IEA
International Epidemiological Association

International Journal of Epidemiology, 2022, 355–363
https://doi.org/10.1093/ije/dyac018
Advance Access Publication Date: 1 March 2022
Opinion

OXFORD

Opinion

Are E-values too optimistic or too pessimistic? Both and neither!

Arvid Sjölander[1]* and Sander Greenland[2]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Solna, Sweden and [2]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA

*Corresponding author. Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Nobels väg 12A, 171 65 Solna, Sweden. E-mail: arvid.sjölander@ki.se

Received 29 September 2021; Editorial decision 17 January 2022; Accepted 25 January 2022

Key words: Bias, causal inference, confounding, E-value, quantitative bias analysis, sensitivity analysis

Introduction

Ding and VanderWeele,[1] hereafter DV, proposed a method to assess the sensitivity of observed associations to uncontrolled confounding. Briefly, this method requires the analyst to provide guesses of two sensitivity parameters RR_{UD} and RR_{EU}, loosely defined as the maximal strength of association that an uncontrolled (set of) confounder(s) may have with the outcome and with the exposure, respectively. DV derived a lower bound for the causal exposure–outcome risk ratio as a function of these sensitivity parameters and the observed exposure–outcome risk ratio. By setting the sensitivity parameters to values that are considered plausible for the study at hand, one obtains a lower bound for the causal risk ratio.

In a subsequent paper, VanderWeele and Ding[2] coined the term 'E-value' for the common value $RR_{UD} = RR_{EU}$ of the sensitivity parameters that gives a lower bound equal to 1; i.e. the E-value shows the minimum size these sensitivity parameters must have if they are equal and the confounding that they produce is exactly the inverse of the observed association. For an observed risk ratio RR^{obs}_{ED} above 1, the E-value turns out to be $RR^{obs}_{ED} + \left(RR^{obs}_{ED}\left(RR^{obs}_{ED} - 1\right)\right)^{1/2}$; it measures how strong an association the uncontrolled confounder must have with exposure and outcome to entirely explain away RR^{obs}_{ED}. The larger the E-value, the stronger the required uncontrolled confounding, and (presumably) the more trustworthy the result of the study.

DV's two papers have quickly become influential; as of 28 September 2021, they have 314 and 1431 citations, respectively, according to Google Scholar. In a systematic literature review up to the end of 2018, Blum *et al.*[3] found 87 papers presenting 516 E-values, and E-values have now been recommended as a main basis for evaluating residual confounding.[4] However, the E-value has also been subject to quite intense debate and criticism.[5–13] The critics agree that the E-value is merely a transform of the observed risk ratio and thus can mislead because it uses 'no background or data information on confounders or prevalences, and no expectations about unobserved confounders or correlations with controlled confounders'.[8]

We will elaborate on two important ensuing criticisms. Ioannidis *et al.*[5] argued that a high E-value may not provide much assurance because 'if dozens of unknown confounders exist, such a [large] composite effect [i.e. such large values of the sensitivity parameters] might not be totally implausible, even if each confounder's strength of association [with the exposure and the outcome] is modest'. In contrast, Greenland[8] argued that a low E-value may give an unnecessarily pessimistic impression of the study, since 'confounding by unmeasured factors may be weakened considerably due to their associations with strong controlled confounders (e.g. age and sex)', and MacLehose *et al.*[12] argued that 'the calculation of E-values for unknown and unsuspected confounders is an exercise in unwarranted paranoia, given the lack of history of plausible associations in epidemiology being completely refuted by the belated discovery of previously unsuspected confounders '.

At first glance, these ensuing criticisms may seem somewhat contradictory insofar as Ioannidis *et al.*[5] fear the E-value can lead to underestimation of uncontrolled confounding bias, whereas the other two fear the E-value will lead to overestimation of the same bias. We aim to reconcile these two criticisms, and to show that there is no logical conflict between them. We will argue that the true values of DV's sensitivity parameters can be very large and thus exceed what would often be considered a high E-value even if the actual confounding bias is small. From this we further argue that the E-value can be uninformative in that it can lead to false optimism when large and false pessimism when small. In this way, we provide further evidence for the severe concern expressed by MacLehose *et al.*[12] that

> The calculation of E-values for known but unmeasured confounders is irresponsible, as it makes no use of the information on those covariates that make them plausible to view as confounders. A desire for sensitivity analyses without assumptions is a desire to do inference in basic ignorance of background context.

To illustrate our points in a realistic setting, we provide a simulation based on data from a recently published study on non-steroidal anti-inflammatory drugs (NSAIDs) and psychiatric disorders following a cancer diagnosis.[14] In a real study, neither the causal risk ratio nor DV's sensitivity parameters are known. However, in our simulation, we are able to compute these quantities for various scenarios by assuming that a full set of available confounders is sufficient for confounding control, and dividing this set into subsets of controlled and uncontrolled confounders of various size. Thus, we are able to study how the bias of the observed risk ratio, and the E-value and DV's lower bound for the causal risk ratio, vary with different combinations of controlled/uncontrolled confounders.

Sensitivity parameters and lower bound

We first briefly recap the key components of the sensitivity analysis by Ding and VanderWeele.[1] Let E and D be a binary exposure and outcome of interest, respectively, e.g. NSAID use and psychiatric diagnosis before end of follow-up. Let C be a set of measured confounders that are controlled in the analysis. The 'observed' risk ratio, given $C = c$, is defined as:

$$RR^{obs}_{ED|c} = \frac{p(D=1|E=1,C=c)}{p(D=1|E=0,C=c)}.$$

Writing $D(e)$ for the potential outcome for a given patient if the exposure were set to level e, the 'true' (i.e. causal) risk ratio, given $C = c$, is defined as:

$$RR^{true}_{ED|c} = \frac{p\{D(1)=1|C=c\}}{p\{D(0)=1|C=c\}}.$$

Let U be a set of confounders that are not controlled; typically, U would be unmeasured. If C and U together are sufficient for confounding control (and the standard assumptions of consistency and positivity hold), $RR^{true}_{ED|c}$ is equal to the standardized risk ratio:

$$\frac{\sum_u p(D=1|E=1,C=c,U=u)p(U=u|C=c)}{\sum_u p(D=1|E=0,C=c,U=u)p(U=u|C=c)}.$$

DV defined the sensitivity parameters:

$$RR_{EU|c} = \max_u \left\{ \frac{p(U=u|E=1,C=c)}{p(U=u|E=0,C=c)} \right\}$$

and

$$RR_{UD|c} = \max_e \left\{ \frac{\max_u p(D=1|E=e,C=c,U=u)}{\min_u p(D=1|E=e,C=c,U=u)} \right\},$$

and the bounding factor:

$$BF_{U|c} = \frac{RR_{EU|c}RR_{UD|c}}{RR_{EU|c}+RR_{UD|c}-1}.$$

They showed that $RR^{true}_{ED|c}$ is bounded below by:

$$RR^{true}_{ED|c} \geq \frac{RR^{obs}_{ED|c}}{BF_{U|c}},$$

and they defined the E-value as the common value of $RR_{EU|c} = RR_{UD|c}$ for which the lower bound is 1, which simplifies as shown before.

Data

We give a brief explanation of the data here; see Hu *et al.*[14] for details. From the Swedish Cancer Registry, 338 009 patients were identified that were diagnosed with cancer between 1 July 2006 and 31 December 2013. Information was obtained on NSAID use within 1 year before cancer diagnosis (the exposure), dichotomized as 'no/yes', and on diagnosis dates of psychiatric disorders (depression, anxiety or stress) during follow-up (the outcome). In the original analysis of these data, Hu *et al.*[14]

restricted the follow-up to 1 year after cancer diagnosis, with censoring at death or emigration. To increase the number of observed outcome events, we extended the follow-up in our reanalysis to a maximum of 5 years. Fourteen baseline covariates were measured: age, calendar year, cancer type (prostate, breast, gastrointestinal, lung, skin excluding melanoma, melanoma, kidney + bladder, gynaecologic, haematological, other), indicator of inflammatory musculoskeletal disorders within 1 year before cancer diagnosis, indicator of pain and fever within 1 year before cancer diagnosis, residency (east, south, north), chronic disease score (ranging from 0 to 11) based on dispensed medication within 1 year before cancer diagnosis, sex, occupation (blue-collar, white-collar, not working, unknown), indicator of cardiovascular disease at cancer diagnosis, marital status [unmarried, married, divorced, widow(er)], cancer stage (early, local spread, regional spread, metastatic, leukaemia, lymphoma, myeloma, myelodysplastic syndrome, myeloproliferative neoplasm, unknown), indicator of inflammatory system diseases within 1 year before cancer diagnosis, education level (9 years, college, high school, unknown).

After various exclusions described by Hu *et al.*[14] and restriction to patients with complete information on all variables listed above, the data set comprises 288 253 patients of whom 11 357 (3.9%) had psychiatric diagnoses during follow-up. In their original analysis, Hu *et al.*[14] considered both aspirin and non-aspirin NSAIDs as exposures. We restrict attention to the latter, but for simplicity we from now on refer to non-aspirin NSAIDs as just NSAIDs. Among all patients, 63 945 (22.2%) used NSAIDs within 1 year before cancer diagnosis. One may argue that some of the measured covariates (e.g. cancer stage) could possibly act as mediators for the effect of NSAIDs on psychiatric diagnoses, rather than confounders. Distinguishing between mediators and confounders is an important issue, but beyond the scope of this paper. We thus ignore this issue and assume that all baseline covariates are truly confounders. For computational reasons, we categorized age as 24–49, 50–59, 60–69, 70–106 years; calendar year as 2006–2007, 2008–2009, 2010–2011, 2012–2013; and chronic disease score as 0, 1–3, 3–11; and merged cancer stage into stage I (early, myeloproliferative neoplasm, lymphoma), stage II (local spread, myelodysplastic syndrome, myeloma, unknown) and stage III (regional spread, metastatic, leukaemia), and cancer type into type I (melanoma, prostate, skin excl. melanoma), type II (breast, gastrointestinal, kidney + bladder, gynaecologic, haematological, other) and type III (lung). Thus, all 14 confounders were coded as categorical variables, with at most four levels each.

Methods

We started with a set of analyses aimed to describe the associations between the variables in the study. We estimated the pairwise association between the 14 measured confounders, using the bias-corrected version of Cramer's V. This measure of association ranges from 0 (minimal association) to 1 (maximal association). We estimated the association between the measured confounders and NSAID use by regressing the latter on all the former simultaneously, with multivariable logistic regression. We estimated the association between the measured confounders and NSAID use, and time to psychiatric diagnosis, by regressing time to diagnosis on confounders and NSAID use with multivariable Cox proportional-hazards regression.

To illustrate the implications of uncontrolled confounding for the E-value, we carried out a simulation based on the real data. We give a non-technical explanation of the simulation here, and provide technical details in the online Supplementary material (available as Supplementary data at *IJE* online). In the simulation, we defined the binary outcome D as the occurrence of a psychiatric diagnosis before end of follow-up. We provisionally assumed there were no other confounders for the exposure–outcome association than the 14 available confounders. Since our focus is on bias, we ignored statistical uncertainty throughout the simulation. Readers uncomfortable with this approach may either pretend that our sample is identical to the target population or that each patient in the sample represents a large number (e.g. thousands) of patients with identical outcome, exposure and confounder values.

Well-designed observational studies do not select confounders randomly, but according to their perceived importance for the exposure–outcome association. To reflect common practice, we first used a change-in-estimate procedure to order the 14 confounders. For this purpose, we used the function 'chestcox' from the R package 'chest', which adds confounders sequentially to a Cox proportional-hazards model, in each step selecting the confounder from the remaining set that maximizes the relative change in the estimated exposure–outcome hazard ratio.[15]

We next carried out a step-wise analysis in which we simulated control for an increasingly larger set of confounders. In each step $k = 0, \ldots, 14,$ we controlled for the k first confounders selected by the change-in-estimate procedure. Thus, in the notation of the 'Sensitivity parameters and lower bound' section, the set C of controlled confounders increased from the empty set in Step 0 to the set of all 14 available confounders in Step 14, whereas the set U of uncontrolled confounders was taken in the opposite direction. In each step of the analysis, we fitted a logistic regression model for the exposure as a function of the controlled confounders, and a Cox proportional-hazards model for the time to diagnosis as a function of the exposure and the controlled confounders. Using the fitted Cox model, we computed the observed risk ratio in each step k of the analysis, i.e. for each set of controlled confounders.

In a real study, the true risk ratio and DV's sensitivity parameters are typically not estimable from the observed data, because the uncontrolled confounders U are unmeasured. However, in our simulation, U was a subset of the 14 available confounders, which allowed us to use the fitted regression models to compute both the true risk ratio and the sensitivity parameters, in each step k of the analysis. These computations are relatively straightforward applications of the definitions given in the 'Sensitivity parameters and lower bound' section. For instance, the parameter $RR_{EU|c}$ was obtained by using the fitted logistic regression models to compute the ratio $p(U = u|E = 1, C = c)/p(U = u|E = 0, C = c)$ for all possible combinations of the uncontrolled confounders (i.e. for all levels u), then setting $RR_{EU|c}$ equal to the largest ratio among all these combinations. The parameter $RR_{UD|c}$ was obtained in a similar way using the fitted Cox regression models. We refer the reader to the online Supplementary material (available as Supplementary data at *IJE* online) for a detailed explanation of these computations. We used the true risk ratio and the sensitivity parameters, together with the observed risk ratio, to compute the relative bias of the observed risk ratio, the E-value and the lower bound for the true risk ratio.

Since the outcome is rare, we approximated risk ratios for the binary outcome D with cumulative hazard ratios for the time to diagnosis (Rothman *et al.*,[16] Ch. 4). Under the Cox proportional-hazards model, the observed risk ratio $RR^{obs}_{ED|c}$, the true risk ratio $RR^{true}_{ED|c}$ and the sensitivity parameter $RR_{UD|c}$ thus became approximately constant across levels c of the controlled confounders C. However, the sensitivity parameter $RR_{EU|c}$ was not constant across c under the logistic regression model that we used. Since DV's lower bound for the true risk ratio holds within each level c, we used the smallest value of $RR_{EU|c}$ over all levels c when computing the lower bound in each step k, thus giving the best (i.e. largest) bound among all c-specific bounds.

Results

Table 1 shows the estimated pairwise associations among the 14 measured confounders. These range from 0 (e.g. calendar year vs sex) to 0.40 (e.g. cancer type vs sex). Table 2 shows the estimated odds ratios between the measured confounders and NSAID use, with corresponding P-values. The odds ratios range from 0.66 (age 70–106 years) to 2.42 (musculoskeletal disorder), and most have $P < 0.001$ showing that the data are far from what would be expected from a randomized trial. Table 3 shows the estimated hazard ratios associating time to psychiatric diagnosis with the measured confounders plus NSAID use, with corresponding P-values. The hazard ratios range from 0.35 (age 70–106 years) to 2.13 (cancer type III), and again most have $P < 0.001$. These results indicate that the associations span a fairly wide range. The estimated hazard ratio when comparing NSAID-user to non-users controlling for all measured confounders is 1.15 with 95% compatibility ('confidence') interval[17–19] (1.10, 1.20). Assuming no other

Table 1 Pairwise associations between the 14 measured confounders

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	0.01	0.1	0.1	0.06	0.03	0.26	0.13	0.39	0.18	0.23	0.07	0.03	0.17
2	0.01	1	0.01	0.02	0.01	0.01	0.02	0	0.01	0.02	0.01	0.07	0	0.03
3	0.1	0.01	1	0.03	0.02	0.02	0.04	0.4	0.04	0.04	0.07	0.32	0	0.04
4	0.1	0.02	0.03	1	0.02	0.01	0.1	0.02	0.08	0.05	0.05	0.04	0.06	0.03
5	0.06	0.01	0.02	0.02	1	0.02	0	0.04	0.04	0	0.02	0.02	0.02	0.03
6	0.03	0.01	0.02	0.01	0.02	1	0.03	0.01	0.04	0.01	0.04	0.01	0	0.07
7	0.26	0.02	0.04	0.1	0	0.03	1	0.05	0.22	0.32	0.11	0.03	0.07	0.13
8	0.13	0	0.4	0.02	0.04	0.01	0.05	1	0.03	0.08	0.22	0.27	0.03	0.03
9	0.39	0.01	0.04	0.08	0.04	0.04	0.22	0.03	1	0.15	0.16	0.04	0.03	0.24
10	0.18	0.02	0.04	0.05	0	0.01	0.32	0.08	0.15	1	0.08	0.03	0.03	0.08
11	0.23	0.01	0.07	0.05	0.02	0.04	0.11	0.22	0.16	0.08	1	0.03	0.02	0.1
12	0.07	0.07	0.32	0.04	0.02	0.01	0.03	0.27	0.04	0.03	0.03	1	0.01	0.04
13	0.03	0	0	0.06	0.02	0	0.07	0.03	0.03	0.03	0.02	0.01	1	0.01
14	0.17	0.03	0.04	0.03	0.03	0.07	0.13	0.03	0.24	0.08	0.1	0.04	0.01	1

Confounders are ordered: 1 age, 2 calendar year, 3 cancer type, 4 indicator of inflammatory musculoskeletal disorders, 5 indicator of pain and fever, 6 residency, 7 chronic disease score, 8 sex, 9 occupation, 10 indicator of cardiovascular disease, 11 marital status, 12 cancer stage, 13 indicator of inflammatory system diseases, 14 education level.

Table 2 Estimated odds ratios associating the measured confounders with non-steroidal anti-inflammatory drug use obtained from multivariable logistic regression with corresponding *P*-values

Confounder	Estimated odds ratio	*P*
Age (years)		
24–49	1	–
50–59	1	0.943
60–69	0.9	<0.001
70–106	0.66	<0.001
Calendar year		
2006–2007	1	–
2008–2009	0.93	<0.001
2010–2011	0.87	<0.001
2012–2013	0.78	<0.001
Cancer type		
I	1	–
II	1.03	0.032
III	1.35	<0.001
Musculoskeletal disorder		
No	1	–
Yes	2.42	<0.001
Pain or fever		
No	1	–
Yes	1.45	<0.001
Residency		
East	1	–
North	1.04	0.002
South	1.04	<0.001
Chronic disease score		
0	1	–
1–3	1.05	<0.001
4–11	1.04	0.004
Sex		
Male	1	–
Female	1.09	<0.001
Occupation		
Blue-collar	1	–
Not working	0.85	<0.001
Unknown	0.88	0.116
White-collar	0.82	<0.001
Cardiovascular disease		
No	1	–
Yes	0.89	<0.001
Marital status		
Divorced	1	–
Married	0.96	<0.001
Unmarried	0.81	<0.001
Widow(er)	0.84	<0.001
Cancer stage		
I	1	–
II	1.07	<0.001
III	1.08	<0.001
Inflammatory disease		
No	1	–
Yes	1.25	<0.001
Education		
9 years	1	–
High school	1.04	<0.001
College	1.01	0.493
Unknown	1.09	0.034

Confounders are listed in the order selected by the change-in-estimate procedure.

Table 3 Estimated hazard ratios associating the measured confounders with non-steroidal anti-inflammatory drug use and time to psychiatric diagnosis obtained from multivariable Cox proportional-hazards regression with corresponding *P*-values

NSAID use and confounder	Estimated hazard ratio	*P*
NSAIDs		
No	1	–
Yes	1.15	<0.001
Age (years)		
24–49	1	–
50–59	0.7	<0.001
60–69	0.39	<0.001
70–106	0.35	<0.001
Calendar year		
2006–2007	1	–
2008–2009	0.9	<0.001
2010–2011	0.73	<0.001
2012–2013	0.51	<0.001
Cancer type		
I	1	–
II	1.51	<0.001
III	2.13	<0.001
Musculoskeletal disorder		
No	1	–
Yes	1.04	0.093
Pain or fever		
No	1	–
Yes	1.8	<0.001
Residency		
East	1	–
North	0.67	<0.001
South	0.74	<0.001
Chronic disease score		
0	1	–
1–3	1.17	<0.001
4–11	1.33	<0.001

(Continued)

Table 3 Continued

NSAID use and confounder	Estimated hazard ratio	*P*
Sex		
Male	1	–
Female	1.31	<0.001
Occupation		
Blue-collar	1	–
Not working	1.45	<0.001
Unknown	1.23	0.186
White-collar	0.97	0.425
Cardiovascular disease		
No	1	–
Yes	1.18	<0.001
Marital status		
Divorced	1	–
Married	0.7	<0.001
Unmarried	0.86	<0.001
Widow(er)	0.75	<0.001
Cancer stage		
I	1	–
II	1.11	<0.001
III	1.47	<0.001
Inflammatory disease		
No	1	–
Yes	1.12	0.065
Education		
9 years	1	–
High school	0.98	0.343
College	1.05	0.056
Unknown	0.98	0.823

Confounders are listed in the order selected by the change-in-estimate procedure. NSAID, non-steroidal anti-inflammatory drug.

confounders than these 14, this hazard ratio is approximately equal to the true risk ratio for the binary event of a psychiatric diagnosis before end of follow-up.

Our change-in-estimate procedure selected the 14 measured confounders in the order listed in the 'Data' section and indicated by the caption of Table 1. Figure 1 shows the results of the step-wise analysis; the observed risk ratio (top-left panel), the relative bias of the observed risk ratio (top-right panel), the E-value (middle-left panel), the sensitivity parameters $RR_{EU|c}$ (middle-right panel) and $RR_{UD|c}$ (bottom-left panel) and the lower bound for the true risk ratio (bottom-right panel) as functions of the number of controlled confounders. With no confounder control, the observed risk ratio is equal to 1.24, which gives a quite modest bias of (1.24–1.15)/1.24 = 7% and an E-value equal to 1.79. However, even the minimal (over controlled confounder levels c) sensitivity parameter $RR_{EU|c}$ is very large (= 9.93), and the sensitivity parameter $RR_{UD|c}$ is enormous (= 261). As a consequence, the lower bound for the true risk ratio is very small (= 0.13) and virtually uninformative. As the number of controlled confounders increases the bias decreases to 0, the E-value decreases to 1.57, the sensitivity parameters decrease to 1 and the lower bound increases to the true risk ratio. We observe that the bias decreases quickly with the number of controlled confounders; e.g. already at five controlled confounders it is <1%. The sensitivity parameters $RR_{EU|c}$ and $RR_{UD|c}$ decrease fairly quickly as well. Notably though, they remain fairly large and the lower bound consequently remains fairly uninformative, until most of the 14 confounders are controlled for. For instance, at 5 controlled confounders, the lower bound is equal to 0.70 and it does not exceed 1 until 11 out of the 14 confounders are controlled.

Discussion and interpretation of the results

In our simulation, we observed that the sensitivity parameters remained fairly large unless most of the confounders were controlled for, which is what Ioannidis *et al.*[5] were concerned about. Yet we also observed that the bias of the observed risk ratio was quite modest, even when no confounders were controlled for, and that the bias quickly became very small when controlling for a few of the most influential confounders, which is what Greenland[8] and MacLehose *et al.*[12] were concerned about.

To understand why the sensitivity parameters may be large even when the confounding bias is small, consider the mathematical definition of the sensitivity parameter $RR_{UD|c}$ given in the 'Sensitivity parameters and lower bound' section. The risk ratio inside the curly brackets contrasts the most extreme 'opposite' confounder levels in the whole space of the uncontrolled confounders, in the sense that these levels maximize and minimize the risk of the outcome. If there are several uncontrolled confounders, or if some of the uncontrolled confounders have more than a few levels, then there may very well exist an extreme (joint) level of the confounders for which the risk of the outcome is very high and another extreme (joint) level for which the risk is very low, which then gives a large risk ratio $RR_{UD|c}$. This may well happen even if these extreme confounder levels are very rare, in which case the actual confounding bias may be quite modest, as in our example. The argument also applies using $RR_{EU|c}$ and the result of combining extreme but rare joint levels for both parameters can exceed large E-values when little confounding is present.

These results have important practical implications. When using DV's sensitivity analysis to construct a lower bound for the true risk ratio, one might consider very large values of the sensitivity parameters as plausible (except perhaps in situations in which one is convinced that the joint relation of the confounders to the exposure or the outcome is small). Thus, in these situations one should not be reassured by a relatively large E-value, since the true

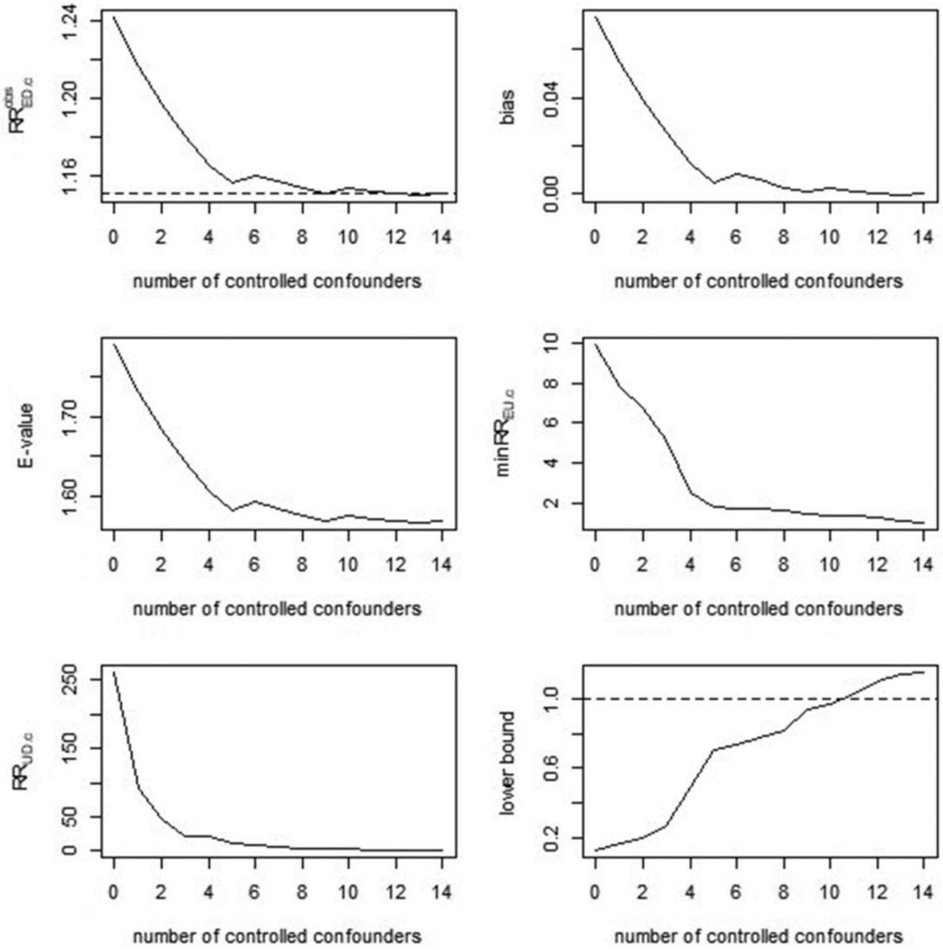

Figure 1 Simulation results. Observed risk ratio (top-left panel; the dashed line indicates the true risk ratio = 1.15), relative bias of the observed risk ratio (top-right panel), E-value (middle-left panel), the sensitivity parameter $RR_{EU|c}$ (middle-right panel) and $RR_{UD|c}$ (bottom-left panel), and the lower bound for the true risk ratio (bottom-right panel) as functions of the number of controlled confounders. The parameters $RR_{UD|c}$ and $RR_{EU|c}$ measure the maximal strength of association that the uncontrolled (set of) confounder(s) *U* may have with the outcome *D* and with the exposure *E*, respectively, conditional on measured confounder levels *c*.

values of the sensitivity parameters may well be even larger. For instance, in most epidemiological studies there could be genetic confounding and there is often a potentially huge number of uncontrolled genotypes. Individually, most of these may have a very small effect on the exposure or the outcome; however, when considering the genome as a whole, there may often exist very rare but extreme combinations of alleles for which the risks of the exposure and outcome are either very small or very large. As a consequence, realistic guesstimates of the sensitivity parameters may be large and so result in a lower bound that is virtually uninformative, whereas the true confounding will be small due to the extreme allele combinations being rare.

This indicates that E-values may need to be interpreted with far more caution than appears to be common practice. In a systematic literature review, Blum *et al.*[3] found 87 papers presenting 516 E-values. They further observed that

> [t]he median E-value was 1.88, 1.82, and 2.02 for the 43, 348, and 125 E-values where confounding was deemed likely by the authors of the papers to affect the results, unlikely to affect the results, or not commented upon, respectively.

The majority of the papers thus considered an E-value of ~1.82 as reassuringly large, which by chance is almost identical to the E-value for our data (= 1.79) with no controlled confounders. Nonetheless, with multiple uncontrolled confounders, the sensitivity parameters can be much higher than 1.82, which is indeed the case for our data. Yet the amount of uncontrolled bias cannot be judged by considering only DV's sensitivity parameters $RR_{EU.c}$ and $RR_{UD.c}$; the joint distribution of all the confounders (controlled and uncontrolled) is crucial and even a very large E-value may correspond to little bias. However, the formal theory of E-values does not reveal this problem, nor does it provide any guidance for judging the plausible magnitude of bias given the amount of controlled confounding and the expected number of uncontrolled confounders.

VanderWeele *et al.*[20] have acknowledged limitations of E-values including the problem of multiple uncontrolled confounders. However, they appeared to reach a somewhat different conclusion than us. They gave a fictitious example in which their sensitivity parameters may plausibly equal 5 and wrote:

> If it is thought plausible that a 5-fold increase in the probability of the outcome could be generated by the unmeasured confounders conditional on the measured covariates, then it is perhaps time to leave that study data alone and pursue other more adequate data sources.

As argued above, we believe that the sensitivity parameters could easily be equal to 5, or much larger, even though the actual confounding bias is small. Thus we do not think that plausible values this large should alone discourage the researcher from pursuing further data analysis, including detailed sensitivity analyses.

As a possible way to deal with large sensitivity parameters, VanderWeele *et al.*[20] proposed considering a hypothetical coarsening of the uncontrolled confounders (e.g. categorization of continuous variables) such that, if control was made for this coarsened set, then confounding bias would not be eliminated, but reduced to an acceptable level of, say, 3%. We agree that such coarsening may likely bring down the size of the sensitivity parameters by an order of magnitude, and may thus be viewed as providing some theoretical justification for the seemingly common practice of being 'reassured' by fairly modest (e.g. ∼1.8) E-values. However, we also fear that a hypothetical coarsening of unmeasured variables may be perceived as rather abstract and that it may thus be very hard for a practitioner to speculate about reasonable values for the sensitivity parameters under such coarsening. Hence, we conjecture that it may be quite unclear what quantitative conclusions can be drawn from such coarsening in practice.

VanderWeele and Mathur[11] also discussed the problem. They wrote:

> The E-value approach, and sensitivity analysis more generally, will be most helpful when there is a single known unmeasured confounder, or when adjustment has been made for all known measured confounders but, of course, with the possibility still of an unknown unmeasured confounder.

Although agreeing with this statement in principle, we believe that the scenarios alluded to in the statement may be very uncommon. In most epidemiological studies there are numerous potential confounders, such as lifestyle factors, socio-economic factors and genetic factors. Even with the 'big data' available today, we cannot realistically hope to fully control all or even most of these factors, especially in light of measurement errors. This does not make observational research futile. By using subject-matter knowledge, one may be able to capture the most important confounders, in which case the obtained estimates may only have little bias, as in our simulation. Nonetheless, DV's sensitivity parameters may still be large, making the E-value rather uninformative.

As argued above, the discrepancy between the DV's bias bounds and the actual bias arises because the parameters in the bound are defined by comparing the most extreme levels of the uncontrolled confounders, without taking the (likely low) prevalence of these extreme confounder levels into account. Indeed, in contrast to many other sensitivity analyses (Rothman *et al.*,[16] Ch. 19), E-value analysis does not require the analyst to make any distributional assumptions about the uncontrolled confounders. Ding and VanderWeele[1] viewed this assumption-free feature as a strong advantage. But as noted by MacLehose *et al.*,[12] it is important to recognize that there is no free lunch: a method that keeps assumptions to a bare minimum may also tend to be too conservative and even uninformative in light of what is likely. This problem may easily go unnoticed for E-value analysis since the extreme distribution used to compute it is invisible in its formulation. Furthermore, what should be considered realistic will vary greatly across contexts. We suspect that these problems will increase as one includes more bias components in the bounding exercise (as in Smith *et al.*[21]).

To illustrate our points we have used real data. These data may not be representative in all possible respects and there may be more or less confounding in other settings. Nonetheless, we think that our data are not atypical insofar as DV's sensitivity parameters are extremely large. As argued above, this feature will likely be present in many other studies—even those in which the confounding bias is modest, as in our example.

In summary, a naïve optimistic analyst may easily underestimate the values of the sensitivity parameters in the E-value and thus get lulled into a false sense of security by the computed E-value. Yet a naïve pessimistic analyst may easily overestimate the amount of confounding bias by overlooking its extreme dependence on the confounder distribution when in fact confounding bias may be quite minor even if the parameters used in the E-value are large and correct. We thus fear that the E-value will often be misleading in one direction or the other and should not be considered a substitute for a contextually well-informed sensitivity analysis.

In closing, we wish to emphasize that we are not against the use of E-values in situations that resemble those depicted by VanderWeele and Mathur[11]: when the concern is with a single known unmeasured potential confounder, the E-value may serve as a simple and useful device to check whether further analyses of its confounding potential are essential, as in that case a large E-value may suffice to move on to other tasks. We also note that E-value developers (e.g. VanderWeele and Mathur[11] and VanderWeele[22]) have, similarly to us, acknowledged the need for contextual reasoning about the relationships among the confounders, exposure and outcome, and also the need for more extensive sensitivity analysis when E-values cannot supply reliable conclusions.

Data availability

The data underlying this article cannot be shared publicly due to the privacy of individuals who participated in the study.

Supplementary data

Supplementary data are available at *IJE* online.

Author contributions

All work for this paper was carried out by A.S. and S.G. jointly, except for the implementation of the analysis, which was done by A.S.

Funding

This work was supported by the Swedish Research Council (2020–01188 to A.S.).

Acknowledgements

The authors thank Paul Gustafson, John Ioannidis, Richard MacLehose, Maya Mathur and Tyler VanderWeele for their helpful comments on the initial manuscript, and Karin Ekström Smedby for kindly sharing the data used in this paper.

Conflict of interest

None declared.

References

1. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology* 2016;7:368–77.
2. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med* 2017;**167**: 268–74.
3. Blum M, Tan Y, Ioannidis J. Use of E-values for addressing confounding in observational studies—an empirical assessment of the literature. *Int J Epidemiol* 2020;**49**:1482–94.
4. Verbeek JH, Whaley P, Morgan RL *et al.* An approach to quantifying the potential importance of residual confounding in systematic reviews of observational studies: a GRADE concept paper. *Environ Int* 2021;157:106868.
5. Ioannidis J, Tan Y, Blum M. Limitations and misinterpretations of E-values for sensitivity analyses of observational studies. *Ann Intern Med* 2019;**170**:108–11.
6. Hamra GB. Re: 'Applying the E value to assess the robustness of epidemiologic fields of inquiry to unmeasured confounding'. *Am J Epidemiol* 2019;**188**:1578–80.
7. Fox MP, Arah OA, Stuart EA. Commentary: the value of E-values and why they are not enough. *Int J Epidemiol* 2020;**49**: 1505–06.
8. Greenland S. Commentary: An argument against E-values for assessing the plausibility that an association could be explained away by residual confounding. *Int J Epidemiol* 2020;**49**:1501–03.
9. Kaufman JS. Commentary: Cynical epidemiology. *Int J Epidemiol* 2020;**49**:1507–08.
10. Poole C. Commentary: Continuing the E-value's post-publication peer review. *Int J Epidemiol* 2020;**49**:1497–500.
11. VanderWeele TJ, Mathur MB. Commentary: Developing best-practice guidelines for the reporting of E-values. *Int J Epidemiol* 2020;**49**:1495–97.
12. MacLehose RF, Ahern TP, Lash TL, Poole C, Greenland S. The importance of making assumptions in bias analysis. *Epidemiology* 2021;**32**:617–24.
13. Gustafson P. To bound or not to bound: is that the question? *Epidemiology* 2021;**32**:635–37.
14. Hu K, Sjölander A, Lu D *et al.* Aspirin and other non-steroidal anti-inflammatory drugs and depression, anxiety, and stress-related disorders following a cancer diagnosis: a nationwide register-based cohort study. *BMC Med* 2020;**18**:238.
15. Wang Z. *Chest: Change-in-Estimate Approach to Assess Confounding Effects. R Package version 0.3.5.* 2021. https://CRAN.R-project.org/package=chest.
16. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*, 3rd edn. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008.
17. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;**567**:305–07.
18. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med Res Methodol* 2020;**20**:244.
19. Cole SR, Edwards JK, Greenland S. Surprise! *Am J Epidemiol* 2021;**190**:191–93.
20. VanderWeele TJ, Ding P, Mathur M. Technical considerations in the use of the E-value. *J Causal Inference* 2019;**7**.
21. Smith LH, Mathur MB, VanderWeele TJ. Multiple-bias sensitivity analysis using bounds. *Epidemiology* 2021;**32**: 625–34.
22. VanderWeele TA. Are Greenland, Ioannidis and Poole opposed to the Cornfield conditions? A defence of the E-value. *Int J Epidemiol* 2022;**51**:364–71.