question, i.e. we put the restriction \( v_{1i} = v_{1i}' \), then \( \Delta_{bc} = \{(\Lambda \Gamma_{bc} - 1) / (\Lambda - \Gamma_{bc})\}^{1/\psi_2} \) and \( \Delta_{nm} = \{(\Lambda \Gamma_{nm} - 1) / (\Lambda - \Gamma_{nm})\}^{1/\psi_1} \) 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 \( \delta_{nm} = \log(\Delta_{nm}) \) changes 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, \( \delta_{bc} \), and on the narrow case status to a marginal case status, \( \delta_{nm} \). In this illustration we assume \( \psi_1 = \frac{3}{4} \), 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, i.e., \( \tilde{\psi}_1 = \frac{1}{5} \). 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 amount of $\delta_{bc} \geq 1.671$ and $\delta_{nm} \geq 6.265$. The second critic would have to assume biases at the amount 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 and that between all colorectal cancer cases and controls is 0.64. 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 versus the proximal colorectal cancer cases is $2.3 \times 10^{-5}$, with the corresponding odds ratio 0.60 (C.I. 0.46 to 0.76)). The p-value for all cases (distal and proximal) versus the matched controls is $5.0 \times 10^{-11}$, with odds ratio 0.62 (C.I. 0.54 to 0.72)) (this result is similar to previously reported odds ratios (Atkin et al., 2010; 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

<table>
<thead>
<tr>
<th></th>
<th>Distal cancer cases</th>
<th>Proximal cancer cases</th>
<th>All colorectal cancer cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>No screening sigmoidoscopy</td>
<td>678</td>
<td>662</td>
<td>1340</td>
<td>2538</td>
</tr>
<tr>
<td>Screening sigmoidoscopy</td>
<td>144</td>
<td>224</td>
<td>368</td>
<td>1097</td>
</tr>
<tr>
<td>Odds ratio from matched sets</td>
<td>0.60 (0.46, 0.76)</td>
<td></td>
<td>0.62 (0.54, 0.72)</td>
<td></td>
</tr>
<tr>
<td>p-value from matched sets</td>
<td>$2.3 \times 10^{-5}$</td>
<td></td>
<td>$5.0 \times 10^{-11}$</td>
<td></td>
</tr>
</tbody>
</table>
of Section 4 we consider two sensitivity parameters $\Gamma_{nm}$ and $\Gamma_{bc}$ 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 $\bar{\alpha} = 0.20$. 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 com-
pared to evidence from an analysis that only looks at the screening rates between all colorectal cancer cases and controls.

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, we pair matched narrow cases to marginal cases 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, e.g., full matching (Hansen, 2004) or 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, i.e., 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 non-malarial causes of coma and many children in malaria
endemic areas have non-symptomatic 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.

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.

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 $Z_{n(s)}$. Given the strata, from equation (4.1) we have that the maximum p-value of the narrow versus marginal comparison, $P_{nm,\Gamma_{nm}}$, is computed based on the conditional distributions $\{\{Z_{n(s)} | \sum_{i \in [ns]} Z_{n(s)} + \sum_{i \in [ms]} Z_{m(s)}\}\}$. Combining these facts, we get the first result that marginally $P_{nm,\Gamma_{nm}}$ is a function of $\sum_{i \in [ns]} Z_{n(s)} + \sum_{i \in [ms]} Z_{m(s)}$.

Next we note that $T_{bc}$ is a function of $\sum_{i \in [ns]} Z_{n(s)} + \sum_{i \in [ms]} Z_{m(s)}$. Now, by looking at equation (4.2), $P_{bc,\Gamma_{bc}}$ is computed based on the family of conditional distributions $\{\{\sum_{i \in [ns]} Z_{n(s)} + \sum_{i \in [ms]} Z_{m(s)} | \sum_{i \in [ns]} Z_{n(s)} + \sum_{i \in [ms]} Z_{m(s)} + \sum_{i \in [cs]} Z_{c(s)}\}\}$. Consequently, $P_{bc,\Gamma_{bc}}$ is determined by the number of exposed cases $\{\sum_{i \in [ns]} Z_{n(s)} + \sum_{i \in [ms]} Z_{m(s)}\}$ and the total number of exposed individuals $\{\sum_{i \in [ns]} Z_{n(s)} + \sum_{i \in [ms]} Z_{m(s)} + \sum_{i \in [cs]} Z_{c(s)}\}$. But it is enough to know whether each control is exposed or not, i.e., $Z_{c(s)}$, to know the number of exposed cases when we have the information on total number of exposed units. The result is hence proved. ■
Proof of Lemma 5.3. For part (i) and (ii) note that p-values or their upper bounds are valid p-values thus are stochastically larger than $Unif[0,1]$. Part (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 $\sum_{i \in [n]} Z_{n(i)} + \sum_{i \in [m]} Z_{m(i)}$ the random variables $Z_{n(i)}$ and $Z_{c(i)}$ are independently distributed, by Lemma 5.2 the conditional distribution in the statement of the lemma is same as $[P_{nm} \Gamma_{nm} \mid \sum_{i \in [n]} Z_{n(i)} + \sum_{i \in [m]} Z_{m(i)}]$. Now the result follows from part (i) of Lemma 5.3.

Proof of Lemma 5.5. We can write for any $0 \leq p, q, \leq 1$, the conditional probability as,

$$\Pr(P_2 \leq q \, | \, P_1 \leq p) \overset{C_1}{=} \Pr(P_2 \leq q \, | \, \{V_1 : P_1 \leq p\})$$

$$= \mathbb{E}[\Pr(P_2 \leq q \, | \, V_1) \, | \, \{V_1 : P_1 \leq p\}]$$

$$\overset{C_2}{\leq} \mathbb{E}[q \, | \, \{V_1 : P_1 \leq p\}] = q.$$

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

REFERENCES


B. KARMAKAR, C. A. DOUBENI AND D. S. SMALL


**DEPARTMENT OF STATISTICS**
**The Wharton School**
**University of Pennsylvania**
**400 Jon M. Huntsman Hall**
**3730 Walnut Street**
**Philadelphia, PA 19104-6340**
**E-mail: bikramk@wharton.upenn.edu**
**dsmall@wharton.upenn.edu**

**DEPARTMENT OF FAMILY MEDICINE**
**AND COMMUNITY HEALTH**
**Perelman School of Medicine**
**University of Pennsylvania**
**Gate 2, 3400 Spruce Street**
**Philadelphia, PA 19104, U.S.A.**
**E-mail: chyke.doubeni@uphs.upenn.edu**