International Journal of Epidemiology, 2020, 1–13
doi: 10.1093/ije/dyz261
Original article

Original article

Use of E-values for addressing confounding in observational studies—an empirical assessment of the literature

Manuel R Blum , [1,2,3†] Yuan Jin Tan [1,3†] and John P A Ioannidis [1,3,4,5,6]

[1]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA, [2]Department of General Internal Medicine, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland, [3]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA, [4]Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA, [5]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA and [6]Department of Statistics, Stanford University School of Humanities and Science, Stanford, CA, USA

Corresponding author. Stanford Prevention Research Center, 1265 Welch Rd, Medical School Office Building, Room X306, Stanford CA 94305, USA. E-mail: jioannid@stanford.edu

[†]These authors contributed equally.

Editorial decision 25 November 2019; Accepted 6 December 2019

Abstract

Background: E-values are a recently introduced approach to evaluate confounding in observational studies. We aimed to empirically assess the current use of E-values in published literature.

Methods: We conducted a systematic literature search for all publications, published up till the end of 2018, which cited at least one of two inceptive E-value papers and presented E-values for original data. For these case publications we identified control publications, matched by journal and issue, where the authors had not calculated E-values.

Results: In total, 87 papers presented 516 E-values. Of the 87 papers, 14 concluded that residual confounding likely threatens at least some of the main conclusions. Seven of these 14 named potential uncontrolled confounders. 19 of 87 papers related E-value magnitudes to expected strengths of field-specific confounders. The median E-value was 1.88, 1.82, and 2.02 for the 43, 348, and 125 E-values where confounding was felt likely to affect the results, unlikely to affect the results, or not commented upon, respectively. The 69 case-control publication pairs dealt with effect sizes of similar magnitude. Of 69 control publications, 52 did not comment on unmeasured confounding and 44/69 case publications concluded that confounding was unlikely to affect study conclusions.

Conclusions: Few papers using E-values conclude that confounding threatens their results, and their E-values overlap in magnitude with those of papers acknowledging susceptibility to confounding. Facile automation in calculating E-values may compound the already poor handling of confounding. E-values should not be a substitute for careful consideration of potential sources of unmeasured confounding. If used, they should be interpreted in the context of expected confounding in specific fields.

Key words: E-value, confounding, literature review, observational study, sensitivity analysis

Key Messages

- The E-value is a new standardized approach for sensitivity analyses on confounding in observational studies, but it presents conceptual and validity problems.
- Our empirical analysis shows that the vast majority of authors concluded, based on E-values, that uncontrolled confounding was unlikely to be a major concern.
- There is no clear demarcation in magnitude between E-values used to acknowledge susceptibility to confounding and those that do not.
- Facile automation in calculating E-values may compound the already poor handling of confounding.

Introduction

Confounding due to unmeasured or uncontrolled confounders creates serious problems for the interpretation of results from observational studies.[1,2] Many approaches have been developed to address the issue of confounding in observational studies through sensitivity analyses.[3–6] One recently proposed approach by VanderWeele and Ding introduced the E-value, the minimum magnitude of association that an unmeasured confounder needs to have with both the exposure and the outcome to fully explain away the observed exposure-outcome association, conditional on the measured covariates.[7,8]

However, in a recent conceptual paper published in *Annals of Internal Medicine*, we have outlined several major conceptual problems of E-values.[9] E-values are based on often highly speculative assumptions, such as assuming that exposure prevalence and confounder prevalence are at the point that maximizes confounding. Also, E-values have a monotonic relationship with effect estimates[10] and therefore provide little additional information. Importantly, no guidance exists on what range of E-values is acceptable to stop worrying about residual confounding.

We are concerned that E-values may be used in the scientific literature as a misleading alternative to critical appraisal and careful thinking about the handling and reporting of confounding in observational data, but empirical data on the use of E-values are lacking. The purpose of the current study was therefore to make an empirical assessment of the current use of E-values in scientific publications.

Methods

Identification and selection of studies

We used Google Scholar to identify all publications that were published until the end of 2018 and that cited at least one of the two original publications that first introduced the concept of E-values.[7,8] The search was performed with no restrictions on 24 July 2019.

Data collection

Two reviewers (Y.J.T., M.R.B.) extracted data from each identified study's data on study characteristics (study type, protocol registration as reported by authors), E-values [number and individual values of E-values for effect estimates and for lower/upper confidence limits (CLs) of effect-measure estimates (in particular the CL closer to the null), verbatim language of E-value use (in the Abstract, Methods, Results and Discussion, respectively)], any other types of sensitivity analyses performed to assess confounding and any mentioning of specific uncontrolled or only partially controlled confounders. Studies were classified into different scientific fields according to the Web of Science Categories (studies could be classified into multiple categories). Supplementary Table 1, available as Supplementary data at *IJE* online, contains all extracted data.

Synthesis and appraisal

We calculated E-values for lower or upper CLs of effect estimates (the CL closest to the null) where authors did not provide them in the publications, using the web-based tool provided at [https://www.evalue-calculator.com]. We did not calculate E-values for papers not reporting any E-value, nor for papers reporting the joint bounding factor through large tables as proposed by Ding and VanderWeele.[8] We categorized studies according to their author-defined, E-value-based conclusion about the threat of residual confounding from uncontrolled variables to the main results as either 'unlikely to affect', 'likely to affect',

or 'no comment'. Where studies presented multiple E-values with differing conclusions, categorizing was done at the individual E-value level. We categorized the conclusions of additional sensitivity analysis results as 'supporting main conclusions', 'not supporting main conclusions', or 'no comment'.

Matched control studies

We compared effect sizes obtained from papers within our review ('case publications') with effect sizes from a matched set of control publications that had been published in the same journal and issue but where the authors had not reported any E-values. Effect sizes were extracted from the abstract of these control studies. The matched analysis was restricted to case publications which reported at least one odds ratio, risk ratio, or hazard ratio in their abstracts for which an E-value had been calculated anywhere in the paper. For each case, we found a control publication that also had at least one odds ratio, risk ratio, or hazard ratio in its abstract but had not calculated any E-value anywhere in the paper. There were a total of 69 case-control pairs. The control publications were selected by going forwards (i.e. in the next pages in the same issue) within issues containing case publications. If no control publications were found that reported the above effect measures in their abstracts, we searched backwards (i.e. in the preceding pages within the same issue). If no control publications were found within the same issue, the subsequent issue was searched, until an eligible control publication could be identified. If effect sizes were less than one, we used their inverse in this comparison. Whenever there were multiple eligible effect sizes reported in an abstract, the median effect sizes were calculated. Comparison between the case and control publications for the log median effect sizes used a paired t test. We also assessed whether case versus control publications stated that their results were likely to be affected by confounding and whether they mentioned any specific uncontrolled potential confounders that were felt to threaten at least some of the main conclusions. Supplementary Table 2, available as Supplementary data at *IJE* online, contains extracted data used in the case-control comparison along with references of included case and control publications.

Data cleaning, analysis and plotting

The following R packages were used for data cleaning and analysis: *ggplot2*,[11] *tidyr*[12] and *gridExtra*.[13] Publish or Perish was used to extract titles and publication information from Google Scholar.[14] Datasets and scripts used in this project have been deposited online on the Open Science Framework at [https://osf.io/4vxph/?view_only=472957acd3af4687b3702889df5e0776].

Results

We identified 159 articles that cited one of the original papers introducing E-values. We excluded two articles to which we were unable to obtain access. Of the remaining 157 articles, 70 did not apply E-values to specific data. Of the 87 articles that did apply E-values to specific data, 33 reported E-values only for point effect estimates and 50 also included calculations for a CL, always for the 95% confidence interval. Four studies presented only E-values for the CL and not the point estimate. We therefore calculated the E-values for the 95% CL of the point estimates in these 33 articles. A total of 516 E-values for the point estimate were presented in the 87 papers.

Additional key study characteristics are shown in Table 1. The majority of studies using E-values were cohort studies, and the majority of E-values were calculated from risk ratios. Four were pre-registered. Overall, 51 papers concluded that residual confounding of a magnitude indicated by the respective E-values in those publications was unlikely, 11 concluded that residual confounding could affect the results, three papers concluded that some associations could be affected by confounding and others were unlikely to be affected and 22 made no comment. The exact relevant phrasing is shown in Table 2.

Of 87 articles reporting E-values, 34 named specific variables that could be confounders but were not accounted for in the analysis (either not accounted for at all or accounted for in some suboptimal form that could still leave residual confounding). One such variable was named in nine articles, two were named in eight articles and more than two were named in 17 articles. The maximum number of potential confounders named was six. Of these 34 articles, 18 concluded that residual confounding from the known but uncontrolled variables was unlikely to affect the main results, two articles made this conclusion for parts of the main results, five articles concluded that residual confounding from these variables was likely to affect the main results and nine articles did not comment on this. Articles that concluded that residual confounding was likely (based on the E-value) did not differ in their appraisal of the impact of specific residual confounders versus articles that concluded that residual confounding was unlikely (Table 3).

19 of the 89 studies (21.3%) related the magnitude of the E-value to the expected strength of confounders in the field, to determine if a known but unmeasured confounder was likely to pose a threat to validity. Of these, 15 articles concluded that it was unlikely for confounders equivalent

Table 1. Key characteristics of studies reporting E-values

Number of eligible studies	87
Number of E-values reported/study, median (IQR)	2 (1.0–3.0)
Number of registered studies, *n* (%)	4 (5)
Types of outcome used for E-value calculation, *n* (%)	
Risk ratio	277 (53)
Odds ratio	92 (18)
Hazard ratio	107 (21)
Standardized mean difference	6 (1)
Risk difference	1 (0.2)
Linear regression coefficient	33 (6)
Study types, *n* (%)	
Case-control	4 (5)
Cohort	67 (77)
Commentary	7 (8)
Cross-sectional	4 (5)
Meta-analysis	2 (2)
Nested case-control	1 (1)
Randomized controlled trial	2 (2)
Common fields of study, *n* (%)[a]	
Medicine, general & internal	28 (32)
Public, environmental & occupational health	15 (17)
Paediatrics	7 (8)
Infectious diseases	6 (7)
Oncology	6 (7)
Obstetrics & gynaecology	5 (6)
Rheumatology	4 (5)
Cardiac & cardiovascular systems	3 (3)
Endocrinology & metabolism	3 (3)
None	3 (3)
Pharmacology & pharmacy	3 (3)
Psychiatry	3 (3)
Respiratory system	3 (3)
Substance abuse	3 (3)

[a]Fields of study were derived from the Web of Science Categories. Each study may be classified under multiple such categories, not shown are 21 categories with 1–2 studies each (total of 27).

or greater in magnitude than the observed E-values to exist, and that residual confounding was unlikely to be a threat. Three articles concluded that the magnitude of the E-value was not implausibly large and residual confounding could pose a threat to validity. The remaining study concluded that confounding was a threat for a portion of the main results. The exact phrasing used is shown in Table 4.

Figure 1 shows the range of E-values for point effect estimates and of E-values for the 95% CL closest to the null for the claims listed in Table 2. As shown, there is no clear demarcation of what magnitude of E-values is large enough to herald protection from confounding. For example, a population-based study investigating perinatal infant infections in women with and without systemic lupus erythematosus, with E-values of 1.9 to 6.1 for the effect size point estimates and 1.0 to 3.0 for the 95% CL, claimed that confounding may still affect the results, specifically citing E-values such as 2.1 and 1.9.[15] However, another study, investigating the effect of interpregnancy intervals on adverse perinatal outcomes, specifically cited E-values of 1.11 to 3.5 for the effect size point estimates and 1.0 to 3.06 for the 95% CL to support the claim that confounding may not affect the results.[16] In all cases where confounding was deemed likely to affect the results, the E-value was <4.0 for the 95% CL. However, only 10% (43 of 440) of all E-values <4.0 for the 95% CL were translated as 'confounding likely to affect the results', whereas 71% were translated as 'confounding unlikely to affect the results' and for the other 19%, there was no comment made. For the 239 E-values <1.5 for the 95% CL, the proportions were very similar, i.e. 11%, 71%, and 18%, respectively.

The median point estimate E-value was 1.88 [interquartile range (IQR), 1.48 to 2.52], 1.82 (1.4 to 2.71) and 2.02 (1.46 to 3.68), for the 43, 348, and 125 E-values where confounding was deemed likely to affect, unlikely to affect the results, or not commented upon, respectively. The median E-value for the 95% confidence interval was 1.28 (1.02 to 1.74), 1.49 (1 to 2.04), and 1.65 (1.16 to 3.03), respectively; 355 of 478 E-values were calculated for statistically significant point estimates, representing 74.3% of point estimates for which statistical significance was clearly reported.

Of 87 articles with reported E-values, 73 addressed binary exposures and 14 had continuous exposures. Of the 14 studies focusing on continuous exposures, three studies expressed the E-values on a per standard deviation scale, one expressed E-values per 10-unit increase in exposure and 10 compared only the highest and lowest quantiles. The vast majority of the E-values were calculated for ratio measures, with only seven (1%) of the E-values calculated for measures such as risk difference and standardized mean difference.

Of 87 articles reporting E-values, 49 presented the results of at least one other sensitivity analysis (range, 1 to 5). A total of 39 sensitivity analyses varied the analysis model assumptions or the analytical methods, 22 restricted the study sample, 10 varied the exposure definition, and five varied the outcome definition (Supplementary Table 1, available as Supplementary data at *IJE* online). Of these 49 articles, 33 claimed that E-values showed results unlikely to be influenced by uncontrolled confounders. Of the remaining 16 articles, two claimed that E-values showed results likely to be influenced by uncontrolled confounders, 13 made no comment, and one presented diverging conclusions for different E-values. Of the 49 articles presenting at least one other sensitivity analysis, 46 reported that the additional sensitivity analyses supported the main conclusion

Table 2. E-values used to conclude that confounding is unlikely or likely to affect the results

Authors' conclusion on validity threat[a]	Ref. (study)	Wording
Unlikely	S8	. . . the evidence of association seems reasonably strong because substantial unmeasured confounding would be needed to explain away the observed association
	S9	. . . our sensitivity analyses suggest that substantial confounding is highly unlikely
	S11	These results demonstrate that substantial unmeasured confounding would be needed to reduce the observed associations to null
	S13	Our sensitivity analyses suggested that an unmeasured confounder would need to be large . . . to explain away the observed associations
	S18	To our best knowledge we are unaware that there is such a strong confounder between magnesium and the prevalence of chondrocalcinosis
	S21	. . . sensitivity analyses revealed that it would take very strong confounding to negate the associations observed in this study
	S22	. . . we found our result was somehow robust . . . to potential unmeasured confounder(s)
	S24	We are not aware of any potential confounder of this strength that is not already included in the model
	S25	. . . relatively substantial residual unmeasured confounding was needed to explain away the observed significant associations
	S26	. . . our E-value sensitivity analysis suggested that any unmeasured factor would need to be exceptionally strongly correlated . . . to explain away our study findings
	S27	A putative unmeasured or unknown confounder or a set of confounders . . . would have to exhibit a very strong association . . . to explain away the observed association
	S31	. . . relatively influential unmeasured confounders would be needed to negate the observed associations
	S32	This is unlikely to be the case, . . . suggesting that residual confounding has been minimized and that the results are unlikely to be biased
	S37	. . . sensitivity analyses suggest that this is unlikely to explain the study findings
	S38	. . . unlikely that there are any confounders with sufficient magnitude to explain away the mostly high HRs presented in this study
	S41	Such substantial confounding by unmeasured factors seems unlikely. . .
	S43	. . . it is not likely that our findings could be completely attributable to residual or unmeasured confounding
	S44	. . . unmeasured confounding of considerable strength would be needed to fully explain the observed associations
	S48	. . . our findings are robust to potential unmeasured confounders
	S51	. . . while residual confounding is possible, it is unlikely to explain the entire association
	S53	. . . this unmeasured variable would need to be moderate in size. . .
	S54	. . . an effect estimate of a level of 1.6, which is considered quite high. . .
	S67	. . . provides some reassurance that even a modest RR of 1.2 is relatively robust to unmeasured confounding
	S72	. . . such residual confounder effect needs to be quite strongly associated with exposure or outcome
	S81	. . . sensitivity analyses suggested that these and other potential confounding factors would have to be strong to explain the observed association
	S84	. . . would have to increase both the likelihood of being in the uppermost quartile for either sedentary characteristic and the risk for all-cause mortality by 2.0- to 3.0-fold above the measured covariates. This would constitute substantial confounding
	S105	Such substantial confounding by unmeasured factors seems unlikely
	S106	For an unmeasured confounder to fully explain away the association . . . it would have to both increase the likelihood of service attendance and decrease the likelihood of depression by 2.1-fold, above and beyond the measured covariates, which may not be likely
	S108	Analyses estimating the effect of such unmeasured confounders revealed that each of the confounder–interval CRC and confounder–race/ethnicity associations would need to exceed 1.50 to substantially alter our main findings, a level not observed in most studies of overall CRC risk

(Continued)

Table 2. Continued

Authors' conclusion on validity threat[a]	Ref. (study)	Wording
	S119	Such substantial confounding by unmeasured factors seems unlikely
	S123	Sensitivity analyses indicate that the most important observed associations . . . are probably not a result of any potential effect of unmeasured confounders
	S124	The only measured confounder of similar magnitude was prepregnancy obesity, so it is unlikely that our associations could be explained away by an unmeasured confounder
	S125	For our study, we believe that this degree of confounding remains unmeasured is implausible and, as such, do not believe that our conclusions would change
	S126	Given that this risk ratio is much greater than any observed for known macrovascular disease risk factors examined in the current study . . . it is implausible that an unmeasured confounder exists that can overcome the effect of bariatric surgery observed in the current analysis study
	S127	Substantial confounding would be required to fully explain the observed association
	S128	Residual confounding likely persists, but sensitivity analyses suggest that this is unlikely to explain the study findings
	S132	Using the E-values, we found that an unmeasured confounder needs to be associated with both diet and asthma by a risk ratio of roughly 2.0 (at least) to explain away the association, which we believe is unlikely
	S134	For an unmeasured characteristic to render the described association . . . nonsignificant, it would have to show an adjusted risk ratio of at least 1.82 . . .. while we cannot rule out the presence of such an unmeasured confounder, of the measured patient characteristics included in our multivariable model, no covariate satisfied this requirement
	S135	. . . for unmeasured confounding to change our results such that allopurinol's true effect is harmful would be unlikely (E value, 2.84)
	S137	. . . an unmeasured confounder would need to be associated with allergy and complicated appendicitis by approximately 2.9-fold for the OR (and approximately 1.9-fold for the 95% CI) above and beyond the measured variables to explain the observed effect of allergy. Socioeconomic variables were not available, and information about primary care visits were introduced in the countywide electronic medical records only from 2015 onward, corresponding to approximately one-sixth of the patients; however, we believe that the inclusion of self-reported duration of symptoms in the multivariable model compensates well for this limitation
	S138	There was moderate evidence suggesting the associations of forgiveness with psychosocial well-being and mental health outcomes were likely robust to unmeasured confounding. . .. similarly strong unmeasured confounding between forgiveness and other psychological and mental health outcomes would be needed to explain away the observed associations, suggesting that these associations are somewhat robust to unmeasured confounding
	S140	The high E-values from the sensitivity analysis suggest that it is unlikely that unobserved confounders would nullify the conclusions for the high-risk patients
	S143	. . . our findings are unlikely to be fully explained by unobserved confounding given that only a variable with a strong association with both water insecurity and probable depression could completely explain away the estimated association
	S146	Results from the sensitivity analysis also suggest that a number of the observed associations are relatively robust to potential unmeasured confounding
	S147	However, we do not know of any genetic factor(s) with an RR $\geq$ 1.25; thus it is unlikely there is a confounding factor of this size or larger
	S149	Notably, the high E-value validates the strength and robustness of the observed findings to the presence of an unmeasured cofounder
	S152	The E-value indicated that our treatment effect has moderate robustness to unmeasured confounders
	S155	. . . an unmeasured confounder associated with both PCB serum levels and hypertension onset could bias the estimate substantially with a risk ratio of 4.25 or higher for each of them. To move the confidence interval to include the null, an unmeasured confounder should be associated with PCB exposure and hypertension with a risk ratio of 1.92. The magnitude of this effect seems however to be higher than that commonly found for dietary patterns that could be related to both PCB serum levels and hypertension

(Continued)

Table 2. Continued

Authors' conclusion on validity threat[a]	Ref. (study)	Wording
	S157	Such a confounder would need to be associated with our intervention and mortality by the same relative risk of ≥1.85, and not lower . . . we cannot identify such a strong unmeasured factor or a differential exposure that might account for the observed association
	S158	. . . the strength of association on the risk ratio scale that an unmeasured confounder would need to have with both early retention and mortality, conditional on the measured covariates . . . would need to be at least 3.7 (for the point estimate to be 1) or 2.9 (for the upper limit of the 95% CI to include 1); these are substantial, and suggest that early retention does in fact decrease the risk of mortality
	S159	. . . it is unlikely that these variables would have an effect on cancer risks strong enough to explain away the observed association
Likely	S16	Unmeasured frailty might also explain the observed HR of 0.86 for mortality, if frailty were independently associated with 1.46-fold increased risks of having a heart rate <70 beats/min and mortality
	S17	Other factors . . . could be responsible for the observed mediation effect . . . given how sensitive these results were to unmeasured confounding
	S23	It is quite possible that such an unmeasured confounder exists, which is a limitation of our study
	S33	. . . even minor uncontrolled risk and protective factors . . . may, in combination, bias. . . findings
	S40	Relatively weak unmeasured confounding could overturn the exposure-outcome association
	S47	An unmeasured confounder could move the lower confidence bounds on these associations below the null
	S52	Such level of confounding is not uncommonly large and would render the results no longer significant
	S141	An unmeasured confounder associated with both tamsulosin exposure and development of dementia by an RR of 1.62 does not seem implausible. Furthermore, the confidence interval around the HR . . . could be moved to include the null by an unmeasured confounder that was associated with both the treatment and the outcome by an RR of 1.54-fold each . . . again, such an unmeasured confounder does not seem implausible
	S144	However, given the abovementioned RCTs and that our result was sensitive to potential unmeasured confounders E-factor 1.74, we feel that our results should not be interpreted as conclusive and that further modern RCTs are needed
	S154	We . . . observed that even modest unmeasured confounding could be an alternate explanation for the apparent comparability of 8- and 12-week regimens not only in our study but also prior studies
	S156	The E-values indicate that a fairly small effect size of the unmeasured confounders might be sufficient to explain the reported associations
Both	S5	. . . not likely to represent a threat to the adjusted models; although the assessment of unmeasured confounding for the RR model did point to some imprecision in the confidence intervals for the intervention-outcome associated identified
	S14	The possibility for a confounding as strong as 21.7 is very low as the effect sizes of 3-fold or more are not particularly common in biomedical and social sciences research . . . however, the E-value of 1.4 for the CI indicates that substantial confounder associations with caffeine intake and oral clefts could potentially move the CI to include 1
	S39	Sensitivity analysis showed relatively high E-values . . . suggesting that these associations are unlikely due to unmeasured confounders. Conversely, the E-values for ICU were lower, suggesting that the observed associations could be ruled out by an unmeasured confounder
No comment	S1, S6, S19, S42, S45, S46, S50, S55, S104, S117, S129, S130, S131, S133, S136, S139, S142, S145, S148, S150, S151, S153	

Comment: Only explicit statements are included in this table.

[a]This represents the original authors' E-value-based conclusion about the validity threat from uncontrolled confounding.

and one article made no comment; 21 of the 49 studies employing other sensitivity analyses also mentioned additional, uncontrolled, likely confounders. A cross-tabulation of E-value-based conclusions and additional sensitivity analysis-based conclusions is also presented in Table 3.

Median effect sizes were similar between those reported in studies included in our review and those calculated from matched control studies where the authors had not reported any E-values. The median effect sizes were larger in the case publication than the control in 31 pairs and

Table 3. Cross-tabulation of E-value-based conclusions and additional sensitivity analyses-based conclusions

Conclusion based on E-values	Number of articles	Number of uncontrolled potential confounders mentioned				Number of articles	Conclusion from other sensitivity analyses		
		None	1	2	>2		Supports main conclusion	Does not support main conclusion[a]	No comment
Unlikely to affect[b]	51	33	4	4	10	33	33	0	0
Likely to affect[b]	11	6	2	2	1	2	2	0	0
Both	3	1	0	1	1	1	1	0	0
No comment	22	13	3	1	5	13	10	1	2

[a]Where the trend observed was nullified in at least one of the results undergoing sensitivity analyses.

[b]The proportions of papers with 0, 1, 2 uncontrolled confounders mentioned in these two groups were compared using a two-sided Cochrane Armitage chi squared test for trend. Z = -1.2796, dim = 3, *P*-value = 0.200.

smaller in the case publication than the control in 38 pairs. The median of median effect size was 1.68 in the control group versus 1.61 in the case group (*P*-value for the two-tailed paired t test = 0.13, Figure 2).

In the 69 matched pairs, 44 case publications stated that confounding is unlikely to affect the results, five stated that confounding is likely to affect the results, two had mixed statements and 18 made no comment. Conversely, among control publications, only three stated that confounding was unlikely to affect the results, 14 stated that confounding was likely to affect the results and 52 made no comment. The cross-tabulation of pairs in their claims about confounding is shown in Table 5. Uncontrolled potential confounders that were felt to threaten at least some of the main conclusions were listed in only 26 case studies and 16 control studies.

Discussion

In this study, we sought to make an empirical assessment of the current use of E-values in the scientific literature since they were first introduced in 2016.[8]

We observed very large overlap in the E-value ranges between studies that deemed residual confounding to be likely versus unlikely to threaten the main conclusions. This demonstrates that there is no clear guidance on what constitutes a large enough E-value to stop being concerned about potential threat of residual confounding. The original paper introducing the E-value does not provide ranges of E-values that are deemed large or small, but mentions that 'the investigator ... merely reports how strongly an unmeasured confounder must be related to the treatment and outcome to explain away an effect estimate; readers or other researchers may then assess whether the confounder associations of that magnitude are plausible'.[7] However, we argue that it is practically impossible for readers of the research articles or other researchers to tell with certainty what is a large, modest or small E-value. We thus recommend that authors who still decide to use E-values should provide guidance on their interpretation, relating their magnitude to expected magnitudes of known but uncontrolled confounders in the specific field of study. Even in the case of effect sizes, where we have accumulated extensive, century-long experience of what (field-specific) effect sizes look like—e.g. typically effects of 0.2, 0.5 and 0.8 on a standardized scale are classically considered small, modest, and large, respectively[17]—full consensus on what constitutes a large effect may still be difficult to reach. For example, the GRADE tool considers a relative risk of 3 (which would correspond to an E-value of 5.5) as the cut-off for a large effect,[18] whereas others have proposed even a relative risk of 10 or 12 (which would correspond to an E-value of 19.5 and 23.5) for situations where observational data suffice and randomized trials are unnecessary and impossible to do.[19,20] Importantly, sometimes residual confounding cannot be excluded even for effects sizes that are deemed to be large (thus also for E-values that are as large), although, in other cases, residual confounding may not be important even for effect sizes that are small (thus also for E-values that are small). Authors who decide to use E-values, should pre-specify and explain at a minimum what is the context for their interpretation.

We worry that E-values may be used to dismiss the threat of residual confounding without sufficient and robust consideration. Indeed, few studies using E-values concluded that residual confounding could influence their conclusions. Moreover, the majority of studies included in this review did not discuss any specific confounders, and very few studies related the magnitude of the E-values to the expected strength of known confounders in the field. Whereas most epidemiological studies do not fully discuss

Table 4. Studies relating the E-value magnitude to the expected strength of confounders in the field

Authors' conclusion on validity threat[a]	Ref. (study)	Wording
Unlikely	S18	To our best knowledge we are unaware that there is such a strong confounder between magnesium and the prevalence of chondrocalcinosis
	S24	As in every prospective study, the possibility of residual confounding cannot be excluded despite careful adjustments for important available covariates. However, such confounding needs to be associated with the exposure metabolite and the outcome, to an extent, as described by the calculated E-values for the respective metabolites. We are not aware of any potential confounder of this strength that is not already included in the model
	S26	First, we considered and adjusted for a number of sociodemographic, psychosocial and clinical factors. Second, our E-value sensitivity analysis suggested that any unmeasured factor would need to be exceptionally strongly correlated with both CBA and severe food insecurity at 12 months to explain away our study findings. Last, we know of no geographical trends, events or differences that would have resulted in such drastic reductions in reported household food insecurity in the CBA group but not the clinic-based group
	S27	Given that the analyses already accounted for known confounders and employed an active comparator group, to cancel the results, any uncontrolled confounder would also have to be independent of the confounders already adjusted for and is unlikely to exist with sufficient strength required to nullify the association of PPI and risk of eGFR less than 60 ml/min/1.73 m^2
	S32	The strength of association between unmeasured or residual confounding with exposure and outcome would have be greater than 2 beyond the model for the result of 1.38 to be untrue. This is unlikely to be the case and further, the risk estimates for diclofenac and naproxen are similar to prior studies suggesting that residual confounding has been minimized and that the results are unlikely to be biased
	S51	While it is conceivable that unmeasured dietary factors are associated with seafood intake by a risk ratio of >2.13-fold, we are unaware of any studies linking any specific dietary or lifestyle factors to fecundity by a risk ratio (or FOR) >2.13-fold. For context, the FOR comparing women <27y to ≥35y in this cohort was 1.96. Thus, while residual confounding is possible, it is unlikely to explain the entire association
	S72	Although we adjusted for important covariates as proxy for socioeconomic status, there may be an influence of residual or unmeasured confounding. However, such residual confounder effect needs to be quite strongly associated with exposure or outcome (RR ∼2.30). This is unlikely given that known socioeconomic factors, such as education level, are only weakly associated with hip fractures
	S81	If both sources of confounding were equally strong, the required RR for each would have to be between 1.7 and 1.8. We cannot point to such an unmeasured factor
	S108	Analyses estimating the effect of such unmeasured confounders revealed that each of the confounder–interval CRC and confounder–race/ethnicity associations would need to exceed 1.50 to substantially alter our main findings, a level not observed in most studies of overall CRC risk
	S119	In sensitivity analysis, for an unmeasured confounder to explain the effect of religious service attendance on suicide, it would have to both increase the likelihood of religious service attendance and decrease the likelihood of suicide by greater than 10-fold above and beyond the measured covariates. Such substantial confounding by unmeasured factors seems unlikely
	S126	The sensitivity analysis … indicated that the observed 5-year HR … could only be explained by an unmeasured confounder that was associated with both receipt of bariatric surgery and risk of macrovascular disease by a risk ratio of more than 2.72 … given that this risk ratio is much greater than any observed for known macrovascular disease risk factors examined in the current study, such as hypertension, diabetes, or hyperlipidaemia, it is implausible that an unmeasured confounder exists that can overcome the effect of bariatric surgery observed in the current analysis study
	S134	For an unmeasured characteristic to render the described association between statin continuation and mortality nonsignificant, it would have to show an adjusted risk ratio of at least 1.82 (or 1.63 to explain away the lower confidence limit) with both the exposure and the outcome separately. While we cannot rule out the presence of such an unmeasured confounder, of the measured patient characteristics included in our multivariable model, no covariate satisfied this requirement

(Continued)

Table 4. Continued

Authors' conclusion on validity threat[a]	Ref. (study)	Wording
	S147	The direct and indirect effect of sex on incident knee OA, for example through tibia mode 2, were 1.56 and 0.96, and the corresponding E-values were 2.49 and 1.25, respectively. However, we do not know of any genetic factor(s) with an RR $\geq$ 1.25; thus it is unlikely there is a confounding factor of this size or larger
	S155	An unmeasured confounder associated with both PCB serum levels and hypertension onset could bias the estimate substantially with a risk ratio of 4.25 or higher for each of them. To move the confidence interval to include the null, an unmeasured confounder should be associated with PCB exposure and hypertension with a risk ratio of 1.92. The magnitude of this effect seems however to be higher than that commonly found for dietary patterns that could be related to both PCB serum levels and hypertension
	S157	Such a confounder would need to be associated with our intervention and mortality by the same relative risk of $\geq$1.85, and not lower. A confounder with less strong associations could not explain away this mortality reduction. We cannot identify such a strong unmeasured factor or a differential exposure that might account for the observed association
Likely	S23	In the primary analysis, an unmeasured confounder with an association with both exposure and outcome of at least 1.41 (for the point estimate to be 1) or 1.17 (for the upper limit of the 95% CI to include 1) would be needed to explain away the observed association. It is quite possible that such an unmeasured confounder exists, which is a limitation of our study
	S52	Thus, the observed HR of 0.75 could be explained by a confounder associated with both canakinumab and the primary outcome that has a risk ratio of 1.63 or above. Such level of confounding is not uncommonly large and would render the results no longer significant
	S141	An unmeasured confounder associated with both tamsulosin exposure and development of dementia by an RR of 1.62 does not seem implausible. Furthermore, the confidence interval around the HR point estimate of 1.17 could be moved to include the null by an unmeasured confounder that was associated with both the treatment and the outcome by an RR of 1.54-fold each, above and beyond the measured confounders. Again, such an unmeasured confounder does not seem implausible
Both	S14	The possibility for a confounding as strong as 21.7 is very low as the effect sizes of 3-fold or more are not particularly common in biomedical and social sciences research. For caffeine consumption, the E-value of 11.3 for the estimate is quite robust; however, the E-value of 1.4 for the CI indicates that substantial confounder associations with caffeine intake and oral clefts could potentially move the CI to include 1

[a]This represents the original authors' E-value based conclusion about the validity threat from uncontrolled confounding.

the implications of confounding anyhow,[21] we argue that calculating an E-value provides no additional insight unless careful and thorough consideration is given to potential uncontrolled confounders and their expected strength, as well as on how to handle these confounders. Our matched assessment of control publications where the authors had not used E-values shows that articles which presented E-values dealt with similar effect sizes (and thus would calculate similar E-values) as control articles. However, articles using E-values were far more likely to conclude that confounding is unlikely to affect the study inferences. Almost two-thirds of the articles using E-values reached this conclusion, whereas this was rarely seen in control publications where E-values were not used. The large majority of control publications did not comment at all on the potential for confounding to influence the study conclusions, but among the publications that did comment, authors were more likely to conclude that unmeasured confounding could affect the study inferences. Authors who used E-values might be more cognizant of the need to address unmeasured confounding in observational studies; however, the disparity in conclusions drawn about this unmeasured confounding is striking. It is possible that authors may view the calculation of E-values, regardless of magnitude, as an alibi to dismiss the threat of confounding.

The majority of the studies that we assessed did not consider that multiple uncontrolled confounders could exist, which in combination might pose credible threat to the validity of the results even when no single uncontrolled confounder comes close to the E-value magnitude. Also, even the studies that did list multiple uncontrolled confounders did not mention whether combined confounding from these factors could reach the E-value's magnitude to nullify the found associations.

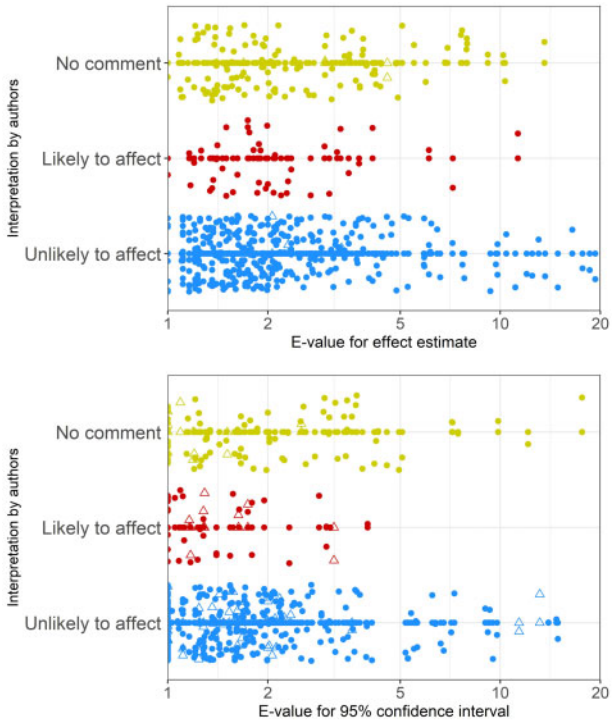

Figure 1. E-values for the point estimates of effect sizes and for the 95% CL closest to the null. When the E-value was not calculated in the original paper either for the effect size or the 95% CL, we have calculated it ourselves (shown with triangles instead of circles). The E-values are plotted on a natural log scale for increased readability at values close to 1. Colour is blue for studies that claim that confounding is unlikely to affect the results ($n = 348$), red for those that claim that confounding may affect the results ($n = 43$) and yellow for those that do not comment ($n = 125$). E-values larger than 20 were omitted for readability (two E-values for point estimates and one E-value for CL were omitted for categories 'No comment' and 'Unlikely to affect', each).

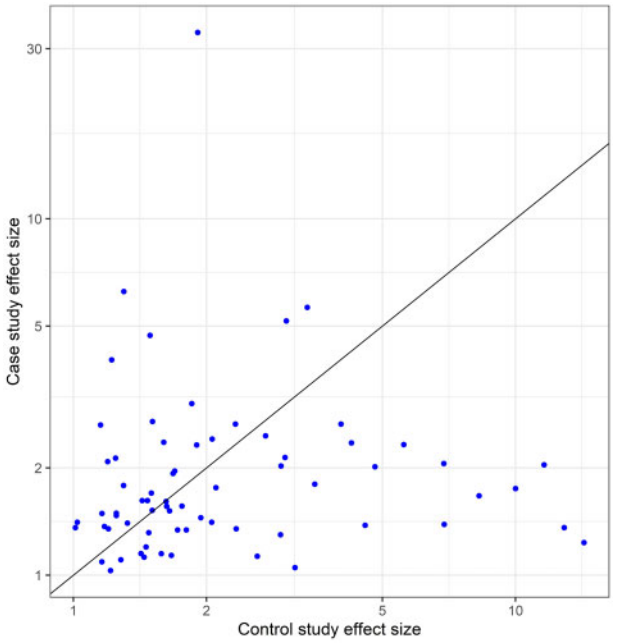

Figure 2. Median effect sizes (for relative risks presented in the abstract of a paper) are plotted for 69 matched case-control pairs, with the case publication effect size on the y axis and the control publication effect size on the x axis. The effect sizes are plotted on a natural log scale. The diagonal line indicates the points at which case study effect size is equivalent to control study effect size. Effect sizes less than 1 have been inverted. The natural log median effect sizes for each group were compared using a matched two-tailed t test, Mean of differences = 0.15 (95% CI -0.04–0.35), df = 68, *P*-value = 0.13.

On the topic of interpretation of E-values, the original paper proposing E-values advocated that 'reporting the E-value for the limit of the CI closest to the null is good practice'.[7] The authors contended that this approach helped address the statistical uncertainty around the observed estimate. In our sample, a sizeable proportion of studies only calculated E-values for point estimates and not CLs. In the broader context, however, this proposed emphasis on quantifying the degree of confounding required to nullify the observed effect could potentially be associated with the same pitfalls seen in null hypothesis significance testing.[22] Perhaps the value proposition of the E-value methodology should lie in quantifying the degree of bias in order to derive the best possible estimate of the effect size, instead of being used as a convenient means to claim immunity against unmeasured confounding.

Besides the above-mentioned difficulties in the interpretation, we identified additional notable observations which may present opportunities for the misuse of E-values. The first such opportunity is in the selective emphasis of favourable E-values within articles which may impart interpretation spins. For example, higher E-values were observed to be emphasized in the discussion over lower ones in a study that calculated multiple E-values.[23]

Second, we observed that most studies using continuous exposures only compared the highest and lowest quantile, which would have larger E-values than smaller (e.g. per unit) exposure contrasts. It has been documented that when relative risks for continuous exposures are weak, authors tend to select more extreme contrasts that will make the risks seem numerically large.[24] It is possible that a similar practice could be extrapolated to E-values, to inflate their magnitude in order to downplay the risk of unmeasured confounding.

We wish to highlight some other miscellaneous observations that could lead to misuse. Perusal of the verbatim statements shows that reasoning was often weak and E-values did not strengthen it. One study had even applied E-values to the analysis of a randomized trial,[25] apparently a redundant practice considering that randomization should hopefully remove the need to apologize for confounding from uncontrolled baseline variables. Additionally, the vast majority of studies calculated E-values only for estimates of relative effect measures,

Table 5. Cross-tabulation of conclusions about confounding for matched case-control pairs ($n = 69$)

		Conclusions about confounding from case publications			
		Unlikely to affect	Likely to affect	Both	No comment
Conclusions about confounding from control publications	Unlikely to affect	1	0	0	2
	Likely to affect	11	1	0	2
	Both	0	0	0	0
	No comment	32	4	2	14

neglecting estimates of absolute difference effect measures. This trend of focusing only on the multiplicative scale and not the additive scale might indicate that incomplete and thus potentially misleading assessment of uncontrolled confounding could be prevalent, given the choice of effect measures selected for E-value calculations.[26]

This study has several limitations. First, we included studies citing one of the two papers that first introduced E-values and may have missed works that applied E-values without citation. In addition, the E-value concept is relatively novel, and its use should be assessed again at a later stage. Second, our control group may not be a perfect comparator. The control group of observational studies that did not use E-values was matched based on journal and issue. We were unable to match on field of study or research question, as the degree of granularity on which to match on these variables would likely have been arbitrary. Even so, however, handling of confounding in many epidemiological studies is known to be poor in general across the literature,[21] and our study suggests that studies that use E-values also do not fare well in their handling of confounding. This tool may therefore be adding an extra layer of opportunities for misleading inferences. Third, it is possible that there may be selective reporting of E-values, where E-values are more likely to be reported when their magnitude falls in line with the author's intended conclusion. As a result, the sample of E-values analysed in this study could be biased.

Given the observed difficulties in interpreting E-values and potential for misuse, alternative recommendations are needed for addressing confounding in observational studies. Thorough reporting of what has been done, as recommended by STROBE,[27] is an important first step but not sufficient. Existing tools of sensitivity analyses[3–6] may be employed with careful consideration about their assumptions and study-specific applicability. In our sample, different sensitivity analyses were used in extremely diverse ways and they rarely changed the main conclusions. One wonders whether this signifies robustness of these conclusions or the fact that sensitivity analyses are reported mostly when they align with the key message that the authors pre-emptively want to infer. We argue that known confounders should be systematically assessed for their prevalence and magnitude.[9] This may require the availability of careful, field-wide, systematic, umbrella reviews where all known confounders are examined.[28] The results of these reviews may then be used consistently by many studies in the same field. Alternatively, if what to adjust for and how is left entirely to the discretion of the individual researcher, very divergent choices may be made and results may be largely dependent on these choices.[29–31]

Conclusion

Few papers using E-values conclude that confounding may influence their results, and E-values are largely overlapping in magnitude regardless of what is concluded. E-values may compound the already poor handling of confounding, and they may provide a false sense of confidence in the robustness of observational associations against uncontrolled confounding. Simply reporting an E-value as an indicator of susceptibility to confounding, though convenient, should not be a substitute for critical assessment of the influence and magnitude of specific confounders in specific fields of study.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

METRICS has been supported by a grant from the Laura and John Arnold foundation. M.R.B.'s work is supported by a grant from the Swiss National Science Foundation (P2BEP3_175289). Y.J.T. is supported by the Stanford Graduate Fellowship from Stanford University.

Author Contributions

J.P.A.I. had the original idea and all three authors worked on the concept. Y.J.T. and M.R.B. collected data. All authors analysed data and interpreted them. M.R.B. and Y.J.T. wrote the paper and all three authors revised and approved it.

Conflict of interest: None declared.

References

1. Groenwold RHH, Sterne JAC, Lawlor DA, Moons KGM, Hoes AW, Tilling K. Sensitivity analysis for the effects of multiple unmeasured confounders. *Ann Epidemiol* 2016;**26**:605–11.
2. Fewell Z, Davey Smith G, Sterne J. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol* 2007;**166**:646–55.
3. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* 1998;**54**:948–63.
4. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J R Stat Soc Ser B Methodol* 1983;**45**:212–18.
5. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst* 1959;**22**:173–203.
6. Rothman K, Lash T, Greenland S. *Modern Epidemiology*. 3rd edn. https://www.rti.org/publication/modern-epidemiology-3rd-edition (11 December 2018, date last accessed).
7. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med* 2017;**167**:268.
8. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology* 2016;**27**:368–77.
9. Ioannidis JPA, Tan YJ, Blum MR. Limitations and misinterpretations of E-values for sensitivity analyses of observational studies. *Ann Intern Med* 2019;**170**:108.
10. Localio AR, Stack CB, Griswold ME. Sensitivity analysis for unmeasured confounding: E-values for observational studies. *Ann Intern Med* 2017;**167**:285–86.
11. Wickham H, Chang W, Henry L *et al*. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. 2018. https://CRAN.R-project.org/package=ggplot2 (9 November 2018, date last accessed).
12. Wickham H, Henry L. *RStudio. tidyr: Easily Tidy Data With "Spread()" and "Gather()" Functions*. 2018. https://CRAN.R-project.org/package=tidyr (9 November 2018, date last accessed).
13. Auguie B, Antonov A. gridExtra: Miscellaneous Functions for "Grid" Graphics. 2017. https://CRAN.R-project.org/package=gridExtra (12 December 2018, date last accessed).
14. Harzing AW. *Publish or Perish*. 2007. https://harzing.com/resources/publish-or-perish (12 December 2018, date last accessed).
15. Bender Ignacio RA, Madison AT, Moshiri A, Weiss NS, Mueller BA. A population-based study of perinatal infection risk in women with and without systemic lupus erythematosus and their infants. *Paediatr Perinat Epidemiol* 2018;**32**:81–9.
16. Zhang L, Shen S, He J *et al*. Effect of interpregnancy interval on adverse perinatal outcomes in Southern China: a retrospective cohort study, 2000-2015. *Paediatr Perinat Epidemiol* 2018;**32**:131–40.
17. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Hillsdale, NJ: L. Erlbaum Associates, 1988.
18. Guyatt GH, Oxman AD, Vist GE *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;**336**:924–26.
19. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007;**334**:349–51.
20. Djulbegovic B, Glasziou P, Klocksieben FA *et al*. Larger effect sizes in nonrandomized studies are associated with higher rates of EMA licensing approval. *J Clin Epidemiol* 2018;**98**:24–32.
21. Hemkens LG, Ewald H, Naudet F *et al*. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *J Clin Epidemiol* 2018;**93**:94–102.
22. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and an alternative. *J Wildl Manag* 2000;**64**:912–23.
23. Tu C-Y, Hsia T-C, Fang H-Y *et al*. A population-based study of the effectiveness of stereotactic ablative radiotherapy versus conventional fractionated radiotherapy for clinical stage I non-small cell lung cancer patients. *Radiol Oncol* 2017;**52**:181–88.
24. Kavvoura FK, Liberopoulos G, Ioannidis J. Selection in reported epidemiological risks: an empirical assessment. *PLOS Med* 2007;**4**:e79.
25. Marsden J, Goetz C, Meynen T *et al*. Memory-focused cognitive therapy for cocaine use disorder: theory, procedures and preliminary evidence from an external pilot randomised controlled trial. *EBioMedicine* 2018;**29**:177–89.
26. Poole C. On the origin of risk relativism. *Epidemiology* 2010;**21**:3–9.
27. Vandenbroucke JP, Elm E von, Altman DG, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLOS Med* 2007;**4**:e297.
28. Ioannidis J. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *CMAJ* 2009;**181**:488–93.
29. Stang PE, Ryan PB, Overhage JM, Schuemie MJ, Hartzema AG, Welebob E. Variation in choice of study design: findings from the epidemiology design decision inventory and evaluation (EDDIE) survey. *Drug Saf* 2013;**36**:15–25.
30. Patel CJ, Burford B, Ioannidis J. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol* 2015;**68**:1046–58.
31. Silberzahn R, Uhlmann EL, Martin DP *et al*. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv Methods Pract Psychol Sci* 2018;**1**:337–56.