Biometrika (2018), **xx**, to appear, *pp*. 1–15
Printed in Great Britain

Integrating the evidence from evidence factors in observational studies

BY B. KARMAKAR

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, U.S.A.
bikramk@wharton.upenn.edu

B. FRENCH

Department of Statistics, Radiation Effects Research Foundation, Hiroshima 732-0815, Japan
french@rerf.or.jp

AND D. S. SMALL

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, U.S.A.
dsmall@wharton.upenn.edu

SUMMARY

A sensitivity analysis for an observational study assesses how much bias, due to non-random assignment of treatment, would be necessary to change the conclusions of an analysis that assumes treatment assignment was effectively random. The evidence for a treatment effect can be strengthened if two different analyses, which could be affected by different types of biases, are both somewhat insensitive to bias. The finding from the observational study is then said to be replicated. Evidence factors allow for two independent analyses to be constructed from the same data set. When combining the evidence factors, the type-I error rate must be controlled to obtain valid inference. A powerful method is developed for controlling the familywise error rate for sensitivity analyses with evidence factors. It is shown that the Bahadur efficiency of sensitivity analysis for the combined evidence is greater than for either evidence factor alone. The proposed methods are illustrated through a study of the effect of radiation exposure on risk of cancer. An R package ‘evidenceFactors’ is available from CRAN to implement the methods of the paper.

Some key words: Design sensitivity; Evidence factors; Observational study; Quasi-experiment; Sensitivity analysis.

1. INTRODUCTION

In an observational study, treatment assignment is typically assumed to be effectively random conditional on measured covariates. However, the presence of unmeasured confounding can result in non-random treatment assignment, such that standard analysis methods can provide biased estimates of treatment effects. The potential for measured and unmeasured confounding motivates consideration of sensitivity analyses to assesses how much bias, due to non-random treatment assignment, would be necessary to change the conclusions of a randomization infer-

© 2017 Biometrika Trust

ence (Cornfield et al., 1959; Rosenbaum, 1987; Keele & Minozzi, 2013; Stuart et al., 2013; Ding & VanderWeele, 2016; McCandless & Gustafson, 2017).

Evidence factors – two or more independent tests that could be sensitive to different biases – provide an approach to strengthen the evidence for a treatment effect (Rosenbaum, 2010, 2011). When considering sensitivity analyses with evidence factors, multiple comparisons arise by performing more than one test of the same null hypothesis and by considering different sensitivity parameters. Therefore, multiplicity error must be controlled to obtain valid inference. Previous research has not considered the impact of multiplicity error when generating inference based on evidence factors. Standard methods for multiplicity adjustment, such as a Bonferroni correction, could impose a harsh penalty when there are multiple sources of evidence. Consideration of evidence factors is meant to strengthen the aggregate evidence for a treatment effect, but a punitive penalty for multiple comparisons can hamper the ability to detect a significant treatment effect. Can the attractive benefits of evidence factors analysis always be obtained? This paper provides an affirmative answer. We provide a powerful and computationally fast method for combining evidence factors that controls for multiplicity. We show that, in terms of the Bahadur efficiency of the sensitivity analysis, our approach for combining evidence from multiple sources has better performance than considering any of the sources separately.

2. EXAMPLE: SOLID CANCER INCIDENCE IN ATOMIC BOMB SURVIVORS

Understanding the health effects of radiation exposure is important for establishing recommendations for radiation protection, including limits on occupational exposure to radiation and guidelines for diagnostic and therapeutic use of radiation. Because randomized experiments on humans are unethical, observational studies are a key resource for estimating radiation effects.

In 1945, the United States detonated two atomic bombs over the Japanese cities of Hiroshima and Nagasaki. The Life Span Study investigates the long-term health effects of radiation exposure among survivors of the atomic bombings. The Life Span Study includes: proximal survivors, who were within 3 km of the hypocenter; distal survivors, who were between 3 and 10 km of the hypocenter; and city residents who were not in either city at the time of the bombings, and therefore not exposed to radiation. A survivor's radiation dose is estimated from a dosimetry system that accounts for the survivor's reported location and shielding at the time of the bombing, with the total dose given by the sum of the γ-ray dose and 10 times the neutron dose in units of gray (Gy) (Cullings et al., 2017).

Following Preston et al. (2007), our goal is to evaluate the hypothesis that radiation increases the risk of solid cancer. At the time of the bombings, there were notable differences between Hiroshima and Nagasaki. Hiroshima, with even terrain, was an embarkation port and a site of major military headquarters, whereas Nagasaki, with varied terrain, was a center of heavy industry, with associated air pollution. To minimize heterogeneity, our analysis was limited to those Life Span Study participants from Hiroshima alive and at risk for solid cancer as of January 1, 1958, when population-based cancer registries were established. In addition, we did not consider distal survivors because of concern that distal survivors, who lived in more rural areas, and proximal survivors, who lived in more urban areas, could have different cancer rates for reasons other than radiation dose (Pierce & Preston, 2000).

To assess the effect of radiation exposure on the risk of solid cancer, one could compare proximal survivors with high doses to proximal survivors with low doses; the validity of this comparison relies on the assumption that proximal survivors with low and high doses are similar on all characteristics other than their radiation exposure. This assumption could be violated because the hypocenter was close to the urban center, so that the proximal survivors with high

Table 1. *Solid cancer incidences among atomic bomb survivors in Hiroshima*

distance from hypocenter	radiation dose	number of participants	number of solid cancers	percentage
$< 3000\ m$	Low ($\leq$ 2 Gy)	38932	6,989	18·0
$< 3000\ m$	High ($>$ 2 Gy)	397	140	35·3
$< 3000\ m$	all	39329	7129	18·1
not-in-city	–	19259	3160	16·4

Table 2. *Sensitivity analysis for the hypothesis of no carcinogenic effect of radiation versus harmful effect of radiation*

Γ	Maximum p-value	Γ_1	Γ_2	joint evidence
	High-dose vs. Low-dose survivors	1	1	$1{\cdot}47\times10^{-11}$
1	0·0021	1	1·1	0·00032
1·2	0·0443	1	1·2	0·01420
1·3	0·1166	1·2	1	$2{\cdot}73\times10^{-10}$
		1·2	1·1	0·00491
	Proximal survivors vs. NIC residents	1·2	1·2	0·17117
1	$2{\cdot}35\times10^{-10}$	1·3	1	$6{\cdot}94\times10^{-10}$
1·1	0·0131	1·3	1·1	0·01145
1·2	0·9207	1·3	1·2	0·34688

NIC, not-in-city.

doses tended to be located in more urban areas; also, high-dose survivors might have been comparatively healthier to have survived a high dose (Preston et al., 2007). Alternatively, one could compare cancer rates between proximal survivors and not-in-city residents; the validity of this comparison relies on the assumption that proximal survivors and not-in-city residents are similar on all characteristics other than their radiation exposure. This assumption could be violated, for example, if not-in-city residents were better educated or employed. Although both of these comparisons use the proximal survivors, we will show that under the null hypothesis of no effect, they are nearly independent in the sense that their p-values are stochastically as large as the p-values from two independent comparisons under the null hypothesis, which are uniform on the unit square. This will be discussed more formally in §3·1 and additional details provided in §2.2 and §2.3 of the supplement.

Table 1 provides a summary of these two comparisons. The incidence rate of solid cancer was $18{\cdot}0\%$ among proximal survivors exposed to low doses versus $35{\cdot}3\%$ among those exposed to high doses, amounting to an $18{\cdot}1\%$ incidence rate among all proximal survivors. Among not-in-city residents, the incidence rate was $16{\cdot}4\%$. After matching on age and sex, 58 strata were created for the low-dose versus high-dose comparison among proximal survivors and 30 strata were created for the comparison of all proximal survivors versus not-in-city residents. Both the comparisons give strong evidence suggesting radiation exposure is harmful, with one-sided p-values from Mantel–Haenszel tests 0·0021 and $2{\cdot}35\times10^{-10}$ respectively.

What is the gain from considering two p-values from two analyses? Each p-value is computed based on an assumption of no unmeasured confounders for the given comparison. This assumption could be violated for one comparison but not the other. For example, there might be unmeasured differences between people who lived near the hypocenter of the bomb, high-dose proximal survivors, versus far, low-dose proximal survivors, but not between people who were in or out of the city at the time of the bombing, or vice versa. If both p-values indicate strong evidence against the null hypothesis, then there would have to be unmeasured confounders for both comparisons in order to bring the results into question. For each of the two comparisons, we associate a single sensitivity parameter measuring the bias due to the presence of unmeasured confounders. One can study the effect of potential bias by evaluating the strength of evidence for a treatment effect from the comparison for different values of the sensitivity parameter. This sensitivity parameter is defined as the maximum odds that, among two participants with the same measured confounders, one participant would receive treatment and the other control compared to vice versa because of differences in unmeasured confounders (e.g. Rosenbaum, 1987). In §3·2 a formal definition of this sensitivity parameter is given. When the value of this sensitivity parameter is 1, it indicates the assumption of no unmeasured confounders; a value of 2 would mean that the unmeasured confounders can double the odds of receiving treatment. Let Γ_1 and Γ_2 denote these sensitivity parameters for the two comparisons. Table 2 reports the maximum p-values for the two comparisons for different values of these parameters. When Γ_1 =1·2 and Γ_2 =1·1, both the comparisons reject the null hypothesis with maximum p-values 0·0443 and 0·0131 respectively. The evidences from the two comparisons are sensitive at bias levels Γ_1 =1·3 and Γ_2 =1·2, respectively. Table 2 also reports the joint evidence given (Γ_1, Γ_2) values. The joint evidence is calculated using Fisher's combination method. If $(\Gamma_1, \Gamma_2) = (1{\cdot}2, 1{\cdot}2)$ we fail to reject the null hypothesis. The existing theory of evidence factors only allows us to make these statements about a given pair (Γ_1, Γ_2). An objective statistician would not choose a value of (Γ_1, Γ_2), but rather present the results for a range of values, in particular focusing on the value where the inference is sensitive. Hence, we would like to make a comprehensive statement for a range of values of (Γ_1, Γ_2) while ensuring that the familywise error rate is controlled.

3. EVIDENCE FACTORS: A GENERAL VIEWPOINT

3·1. *Definition of evidence factors*

Suppose we wish to test a hypothesis H_0 and let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two different assumptions under which the hypothesis can be tested. Let the evidence gathered against H_0 based on $\mathcal{A}_1$ be E_1 and the evidence based on $\mathcal{A}_2$ be E_2 after taking out E_1. If E_1 and E_2 are p-values calculated from the data given the assumptions $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively, then E_1 and E_2 would constitute separate evidence factors if they are independent upon the assumption of $\mathcal{A}_1 \cap \mathcal{A}_2$. Henceforth in our discussion by evidence against null we mean the (maximum) p-value. The requirement of independence can be relaxed because the desired property of (E_1, E_2) is: when considered jointly they provide more evidence against H_0 than separately. The pair (E_1, E_2) are called evidence factors if, when both $\mathcal{A}_1$, $\mathcal{A}_2$ and H_0 hold, the joint cumulative distribution function of (E_1, E_2) is stochastically larger than the joint distribution of two independent p-values. As shown in §3·2, this definition implies that, most tests for the null hypothesis using the evidence factors can use the cutoff calculated assuming independence and be valid. Since p-values are uniformly distributed under the null hypothesis, this amounts to having for all $(p_1, p_2) \in [0, 1]^2$

$$\mathrm{pr}(E_1 \leq p_1, E_2 \leq p_2) \leq p_1 \times p_2. \qquad (1)$$

DEFINITION 1. *A set $D \subseteq \mathbb{R}^k$ is called a decreasing set if for any $x, y \in \mathbb{R}^k$ 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, in notation $X \succ Y$, if* $\mathrm{pr}(X \in D) \leq \mathrm{pr}(Y \in D)$ *for all decreasing sets D.*

DEFINITION 2. *The pair (E_1, E_2) is said to form evidence factors for testing H_0 assuming $\mathcal{A}_1$ and $\mathcal{A}_2$ if, $(E_1, E_2) \succ (U_1, U_2)$ under $\mathcal{A}_1 \cap \mathcal{A}_2$ and H_0, for two independent Unif[0,1] random variables U_1 and U_2.*

Since $[0, p_1] \times [0, p_2]$ are decreasing sets, if (E_1, E_2) are evidence factors then (1) is satisfied.

3·2. *Sensitivity analysis and evidence factors*

Consider the sensitivity of the evidences E_1 and E_2 with respect to their corresponding assumptions $\mathcal{A}_1$ and $\mathcal{A}_2$. Let $\Gamma_1 (\geq 1)$ be a real number that quantifies possible deviation from assumption $\mathcal{A}_1$ (Gastwirth, 1992; Hosman et al., 2010; Zubizarreta et al., 2012; Rosenbaum, 2002, §4). For instance, when the assumption $\mathcal{A}_1$ is that the treatment is randomly assigned among the treated and the control units, i.e. there are no unmeasured confounders, Γ_1 would quantify bias in treatment assignment due to possible unmeasured confounders. To make this precise, we discuss one definition of the parameter Γ_1 (Rosenbaum, 2002, §4) here. Let $i = 1, \ldots, n$ be the indices assigned arbitrarily to n units. Let Z_i be the indicator for unit i being in the treatment group. Also, let x_i denote the observed pretreatment covariates while u_i is an unobserved number summarizing the unobserved confounders for unit i (see Rosenbaum, 1987). Finally, suppose unit i if exposed to treatment would have response r_{Ti} and if spared exposure would have response r_{Ci}. Consequently, both r_{Ti} and r_{Ci} are not observed simultaneously (Neyman, 1923). Let $\mathcal{F} = \{(r_{Ti}, r_{Ci}, x_i, u_i) \mid i = 1, \ldots, n\}$. Then the sensitivity parameter Γ_1 will be defined by, $\Gamma_1 = \max_{1 \leq i, i' \leq n; x_i = x_{i'}} \text{pr}(Z_i = 1 \mid \mathcal{F})\text{pr}(Z_{i'} = 0 \mid \mathcal{F})\{\text{pr}(Z_{i'} = 1 \mid \mathcal{F})\text{pr}(Z_i = 0 \mid \mathcal{F})\}^{-1}$. In words, the model for treatment assignment is such that due to unobserved covariates the odds of being treated for two units i and i' with the same observed covariates is allowed to differ at most by a multiplicative factor $\Gamma_1\ (\geq 1)$. When there are no unmeasured confounders, $\Gamma_1 = 1$ and two units similar in terms of their observed covariates have the same probabilities of receiving treatment. Let $\mathcal{A}_1(\Gamma_1)$ denote all treatment assignment distributions that deviate from this randomized assignment, assumption $\mathcal{A}_1$, by bias level at most Γ_1.

In a more general setup, when testing H_0 based on a test statistic T_1, the set $\mathcal{A}_1(\Gamma_1)$ would specify a family of possible distributions $\mathcal{P}_1(\Gamma_1)$ for T_1, and the larger Γ_1 is, the larger the family of distributions $\mathcal{P}_1(\Gamma_1)$ becomes. Define sensitivity parameter $\Gamma_2\ (\geq 1)$ and corresponding $\mathcal{A}_2(\Gamma_2)$ similarly for the second factor. Thus, the larger Γ_j is, the more uncertain we are about the design and $\Gamma_j = 1$ implies that we are certain about the aspect $\mathcal{A}_j$ of the design, i.e. $\mathcal{A}_j(1) = \mathcal{A}_j$ for $j = 1, 2$.

The sensitivity analysis computes the largest possible p-values under $\mathcal{A}_1(\Gamma_1)$ and $\mathcal{A}_2(\Gamma_2)$ as $E_1\{\mathcal{A}_1(\Gamma_1)\}$ and $E_2\{\mathcal{A}_2(\Gamma_2)\}$. Naturally, a larger uncertainty about the design will lead to weaker evidence. Because these assumptions are nested; for $j = 1, 2$, with $\Gamma_j \leq \Gamma_j'$

$$\mathcal{A}_j(\Gamma_j) \subseteq \mathcal{A}_j(\Gamma_j'), \qquad E_j\{\mathcal{A}_j(\Gamma_j)\} \leq E_j\{\mathcal{A}_j(\Gamma_j')\}. \tag{2}$$

Following §3·1, $\{E_1\{\mathcal{A}_1(\Gamma_1)\} \mid \Gamma_1 \geq 1\}$ and $\{E_2\{\mathcal{A}_2(\Gamma_2)\} \mid \Gamma_2 \geq 1\}$ are said to form evidence factors for testing H_0 if for any Γ_1 and Γ_2 the pair $(E_1\{\mathcal{A}_1(\Gamma_1)\}, E_2\{\mathcal{A}_2(\Gamma_2)\})$ constitute evidence factors under the assumptions $\mathcal{A}_1(\Gamma_1)$ and $\mathcal{A}_2(\Gamma_2)$. We use the shorthand notation E_{1,Γ_1} and E_{2,Γ_2} for $E_1\{\mathcal{A}_1(\Gamma_1)\}$ and $E_2\{\mathcal{A}_2(\Gamma_2)\}$, respectively, because there is no ambiguity.

4. Combining evidence

How should we quantify combined evidence against H_0 from the evidence factors? Fisher's method, which is used in §2, is a natural choice.

LEMMA 1. *Under H_0, the distribution of $-2\log\ E_1E_2$ is stochastically smaller than the χ^2-distribution with 4 degrees of freedom.*

Proof. Let U_1, U_2 be two independent Unif[0, 1] random variables. For $0 \leq p, q \leq 1$, $pq = \text{pr}(U_1 \leq p, U_2 \leq q)$. Further, $-2\log\ U_1U_2$ is distributed as χ^2 with 4 degrees of freedom. As $(p, q) \mapsto -2\log\ pq$ is a monotone function in both coordinates by Theorem 6.B.16 of Shaked & Shanthikumar (2007, §6), $-2\log\ E_1E_2$ is stochastically smaller than χ^2_4 distribution. □

Since, by definition, $(E_{1,\Gamma_1}, E_{2,\Gamma_2})$ form evidence factors, the combined evidence for bias levels Γ_1 and Γ_2, calculated using Fisher's method is $E_{\Gamma_1,\Gamma_2} = \mathrm{pr}(\chi^2_4 > -2\log(E_{1,\Gamma_1}E_{2,\Gamma_2}))$.

An alternative method of combining p-values is Zaykin et al. (2002)'s truncated product method, which puts more emphasis than Fisher's method on looking for small p-values. The truncated product for some $\tilde{\alpha} \in (0, 1)$ is defined as

$$E_{\wedge,\Gamma_1,\Gamma_2}{=}1_{E_{1,\Gamma_1}\leq\tilde{\alpha}} \log\, E_{1,\Gamma_1} + 1_{E_{2,\Gamma_2}\leq\tilde{\alpha}} \log\, E_{2,\Gamma_2};$$

with 1_A denoting the indicator of an event A. An evidence factor contributes to $E_{\wedge,\Gamma_1,\Gamma_2}$ only if the evidence from that factor is strong, i.e., less than $\tilde{\alpha}$. Fisher's method corresponds to $\tilde{\alpha} = 1$. Hsu et al. (2013) presented simulations and discussion that suggested that the truncated product method often performs better than Fisher's method in sensitivity analysis. The following lemma studies the null distribution of $E_{\wedge,\Gamma_1,\Gamma_2}$.

LEMMA 2. *Let W be a random variable on $[0, \tilde{\alpha}^2]$ with the distribution function*

$$F_W(w) = 2\tilde{\alpha}(1-\tilde{\alpha})(1 - F_{Exp(1)}[-\log\{w(\tilde{\alpha})^{-1}\}]) + \tilde{\alpha}^2(1 - F_{Gamma(2,1)}[-\log\{w(\tilde{\alpha})^{-2}\}]).$$

Then under H_0 and $\mathcal{A}_1(\Gamma_1) \cap \mathcal{A}_2(\Gamma_2)$, $\exp(E_{\wedge,\Gamma_1,\Gamma_2})$ is stochastically larger than W.

Proof. Hsu et al. (2013) provided a simple argument to prove the lemma in the case where E_{1,Γ_1} and E_{2,Γ_2} were independent. Define $f_\wedge(x, y) = \exp\{1_{x\leq\tilde{\alpha}} \log\, x + 1_{y\leq\tilde{\alpha}} \log\, y\}$. Then, $f_\wedge$ is a monotone nondecreasing function. Because the pair $(E_{1,\Gamma_1}, E_{2,\Gamma_2})$ form evidence factors, by Theorem 6.B.16 of Shaked & Shanthikumar (2007, §6), $\exp(E_{\wedge,\Gamma_1,\Gamma_2}) \succ f_\wedge(U_1, U_2)$. Using this fact and the argument of Hsu et al. the proof of the lemma follows. □

Based on the truncated product $E_{\wedge,\Gamma_1,\Gamma_2}$, H_0 is rejected at level of significance α if $\exp(E_{\wedge,\Gamma_1,\Gamma_2})$ is smaller than the αth quantile of the distribution F_W. The combined evidence E_{Γ_1,Γ_2} is quantified as $F_W(\exp\, E_{\wedge,\Gamma_1,\Gamma_2})$. The choice of $\tilde{\alpha}$ is more subjective. Choosing $\tilde{\alpha}$ =0·10 and 0·20 has been advised (Hsu et al., 2013; Zaykin et al., 2002).

Other methods of combining p-values, where the combination is increasing in E_{j,Γ_j}, can be used, e.g. the mean of normal transformations of the evidences defined as $\Phi\{w^{1/2}\Phi^{-1}(E_{1,\Gamma_1}) + (1-w)^{1/2}\Phi^{-1}(E_{2,\Gamma_2})\}$ (Liptak, 1958). Which method is best for combining p-values remains unsettled. Littell & Folks (1971) show that asymptotically, in terms of Bahadur efficiency, Fisher's combination method is optimal. Won et al. (2009) and Whitlock (2005) both show that with appropriate choice of weights, Liptak's method has more power than Fisher's method. Becker (1994) provides a comprehensive survey of various methods for combining p-values.

5. INTEGRATING EVIDENCE

5·1. *Sensitivity analysis and familywise error rate control*

A sensitivity analysis for increasing, potentially infinite, sequences of $\{\Gamma_{1i} \mid i = 1, \ldots\}$ and $\{\Gamma_{2i} \mid i = 1, \ldots\}$ values involves tests of multiple hypotheses. For a pair $(\Gamma_{1i}, \Gamma_{2i'})$ the hypothesis being tested is

$$H_{0,\Gamma_{1i},\Gamma_{2i'}} : H_0 \cap \mathcal{A}_1(\Gamma_{1i}) \cap \mathcal{A}_2(\Gamma_{2i'}).$$

In words, $H_{0,\Gamma_{1i}\Gamma_{2i'}}$ is the hypothesis that H_0 is true and the deviation from assumption $\mathcal{A}_1$ is at most Γ_{1i} and from $\mathcal{A}_2$ is at most $\Gamma_{2i'}$. Since multiple hypotheses are tested simultaneously, controlling type-I error is a concern. Fortunately, as shown below, the structure of the problem allows us to perform each test at level α while controlling for total error at α.

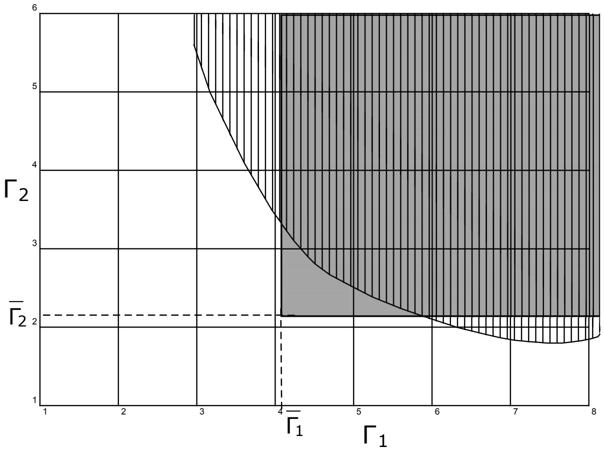

Fig. 1. Illustration of the collection of the hypotheses in a finite sample. Gray area is the null hypotheses and the hatched area depicts the hypotheses not rejected based on the data.

Let $\bar\Gamma_j = \min\{\Gamma_{ji} \mid i = 1, \ldots ; \mathcal{A}_j(\Gamma_{ji}) \text{ is true}\}$ for $j = 1, 2$ with the convention that the minimum of an empty sequence is infinity. When H_0 is true, $H_{0,\Gamma_{1i}\Gamma_{2i'}}$ is true if and only if $\Gamma_{1i} \geq \bar\Gamma_1$ and $\Gamma_{2i'} \geq \bar\Gamma_2$. In Fig. 1 the shaded gray area denotes the set of true null $\mathcal{H}_0 = \{H_{0,\Gamma_{1i},\Gamma_{2i'}} \mid \Gamma_{1i} \geq \bar\Gamma_1, \Gamma_{2i'} \geq \bar\Gamma_2\}$. Let $E_{\Gamma_{1i},\Gamma_{2i'}}$ denote the combined evidence against H_0 under $\mathcal{A}_1(\Gamma_{1i}) \cap \mathcal{A}_2(\Gamma_{2i'})$. In sensitivity analysis on a single parameter, take Γ_1 for example, under H_0 a false rejection at level $\Gamma_{1i} \geq \bar\Gamma_1$ implies false rejection at $\bar\Gamma_1$ which is controlled at level α. Thus the familywise error rate for sensitivity analysis on a single parameter is controlled at the desired level. The following theorem shows that the same argument generalizes for more than one parameter when the parameters correspond to different evidence factors.

THEOREM 1. *Suppose $E_{\Gamma_{1i},\Gamma_{2i'}}$ is a nondecreasing function of individual evidences. Consider the testing procedure where $H_{0,\Gamma_{1i},\Gamma_{2i'}}$ is rejected if and only if $E_{\Gamma_{1i},\Gamma_{2i'}} < \alpha$. Then the probability of rejecting any $\mathcal{H}_0$ is at most α.*

Proof. If H_0 is false there is nothing to prove. Recall property (2) of individual evidences. Now, the joint evidence is nondecreasing in individual evidences. As a consequence of these facts the retained set of hypothesis, i.e. the set of $H_{0,\Gamma_{1i},\Gamma_{2i'}}$ with $E_{\Gamma_{1i},\Gamma_{2i'}} \geq \alpha$, must be an increasing convex set as depicted in form of gridded area in Fig. 1. Thus, for the proposed testing procedure under H_0, pr(any $\mathcal{H}_0$ is rejected) $=$ pr($H_{0,\Gamma_{1i},\Gamma_{2i'}}$is rejected for some $\Gamma_{1i} \geq \bar\Gamma_1$and $\Gamma_{2i'} \geq \bar\Gamma_2$) $\leq$ pr($H_{0,\bar\Gamma_1,\bar\Gamma_2}$ is rejected) $=$ pr($E_{\bar\Gamma_1,\bar\Gamma_2} < \alpha$) $\leq \alpha$. The first inequality follows from convexity of the retention set; the second inequality uses the fact that $E_{\bar\Gamma_1,\bar\Gamma_2} \succ$ Unif[0,1]. □

Methods of combining evidence described in §4 (Fisher's method, the truncated product method) all satisfy the condition of Theorem 1 that $E_{\Gamma_{1i},\Gamma_{2i'}}$ is a nondecreasing function of the individual. However other methods, such as a modified Liptak's method where w is a function of $|\phi^{-1}(E_{j,\Gamma_j})|$ as in Chen & Nadarajah (2014) do not.

The retention set of biases, $\{(\Gamma_{1i}, \Gamma_{2i'}) \mid E_{\Gamma_{1i},\Gamma_{2i'}} \geq \alpha\}$, has a nice structure – it is a convex and increasing set. As a consequence, this set can be computed in $O(\log \max_{j=1,2} |\mathcal{G}_j|)$ time, where $|\mathcal{G}_j|$ is the range for bias on jth evidence factor. This benefit is substantial when there are d evidence factors each with a finite possible bias range $|\mathcal{G}_j|$. Then the complexity, $O(d \log \max_{j=1}^{d} |\mathcal{G}_j|)$, is linear in d as compared to $O(\prod_{j=1}^{d} |\mathcal{G}_j|)$ for linear search algorithms.

The supplement of this paper includes pseudo code for such an algorithm. We have written an R package `evidenceFactors`, available on CRAN, that implements this algorithm along with other methods of this paper.

The above result is more general than stated. One can restrict attention to special subsets of the grid and still ensure multiplicity control. For example Pimentel et al. (2015) discusses testing for pairs $\{(\Gamma_{1i}, 0{\cdot}80\Gamma_{1i}), i = 1, \ldots\}$. Proof of the following corollary is given in the supplement.

COROLLARY 1. *Let $E_{\Gamma_{1i},\Gamma_{2i'}}$ be as in Theorem* 1. *Let $\mathcal{G}$ be a fixed continuous subset of $\{\Gamma_{1i} \mid i \geq 1\} \times \{\Gamma_{2i} \mid i \geq 1\}$ such that $\mathcal{G} \cap \{(1, \Gamma_{2i}) \mid i \geq 1\} \cup \{(\Gamma_{1i}, 1) \mid i \geq 1\}$ is non-empty. Then, the probability that the testing procedure of Theorem* 1 *on $\mathcal{G}$ falsely rejects any hypothesis is at most α.*

5·2. *Design sensitivity, consistency and asymptotic rate*

Most problems of testing have a design sensitivity attached to them, which is an asymptotic measure of power of sensitivity analysis that is not dependent on α (see Rosenbaum, 2004). The design sensitivity is the level of bias above which the power goes to zero as the sample size goes to infinity for any significance level, and below which the power goes to one. The design sensitivity, denoted by $\tilde{\Gamma}$, is the value of the sensitivity parameter at which the corresponding test can asymptotically distinguish a treatment effect with no bias from no treatment effect with bias less than $\tilde{\Gamma}$, but not from no treatment effect with bias larger than $\tilde{\Gamma}$. Let $\tilde{\Gamma}_1$ and $\tilde{\Gamma}_2$ denote the design sensitivity of the first and second kind of biases respectively. Then by definition, for $j = 1, 2$; $E_{j,\Gamma_j} \to 1$ for $\Gamma_j > \tilde{\Gamma}_j$, and $E_{j,\Gamma_j} \to 0$ for $\Gamma_j < \tilde{\Gamma}_j$. All convergence statements here and later are in almost sure sense as the sample size goes to infinity. To explicitly show the dependence on sample size n we write $E_{j,\Gamma_j|n}$ and $E_{\Gamma_1,\Gamma_2|n}$. A consequence of the above is that the joint evidence satisfies: $E_{\Gamma_1,\Gamma_2|n} \to 0$ if $\Gamma_1 < \tilde{\Gamma}_1$ or $\Gamma_2 < \tilde{\Gamma}_2$; $E_{\Gamma_1,\Gamma_2|n} \to 1$ if $\Gamma_j > \tilde{\Gamma}_j$ for both $j = 1, 2$. Pictorially, this means that the gridded area in Fig. 1, which is the collection of hypotheses not rejected, in the limiting case, with the sample size going to infinity, will coincide with the gray rectangular area depicting the collection of true null hypotheses ($\mathcal{H}_0$). Hence, the design sensitivity of the joint conclusion is $(\tilde{\Gamma}_1, \tilde{\Gamma}_2)$.

However, these limits do not provide any information on the rates at which such convergences take place. One can consider the Bahadur slope (Bahadur, 1967), which is the rate of convergence of E_{j,Γ_j} on a logarithmic scale. For example, if it exists, the Bahadur slope for $\Gamma_j < \tilde{\Gamma}_j$ would be $\lim_{n\to\infty} n^{-1} \log E_{j,\Gamma_j|n}$ for the jth evidence factor and for the joint evidence it would be $\lim_{n\to\infty} n^{-1} \log E_{\Gamma_1,\Gamma_2|n}$. Rosenbaum (2015) introduced the Bahadur efficiency of sensitivity analysis in this context. Taking cue from that discussion, we consider the probability of large deviation in rejection and acceptance decisions for the evidences. As shown by Rosenbaum (2015), existence of an exact rate depends on the test statistic used. But an upper bound of the rate can always be considered (Dembo & Zeitouni, 2010, §4·5). Let I_{j,Γ_j} be functions defined on $[0, 1]$ taking non-negative values, including possibly ∞, such that, for any compact subset F of $[0, 1]$, for $j = 1, 2$

$$\limsup_{n\to\infty} n^{-1} \log \mathrm{pr}(E_{j,\Gamma_j|n} \in F) \leq - \inf_{x\in F} I_{j,\Gamma_j}(x). \tag{3}$$

Because $\tilde{\Gamma}_j$ is the design sensitivity of jth factor, if $\Gamma_j > \tilde{\Gamma}_j$ we would expect $I_{j,\Gamma_j}(x) > 0$ for any $x < 1$ and if $\Gamma_j < \tilde{\Gamma}_j$ we would expect $I_{j,\Gamma_j}(x) > 0$ for any $x > 0$. In quantitative terms (3) says, when $\Gamma_j > \tilde{\Gamma}_j$ the probability of rejecting the null based on jth factor is less than ε for sample sizes more than $\log(1/\varepsilon)/\inf_{x\in[\alpha,1]} I_{j,\Gamma_j}(x)$. Similarly if $\Gamma_j < \tilde{\Gamma}_j$ the probability of failing to accept the null based on evidence j is less than ε for $n > \log(1/\varepsilon)/\inf_{x\in[0,\alpha]} I_{j,\Gamma_j}(x)$.

We wish to establish that the joint test has a rate which is larger than that of the individual tests. Theorem 2 requires $(E_{1,\Gamma_1}, E_{2,\Gamma_2})$ to be evidence factors in the following sense

$$(E_{1,\Gamma_1|n}, E_{2,\Gamma_2|n}) \succ (\tilde{E}_{1,\Gamma_1|n}, \tilde{E}_{2,\Gamma_2|n}), \tag{4}$$

where $\tilde{E}_{1,\Gamma_1|n}$ and $\tilde{E}_{2,\Gamma_2|n}$ are independently distributed and $\tilde{E}_{j,\Gamma_j|n}$ have the same distribution as $E_{j,\Gamma_j|n}$. While Definition 2 uses stochastic ordering under H_0, (4) is a more general statement also under the alternative hypothesis.

THEOREM 2. *Suppose I_{j,Γ_j} satisfies* (3) *for $j = 1, 2$. Then with $\alpha < 0.20$, for Fisher's combination*

$$\limsup_{n\to\infty} n^{-1} \log \text{pr}(E_{\Gamma_1,\Gamma_2|n} < \alpha) \leq - \inf_{x:\, x\leq\alpha} \max_{j=1,2} I_{j,\Gamma_j}(x), \tag{5}$$

$$\limsup_{n\to\infty} n^{-1} \log \text{pr}(E_{\Gamma_1,\Gamma_2|n} \geq \alpha) \leq - \inf_{x:\, x\geq\alpha} \max_{j=1,2} I_{j,\Gamma_j}(x). \tag{6}$$

Since $E_{j,\Gamma_j|n}$, for $j = 1, 2$, converges to zero or one almost surely, with $\tilde{\alpha}$ fixed, Fisher's method and truncated product are equivalent for large n. Thus, Theorem 2 holds for the truncated products method as well. Theorem 2 does not assume that evidence factors are well behaved, i.e. does not assume that in (3) the limit of $n^{-1} \log \text{pr}(E_{j,\Gamma_j|n} \in F)$ exists. It allows us to make claims about the worst rates, e.g. in terms of $\limsup n^{-1} \log \text{pr}(E_{\Gamma_1,\Gamma_2|n} \geq \alpha)$ and $\limsup n^{-1} \log \text{pr}(E_{\Gamma_1,\Gamma_2|n} < \alpha)$. If in (3), $\limsup$ can be replaced by $\lim$ and equality in place of inequality, hence the exact rates of rejection and acceptance for the factors exists, then both (5) and (6) hold with $\limsup$ replaced by $\lim$. Theorem 2 can be interpreted as: the joint evidence requires a smaller sample size to make the correct decision than the factors considered separately. An illustration of this result is given through simulation in §6. The proof of Theorem 2 is given in the supplement.

If the evidence factors are well behaved, more accurate statements about the rates can be made. Theorem 3 indicates that if individual factors have Bahadur slopes, then the Bahadur slope of the joint evidence is again better than the individuals.

THEOREM 3. *Suppose for a pair (Γ_1, Γ_2), there exits two non-negative numbers r_{1,Γ_1} and r_{2,Γ_2} such that: (i) $n^{-1} \log E_{1,\Gamma_1|n} \to -r_{1,\Gamma_1}$, and (ii) $n^{-1} \log E_{2,\Gamma_2|n} \to -r_{2,\Gamma_2}$. Then for Fisher's combination method, $\lim_{n\to\infty} n^{-1} \log E_{\Gamma_1,\Gamma_2|n} = -(r_{1,\Gamma_1} + r_{2,\Gamma_2})$. Also, if for some non-negative a_{j,Γ_j}, $n^{-1} \log(1 - E_{j,\Gamma_j|n}) \to -a_{j,\Gamma_j}$, for $j = 2$ (or 1), then with (i) (or (ii)), $\lim_{n\to\infty} n^{-1} \log E_{\Gamma_1,\Gamma_2|n} = -r_{\bar{j},\Gamma_{\bar{j}}}$, where $\bar{j} = 1$ (or 2).*

Proof. Recall that $E_{\Gamma_1,\Gamma_2|n} = \text{pr}(\chi^2_4 > -2 \log E_{1,\Gamma_1|n} E_{2,\Gamma_2|n})$. For any t, as $n \to \infty$, $n^{-1} \log \text{pr}(\chi^2_4 > nt^2) \to -t^2/2$. Now under (i) and (ii), $-2n^{-1} \log E_{1,\Gamma_1|n} E_{2,\Gamma_2|n} \to 2(r_{1,\Gamma_1} + r_{2,\Gamma_2})$. Let c and d be any numbers such that $c < (r_{1,\Gamma_1} + r_{2,\Gamma_2}) < d$, then, $2c < -2n^{-1} \log E_{1,\Gamma_1|n} E_{2,\Gamma_2|n} < 2d$ for large enough n. Thus for large n, $n^{-1} \log \text{pr}(\chi^2_4 > 2dn) \leq n^{-1} \log \text{pr}(\chi^2_4 > -2 \log E_{1,\Gamma_1|n} E_{2,\Gamma_2|n}) \leq n^{-1} \log \text{pr}(\chi^2_4 > 2cn)$. Therefore, $-d \leq \liminf_{n\to\infty} n^{-1} \log E_{\Gamma_1,\Gamma_2|n} \leq \limsup_{n\to\infty} n^{-1} \log E_{\Gamma_1,\Gamma_2|n} \leq -c$. Let $c, d \to (r_{1,\Gamma_1} + r_{2,\Gamma_2})$ to get, $\lim_{n\to\infty} n^{-1} \log E_{\Gamma_1,\Gamma_2|n} = -(r_{1,\Gamma_1} + r_{2,\Gamma_2})$.

If, $n^{-1} \log(1 - E_{2,\Gamma_2|n}) \to -a_{2,\Gamma_2}$ and (i) holds, then $2n^{-1} \log E_{1,\Gamma_1|n} E_{2,\Gamma_2|n} \to -2\, r_{1,\Gamma_1}$. The rest of the proof follows the same arguments as above. □

5·3. *Which evidence factor(s) provide evidence?*

In an analysis based on evidence factors, it is useful if the decision to reject the null hypothesis can be attributed to one or both the factors. The closed testing principle of Marcus et al. (1976)

can be used for this purpose. For a pair $(\Gamma_{1i}, \Gamma_{2i'})$, consider three comparisons: (i) $E_{\Gamma_{1i},\Gamma_{2i'}} < \alpha$, (ii) $E_{1,\Gamma_{1i}} < \alpha$, (iii) $E_{2,\Gamma_{2i'}} < \alpha$. If (i), (ii), and (iii) are true, then we reject H_0 based on evidence from both factors. If (i) and (ii) are true, then we reject based on the first factor. Similarly, if (i) and (iii) are true, then we reject based on the second factor. If only (i) is true, then rejection is based on the combined evidence alone, and the rejection decision cannot be attributed to one factor.

We are working in a scenario where it seems plausible that one assumption is true with the other being false. The following argument establishes that the above procedure preserves the probability of rejecting any $\{H_{0,\Gamma_{1i},\Gamma_{2i'}} \mid \Gamma_{1i} \geq \bar{\Gamma}_1\} \cup \{H_{0,\Gamma_{1i},\Gamma_{2i'}} \mid \Gamma_{2i} \geq \bar{\Gamma}_2\}$ at the level α. If H_0 is false there is nothing to prove. Assume H_0 is true. Then, possible scenarios are: (1) H_0 is true and both $\mathcal{A}_1(\Gamma_{1i})$ and $\mathcal{A}_2(\Gamma_{2i'})$ are true; (2) H_0 is true and $\mathcal{A}_1(\Gamma_{1i})$ is true but $\mathcal{A}_2(\Gamma_{2i'})$ is false; and (3) H_0 is true and $\mathcal{A}_2(\Gamma_{2i'})$ is true but $\mathcal{A}_1(\Gamma_{1i})$ is false. For any pair $(\Gamma_{1i}, \Gamma_{2i'})$ at most one of (1)–(3) can be true. When (1) holds, any false rejection implies $\{E_{\bar{\Gamma}_1,\bar{\Gamma}_2} < \alpha\}$, when (2) holds, a false rejection implies $\{E_{\bar{\Gamma}_1} < \alpha\}$, and finally, when (3) holds, a false rejection implies $\{E_{\bar{\Gamma}_2} < \alpha\}$. Thus the familywise error rate is controlled at desired level α.

6. SIMULATION: COMBINED EVIDENCE DOES BETTER IN FINITE SAMPLE

This section aims to verify that the Bahadur efficiency results of §5·2 provide an adequate guide to finite samples. We wish to verify that the combined evidence factor analysis requires a smaller sample size to make the correct decision with high probability than an analysis using either evidence factor alone.

Our simulation is based on the structure of the Life Span Study data (§2). We assume the data have S strata of triplets with exposures zero-dose, low-dose and high-dose. The response is Bernoulli with probability expit(α_s) if exposed zero-dose or expit$(\alpha_s + \beta_l)$ if exposed to low-dose or expit$(\alpha_s + \beta_h)$ for high-dose; here expit$(x) = \exp(x)/\{1 + \exp(x)\}$ and $\beta_l \leq \beta_h$. The strata effect α_s is sampled independently from $N(0{,}0{\cdot}2^2)$. The sample size (n) of §5·2 here is the number of strata (S), and increasing the sample size is equivalent to adding more and more strata while keeping the size of each stratum fixed. Figure 2 summarizes the simulation results in three panels of plots (A)–(C). Each panel corresponds to a separate simulation scenario with varied values of the effects β_l, β_h. Within each panel three plots correspond to three different pairs of values of (Γ_1, Γ_2). Each plot shows the performance of the various tests as the sample size increases. Recall that Γ_1 is the sensitivity parameter for the high versus low dose comparison and Γ_2 is the sensitivity parameter for exposed versus unexposed comparison.

Panel (A) considers the null case $\beta_l = \beta_h = 0$. Recall that the simulation does not impose any bias in treatment assignment. In this situation, even a small amount bias will cause the probability of rejection to go to zero as the number of strata increases. A test is better if the rate at which this probability of rejection, plotted on the vertical axis, goes to zero is as fast as possible. For the graphs of panel (A), the higher the value, the faster, i.e. with less number of strata, we fail to reject the null on average. In plot A1, where $\Gamma_1 = \Gamma_2 = 1.1$, we see that as the number of strata increases the combined evidence narrowly beats both the factors. In plots A2 and A3 one of the two Γ_j values is large. The comparison with larger Γ_j always makes the correct decision; at least in the simulations. This is shown as a horizontal line at infinity. The plots show that the combined evidence dominates.

Panels (B) and (C) consider two scenarios under the alternative hypothesis: only high-dose has an effect, $\beta_l = 0, \beta_h = 0.5$, and both low-dose and high-dose has an effect, $\beta_l = 0.5, \beta_h = 1$, respectively. Here, in both the factors, as the number of strata increases the probability of acceptance will go to zero for bias below the design sensitivity. In these plots, the larger the

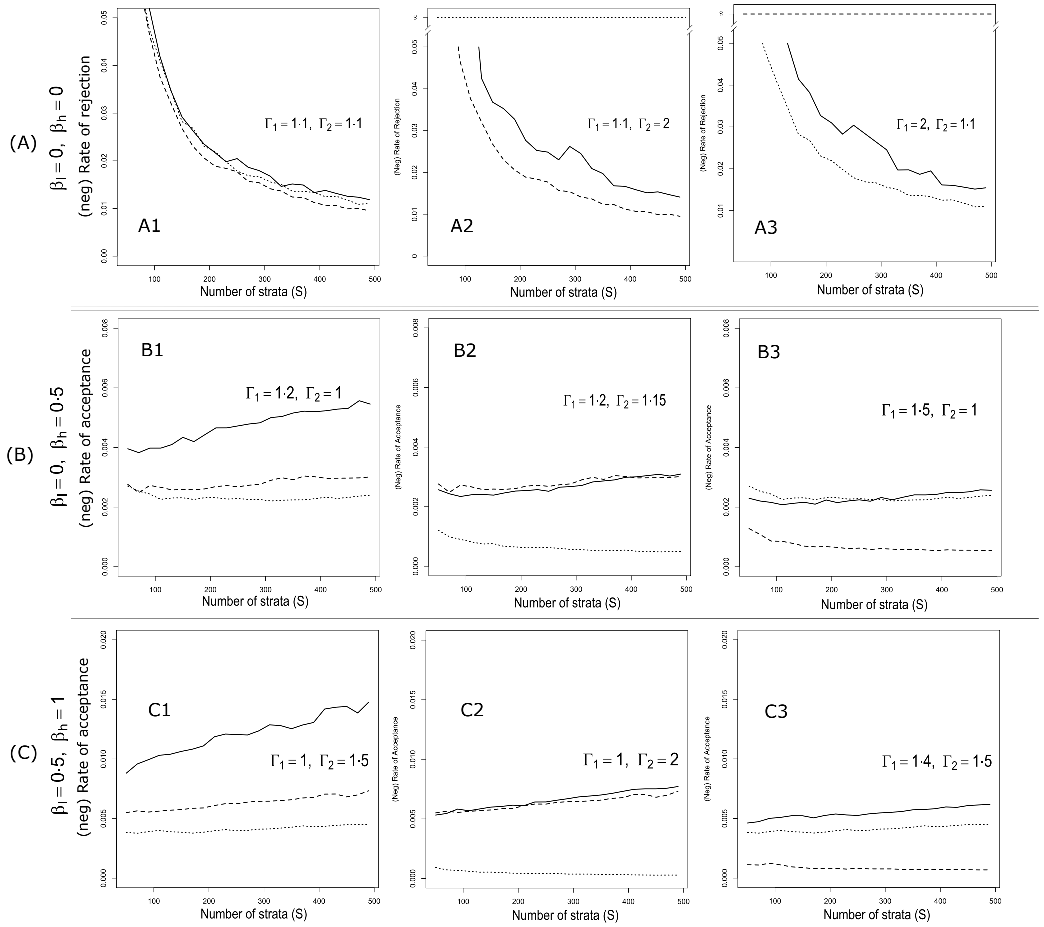

Fig. 2. Graphs for combined evidence (solid), high to low dose (dashes) and exposed versus unexposed (dots). Panel (A) plots the negative of the rate of rejection of the hypothesis, $-\log(\text{probability of rejection})/S$, in the null scenario, against the number of strata (S). Panels (B) and (C) plot the negative of the rate of acceptance of the hypothesis, $-\log(\text{probability of acceptance})/S$ against S. Along the rows Γ_1 and Γ_2 are varied. Results are based on average over 2000 simulations and over a grid of S values in gaps of 20.

graph is on the vertical axis plotting the rate of acceptance, the faster the null is rejected and the smaller the number of strata required to attain a certain power.

In panel (B), the average (attributable) treatment effect in the comparison of exposed to unexposed units is considerably smaller. Consequently, the design sensitivity is smaller for the 385
exposed versus unexposed comparison than that of the high versus low-dose comparison. Plot B1 considers bias levels $\Gamma_1 = 1{\cdot}2$ and $\Gamma_2 = 1$. These Γ values are chosen so that the power is not close to 0 or 1; otherwise we would not be able to compare methods clearly and get a sense of the rate based on the simulations. Correspondingly, as the effect in the comparison of high to low-dose is larger than that of exposed to unexposed comparison, Γ_1 is chosen to be larger 390
than Γ_2. In this plot, the combined evidence dominates both the factors. For the next two plots, one of the two bias parameters are large so that the corresponding analysis is no longer able to detect the treatment effect with high power. Thus in these two scenarios the combined evidence borrows its strength mostly from only one factor. These plots show that as the number of strata increases, the rate for the combined evidence catches up with the better of the two factors. 395

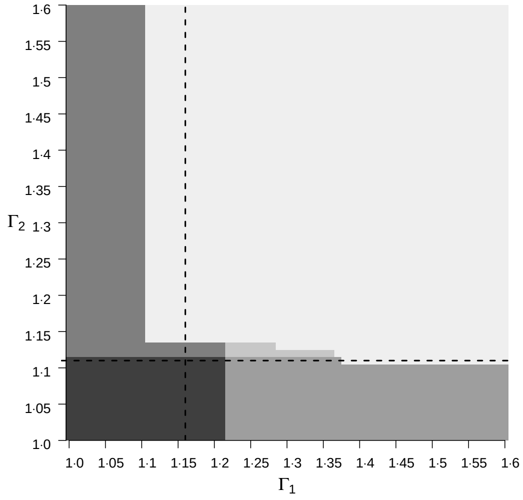

Fig. 3. Result for testing radiation effect on solid cancer. In decreasing order of gradient the colors represent the decision - reject for both comparisons, for high to low dose comparison, for city to not-in-city comparison, without any attribution and do not reject. Bonferroni method rejects the null if Γ_1 <1·16 and Γ_2 <1·11 (dashed lines).

Finally, in panel (C), the design sensitivity is smaller for the high versus low-dose comparison. The plots have similar behavior as in the plots of panel (B). Plot C1 of this panel considers the bias levels $\Gamma_1 = 1$ and $\Gamma_2 = 1{\cdot}5$. The combined evidence has better performance compared to either of the factors. For the last two plots, as in panel (B), one of the bias parameters is taken to be large enough so that the corresponding analysis is no longer able to detect the treatment effect with high power. In both these scenarios we see the combined evidence has comparable performance to the better of the two factors.

7. ANALYSIS OF THE LIFE SPAN STUDY DATA

The analysis to assess whether radiation has any carcinogenic effect consists of two comparisons, one based on comparing all proximal survivors with low and high doses and a second one comparing proximal survivors to not-in-city residents, giving us two evidence factors with $E_1 = 0{\cdot}0021$ and $E_2 = 2{\cdot}35 \times 10^{-10}$, see §2. The fact that these two comparisons form evidence factors is proved explicitly in the supplement.

In the Life Span Study data, the observed confounders are age at exposure and sex. Then the bias levels Γ_1 and Γ_2 measure deviation from the assumptions that there is no unmeasured confounding for the comparisons of high-dose versus low-dose proximal survivors and all proximal survivors versus not-in-city residents, respectively, among individuals in the same strata of age at exposure and sex. The conclusion from the first comparison is sensitive at bias level $\Gamma_1 = 1{\cdot}25$, i.e. we first fail to reject the hypothesis when $\Gamma_1 = 1{\cdot}25$, whereas the conclusion from the second comparison is sensitive at bias level $\Gamma_2 = 1{\cdot}12$. Therefore, to explain the observed associations, an unmeasured confounder, as in §2, would need to have a relatively weaker association with exposure to radiation in comparing all proximal survivors to not-in-city residents than in comparing high-dose to low-dose survivors. Figure 3 presents the results based on evidence factors. The factors are combined using the truncated product method with $\tilde{\alpha} = 0{\cdot}20$ (see §4). The results show that the joint evidence is statistically significant for a carcinogenic effect for $(\Gamma_1, \Gamma_2) = (1{\cdot}35, 1{\cdot}12)$. However this decision cannot be attributed to either of the comparisons; at $(\Gamma_1, \Gamma_2) = (1{\cdot}35, 1{\cdot}12)$ each of the evidences considered separately are sensitive. At $(\Gamma_1, \Gamma_2) = (1{\cdot}1, 1{\cdot}3)$ the null hypothesis is rejected based on the evidence from comparison of proximal survivors with low and high doses. Another method to control for familywise error rate would be to use the Bonferroni correction. This leads to failing to reject the null for $\Gamma_1 \geq 1{\cdot}16$ and $\Gamma_2 \geq 1{\cdot}11$. Clearly, the Bonferroni method is conservative for small bias levels. For instance, at bias levels $(\Gamma_1, \Gamma_2) = (1{\cdot}2, 1{\cdot}13)$, we fail to reject the null after applying Bonferroni correction, but reject the null based on the joint evidence.

The sensitivity parameters, Γ_1 and Γ_2, models the biases in treatment assignment due to imbalance in unmeasured confounders. When calculating the evidence using this model, a near-perfect relationship is assumed between the unmeasured confounders and the response. It is not necessary to assume this near-perfect relationship – the one-parameter model with sensitivity parameter Γ is equivalent to a set of models where with two sensitivity parameters: one that relates the unmeasured confounder to the response, Δ; and one that relates the unmeasured confounder to the treatment, Λ. Rosenbaum and Silber (2009) show that for each Γ in the one-parameter model, there is a curve of Λ and Δ in this two-parameter model that gives equivalent inferences. For example, it follows that $\Gamma = 1{\cdot}25$ is equivalent to an unobserved covariate that doubles the odds of treatment $(\Lambda = 2)$ and doubles the odds of a positive treated-minus-control response difference $(\Delta = 2)$. In the supplement, we provide the technical discussion of this correspondence.

8. Discussion

Unmeasured confounding is a challenge in observational studies. Evidence factors, by constructing multiple independent sources of evidence that are potentially vulnerable to separate sources of unmeasured confounding, help us to either detect potential unmeasured confounding or make our findings more robust to unmeasured confounding. A practitioner might be concerned with loss of power from multiple comparisons when using evidence factors; this paper establishes that if one constructs evidence factors and uses them carefully, as described in Theorem 1, there is no loss in power.

An alternative strategy in the Life Span Study could have been to select one of the two reference groups, distal survivors or non-in-city residents, and present a single analysis (French et al., 2017). We showed that both reference groups can be used to build evidence factors. The combination of the factors provided evidence against the null hypothesis of no carcinogenic effect, which was robust to multiple sources of unmeasured confounding.

Our analysis was limited in the sense that it addressed whether there was a carcinogenic effect of radiation, but did not address the dose-response relationship. Currently, there is strong scientific interest in the shape of the dose-response curve, particularly at lower radiation doses, as well as differences in radiation risk by various demographic and lifestyle factors. The Life Span Study data contain rich individual level information that can be used to model these associations. Future research might seek to build evidence factors, perhaps by comparing survivors across multiple reasons of radiation exposure, to infer about the radiation dose-response and effect modification.

Acknowledgement and Code/Data Availability

The Radiation Effects Research Foundation (RERF), Hiroshima and Nagasaki, Japan is a public interest foundation funded by the Japanese Ministry of Health, Labour and Welfare (MHLW) and the U.S. Department of Energy (DOE). The research was also funded in part through DOE award DE-HS0000031 to the National Academy of Sciences. This publication was supported by RERF Research Protocol RP-S3-16. The views of the authors do not necessarily reflect those of the two governments. The Life Span Study data used in the paper are available for download from www.rerf.or.jp/en/library/data-en/. The code used to analyze the data is submitted along with this paper and available upon request.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of Corollary 1 and Theorem 2, simulation comparison of two combining methods discussed in §4, and pseudo code for an algorithm discussed in §5·1.

References

Becker, B. J. (1994). Combining significance levels. In *The handbook of research synthesis*, Ed. H. Cooper and L. V. Hedges, pp. 215–230. New York: Russell Sage Foundation.

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. of Math. Statist.* **38**, 303–324.

Chen, Z. & Nadarajah, S. (2014). On the optimally weighted z-test for combining probabilities from independent studies. *Comput. Stat. Data Anal.* **70**, 387–394.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer. *J. Nat. Cancer Inst.* **22**, 173–203.

Cullings, H. M., Grant, E. J., Egbert, S. D., Watanabe, T., Oda, T., Nakamura, F., Yamashita, T. et al. (2017). DS02R1: Improvements to atomic bomb survivors' input data and implementation of dosimetry system 2002 (DS02) and resulting changes in estimated doses. *Health Phys.*, **112**, 56–97.

Dembo, A. & Zeitouni, O. (2010). *Large Deviations Techniques and Applications*. Berlin: Springer-Verlag.

Ding, P. & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27**, 368–377.

French, B., Cologne, J., Sakata, R., Utada, M. & Preston, D. L. (2017). Selection of reference groups in the Life Span Study of atomic bomb survivors. *Eur. J. Epidemiol.* **32**, 1055–1063.

Gastwirth, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimet. J.* **33**, 19–34.

Hosman, C. A., Hansen, B. B. & Holland, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4**, 849–870.

Hsu, J. Y., Small, D. S. & Rosenbaum, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Am. Statist. Assoc.* **108**, 135–148.

Keele, L. & Minozzi, W. (2013). How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data. *Political Anal.* **21**, 193–216.

Liptak, T. (1958). On the combination of independent tests. *Magyar. Tud. Akad. Mat. Kutato. Int. Kozl.* **3**, 171–197.

Littell, R. C. & Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *J. Am. Statist. Assoc.* **66**, 802–806.

Marcus, R., Peritz, E. & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

McCandless, L. C. & Gustafson, P. (2017). A comparison of Bayesian and Monte Carlo sensitivity analysis for unmeasured confounding. *Statist. Med.* **36**, 2887–2901.

Neyman, J. (1923). On the spplication of probability theory to agricultural experiments: essay on principles, Section 9. Reprinted in *Statist. Sci.* **5**, 465–472.

Pierce, D. A. & Preston, D. L. (2000). Radiation-related cancer risks at low doses among atomic bomb survivors. *Radiat. Res.* **154**, 178–186.

Pierce, D. A., Vaeth, M. & Shimizu, Y. (2007). Selection bias in cancer risk estimation from a-bomb survivors. *Radiat. Res.* **167**, 735–741.

Preston, D. L., Ron, E., Tokuoka, S., Funamoto, S., Nishi, N., Soda, M., Mabuchi, K. & Kodama, K. (2007). Solid cancer incidence in atomic bomb survivors: 1958 - 98. *Radiat. Res.* **168**, 1–64.

Pimentel, S. D., Small, D. S. & Rosenbaum, P. R. (2015). Constructed second control groups and attenuation of unmeasured biases. *J. Am. Statist. Assoc.* **111**, 1157–1167.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74**, 13–26.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.

Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91**, 153–164.

Rosenbaum, P. R. (2010). Evidence factors in observational studies. *Biometrika* **97**, 333–345.

Rosenbaum, P. R. (2011). Some approximate evidence factors in observational studies. *J. Am. Statist. Assoc.* **106**, 285–295.

Rosenbaum, P. R. (2015). Bahadur efficiency of sensitivity analyses in observational studies. *J. Am. Statist. Assoc.* **110**, 205–217.

Rosenbaum, P. R. & Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Am. Statist. Assoc.* **104**, 1398–1405.

Shaked, M. & Shanthikumar, J. G. (2007). *Stochastic Orders*. New York: Springer.

Stuart, E. A., DuGoff, E., Abrams, M., Salkever, D. & Steinwachs, D. (2013). Estimating causal effects in observational studies using electronic health data: challenges and (some) solutions. *eGEMs* **1**, Article 4.

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18**, 1368–1373.

Won, S., Morris, N., Lu, Q. & Elston, R. C. (2009). Choosing an optimal method to combine p-values. *Statist. Med.*, **28**, 1537–1553.

Zaykin, D., Zhivotovsky, L. A., Westfall, P. & Weir, B. (2002). Truncated product method for combining p-values. *Genet. Epidemiol.*, 22, 170–185.

Zubizarreta, J. R., Neuman, M., Silber, J. H. & Rosenbaum, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *J. Am. Statist. Assoc.* **107**, 901–915.