Journal of Clinical Epidemiology 111 (2019) 32−40

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

The statistical significance of meta-analyses is frequently fragile: definition of a fragility index for meta-analyses

Ignacio Atal[a,b,c,*], Raphaël Porcher[a,b,c], Isabelle Boutron[a,b,c], Philippe Ravaud[a,b,c,d]

[a] *Centre d'Épidémiologie Clinique, Hôpital Hôtel-Dieu, Paris, France*
[b] *Team METHODS, Centre of Research in Epidemiology and Statistics Sorbonne, Paris Cité−CRESS Inserm UMR1153, Paris, France*
[c] *Université Paris Descartes, Paris, France*
[d] *Epidemiology Department, Mailman School of Public Health, Columbia University, New York, NY, USA*

Accepted 26 March 2019; Published online 30 March 2019

Abstract

Objectives: Meta-analyses inform clinical practice by summarizing treatment effect estimates based on results from several trials. However, the statistical significance of a meta-analysis (i.e., whether the pooled treatment effect is statistically significant or not) may rely on the outcome of only a few patients from specific trials in the meta-analysis. We aimed to evaluate the extent to which the statistical significance of meta-analyses can be changed (from statistically significant to nonsignificant, or vice versa) after modifying the event status of patients in specific arms of specific trials.

Methods: We conducted a cross-sectional analysis of meta-analyses of trials with a binary outcome from Cochrane Systematic Reviews. We defined the fragility index of meta-analyses as the minimum number of patients from one or more trials included in the meta-analysis for whom an event-status modification (i.e., changing an event to nonevent or a nonevent to event) would change the statistical significance of the pooled treatment effect. For statistically significant and nonsignificant meta-analyses, we evaluated the fragility index, the ratio between the fragility index and the total number of participants included in the trials, and the ratio between the fragility index and the total number of events.

Results: Our sample comprised 906 meta-analyses: 400 and 506 had statistically significant and nonsignificant pooled treatment effects, respectively. For statistically significant meta-analyses, the median fragility index was 12 (Q1−Q3: 4−33); for 29% the fragility index was 5 or less. Overall, 43% and 9% meta-analyses would have become nonsignificant if the event status was modified for less than 1% of the total participants in one or several specific trials, and for less than 1% of the total number of events, respectively. These proportions were similar for statistically nonsignificant meta-analyses. Overall, the statistical significance of 33% of all meta-analyses depended on the event status of five or fewer participants from one or more specific trials.

Conclusion: The statistical significance of meta-analyses often depends on the outcome of a few patients. The fragility index of meta-analyses may help in interpreting the conclusions of meta-analyses. © 2019 Elsevier Inc. All rights reserved.

Keywords: Fragility index; Meta-analyses; Research methods; *P*-value; Statistical significance; Systematic reviews

1. Introduction

Meta-analyses of randomized controlled trials (RCTs) are considered the highest level of evidence in terms of the effectiveness of interventions. Meta-analyses are observational byproducts of the existing literature that combine a series of results from existing RCTs to derive a pooled treatment effect estimate with an increased statistical power. A meta-analysis with a statistically significant pooled treatment effect is likely to lead to a conclusion concerning the effectiveness of the intervention evaluated. However, the confidence we have in the conclusion of a meta-analysis may also depend on other factors such as the methods used to produce the meta-analysis [1], the quality

Ethics approval and consent to participate: Not applicable.

Conflict of interest: P.R. is the director of Cochrane France. The other authors declare that they have no competing interests.

Consent for publication: Not applicable.

Availability of data and materials: Data for the results are available from the corresponding author on request, and the code used for calculating the fragility index of meta-analyses is shared as open source under the MIT license at http://github.com/iatal/fragility_ma.

* Corresponding author. Centre d'Epidémiologie Clinique, Hôpital Hôtel-Dieu, 1, place du parvis Notre Dame, 75004 Paris, France. Tel.: +33 1 42 34 87 65; Fax: +33 142348790.

E-mail address: ignacio.atal-ext@aphp.fr (I. Atal).

https://doi.org/10.1016/j.jclinepi.2019.03.012
0895-4356/© 2019 Elsevier Inc. All rights reserved.

What is new?

Key findings

- In this cross-sectional study that included 906 meta-analyses of trials with binary outcomes from Cochrane Systematic Reviews, the statistical significance of the pooled treatment effect for 33% of the meta-analyses could have been changed after modifying the event status for five or fewer patients in specific trials.
- For almost half (44.3%) of the meta-analyses, the statistical significance depended on the event status of less than 1% of participants.

What this adds to what was known?

- For randomized controlled trials showing statistically significant results, the fragility index has been defined as the minimal number of patients whose status would have to be modified from a nonevent to an event to change the result to statistically nonsignificant. We extended the notion of fragility index to meta-analyses, defined as the minimum number of patients from one or more trials included in the meta-analysis for which a modification on the event status (i.e., changing events to nonevents, or nonevents to events) would change the statistical significance of the pooled treatment effect (from statistically significant to nonsignificant, or vice versa).
- Based on the fragility index, we showed that the statistical significance of meta-analyses often depends on the outcome of a few patients.

What is the implication and what should change now?

- The fragility index appears as an intuitive means for clinicians to underline the fragility of the statistically significance of meta-analyses and illustrates an additional drawback of the use of arbitrary thresholds (e.g., P-value < 0.05) for defining statistical significance of treatment effects.

of trials included, the number of trials, the heterogeneity of treatment effect estimates across trials, and the precision of each trial for estimating the treatment effect [2,3]. If a meta-analysis showing a statistically significant effect includes an RCT of low quality, with a small sample size, or presenting an extreme treatment effect, we might not conclude on the effectiveness of the treatment evaluated. Furthermore, the statistical significance of a meta-analysis could be changed (e.g., from a statistically significant to a statistically nonsignificant pooled treatment effect) if the results from an RCT were added to (or suppressed from) the meta-analysis, but also if the outcomes of a few patients were modified within the RCTs included in it.

For RCTs showing statistically significant results, the fragility index has been defined to measure the confidence we have in the statistical significance [4]. The fragility index corresponds to the minimal number of patients whose status would have to be modified from a nonevent to an event to change the result to statistically nonsignificant. The statistical significance of an RCT with fragility index of 1 is fragile in the sense that it relies on the event of only one patient. By reviewing a sample of 399 RCTs published in leading medical journals, Walsh et al showed that for 25% of RCTs, the fragility index was 3 or less. Similarly, the statistical significance of meta-analyses of RCTs could be fragile in terms of modifications in the outcome of patients included in the RCTs.

In this work, we extended the notion of fragility index to meta-analyses of RCTs, that is a measure of the confidence we could have in the statistical significance in terms of the minimal number of patients from specific RCTs whose event status would have to be changed to modify the statistical significance of the pooled treatment effect (from statistically significant to nonsignificant, or vice versa). We measured the fragility of the statistical significance of a large sample of meta-analyses from Cochrane Systematic Reviews.

2. Methods

We defined the fragility index for meta-analyses of binary outcomes and extended the definition to statistically nonsignificant meta-analyses. We then evaluated the fragility index for statistically significant and nonsignificant meta-analyses from a large sample of meta-analyses from Cochrane Systematic Reviews.

2.1. *Definition of a fragility index for meta-analyses*

The fragility index for an RCT with a statistically significant result was defined as the minimum number of nonevents that would need to be changed to events in one arm to switch the result to statistically nonsignificant [4].

Based on this notion, we defined a fragility index for meta-analyses of clinical trials showing a statistically significant treatment effect. Meta-analyses combine results from several trials to estimate a pooled point estimate and the 95% confidence interval (CI) of the treatment effect. The statistical significance of a meta-analysis is decided by whether or not the 95% CI of the pooled treatment effect overlaps the null effect. We defined a fragility index for meta-analyses as the minimum number of patients from one or more trials included in the meta-analysis for which a modification in the event status (i.e., changing an event to a nonevent or a nonevent to an event) would change the statistical significance of the pooled treatment effect to statistically nonsignificant (Fig. 1).

The minimal number of event status modifications leading to a statistically nonsignificant pooled treatment effect could be achieved by changing nonevents to events in one group of patients but also events to nonevents in the other group, or a mix of both, and for different groups of patients of different trials. Because the influence of each trial in the pooled treatment effect of the meta-analysis may differ, these event status modifications may need to occur in one or more specific trials.

In addition, we defined a fragility index for nonstatistically significant meta-analyses as the minimum number of patients in one or more specific trials for whom an event-status modification would change the pooled treatment effect of the meta-analysis to statistically significant.

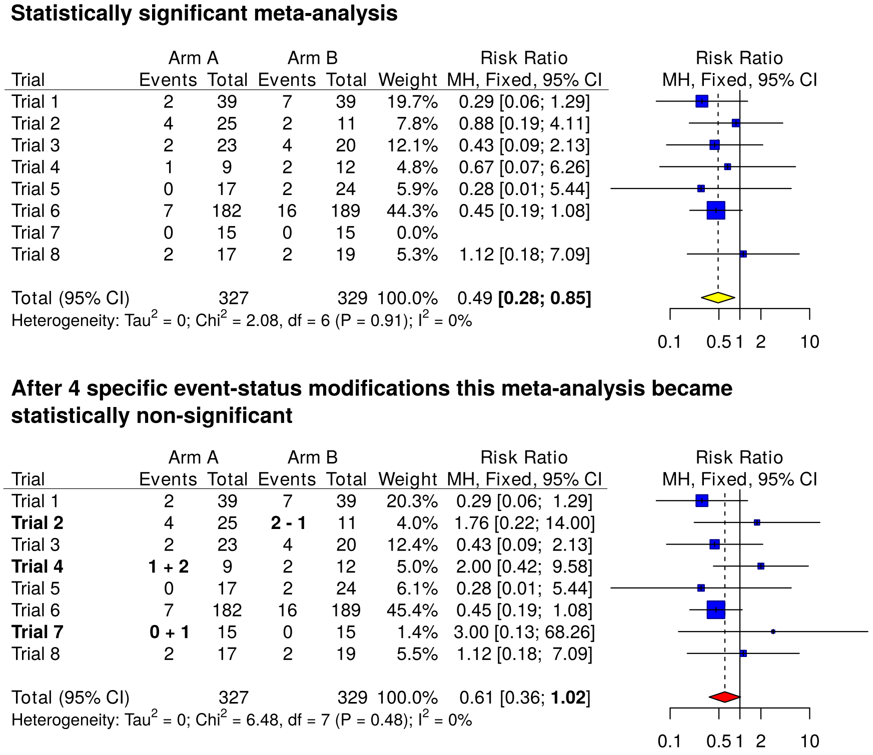
By combining the fragility index of both statistically significant and nonsignificant meta-analyses, we can derive a general fragility index for meta-analyses, corresponding to the minimum number of patients in one or more specific trials for whom an event-status modification would change the statistical significance of the meta-analysis.

2.2. *Method for estimating the fragility index for meta-analyses*

We developed a method to estimate the fragility index of meta-analyses by using a heuristic process to find specific trials and event-status modifications that could change the statistical significance of the pooled treatment effect of the meta-analysis with the smallest number of modifications (Fig. 2). The method is based on an iterative re-evaluation of the statistical significance of the pooled treatment effect of modified meta-analyses, iteratively derived from the original meta-analysis by performing single event-status modifications in each arm of each trial in turn.

For instance, let us consider a meta-analysis combining the results of *N* trials evaluating the risk ratio (RR) of a binary outcome between two treatments A and B. In addition, let us consider that the meta-analysis leads to a statistically significant effect based on a pooled RR with a 95% confidence limit <1 showing that the number of events for patients receiving treatment A was significantly less than that for the group of patients receiving treatment B.

To evaluate the fragility index of that meta-analysis, we sequentially recalculated the 95% CI of the pooled RR after performing all single event-status modifications that would increase the RR, which could be 1) changing a nonevent to an event for patients receiving treatment A, for each single trial, or 2) changing an event to a nonevent for patients receiving treatment B, for each trial. This process leads to $2N$ newly calculated 95% CIs for the pooled RR. If one of the newly calculated CIs overlapped 1, the fragility index of the meta-analysis was 1 because we found a specific trial



Statistically significant meta-analysis

Trial	Arm A Events	Arm A Total	Arm B Events	Arm B Total	Weight	Risk Ratio MH, Fixed, 95% CI
Trial 1	2	39	7	39	19.7%	0.29 [0.06; 1.29]
Trial 2	4	25	2	11	7.8%	0.88 [0.19; 4.11]
Trial 3	2	23	4	20	12.1%	0.43 [0.09; 2.13]
Trial 4	1	9	2	12	4.8%	0.67 [0.07; 6.26]
Trial 5	0	17	2	24	5.9%	0.28 [0.01; 5.44]
Trial 6	7	182	16	189	44.3%	0.45 [0.19; 1.08]
Trial 7	0	15	0	15	0.0%	
Trial 8	2	17	2	19	5.3%	1.12 [0.18; 7.09]
Total (95% CI)		327		329	100.0%	0.49 **[0.28; 0.85]**

Heterogeneity: Tau^2 = 0; Chi^2 = 2.08, df = 6 (P = 0.91); I^2 = 0%

After 4 specific event-status modifications this meta-analysis became statistically non-significant

Trial	Arm A Events	Arm A Total	Arm B Events	Arm B Total	Weight	Risk Ratio MH, Fixed, 95% CI
Trial 1	2	39	7	39	20.3%	0.29 [0.06; 1.29]
Trial 2	4	25	**2 - 1**	11	4.0%	1.76 [0.22; 14.00]
Trial 3	2	23	4	20	12.4%	0.43 [0.09; 2.13]
Trial 4	**1 + 2**	9	2	12	5.0%	2.00 [0.42; 9.58]
Trial 5	0	17	2	24	6.1%	0.28 [0.01; 5.44]
Trial 6	7	182	16	189	45.4%	0.45 [0.19; 1.08]
Trial 7	**0 + 1**	15	0	15	1.4%	3.00 [0.13; 68.26]
Trial 8	2	17	2	19	5.5%	1.12 [0.18; 7.09]
Total (95% CI)		327		329	100.0%	0.61 [0.36; **1.02**]

Heterogeneity: Tau^2 = 0; Chi^2 = 6.48, df = 7 (P = 0.48); I^2 = 0%

Fig. 1. Example of a statistically significant meta-analysis with fragility index of 4. Data presented in this example corresponds to a real meta-analysis from a Cochrane Systematic Review. The measure used for evaluating the treatment effect (here risk ratio) and for deriving the pooled treatment effect (here Mantel-Haenszel with a fixed-effects model) was the same as that used in the Cochrane Systematic Review containing the data. The fragility index for the meta-analysis was then derived relative to the measure and method used in Cochrane Systematic Reviews for each individual meta-analysis.

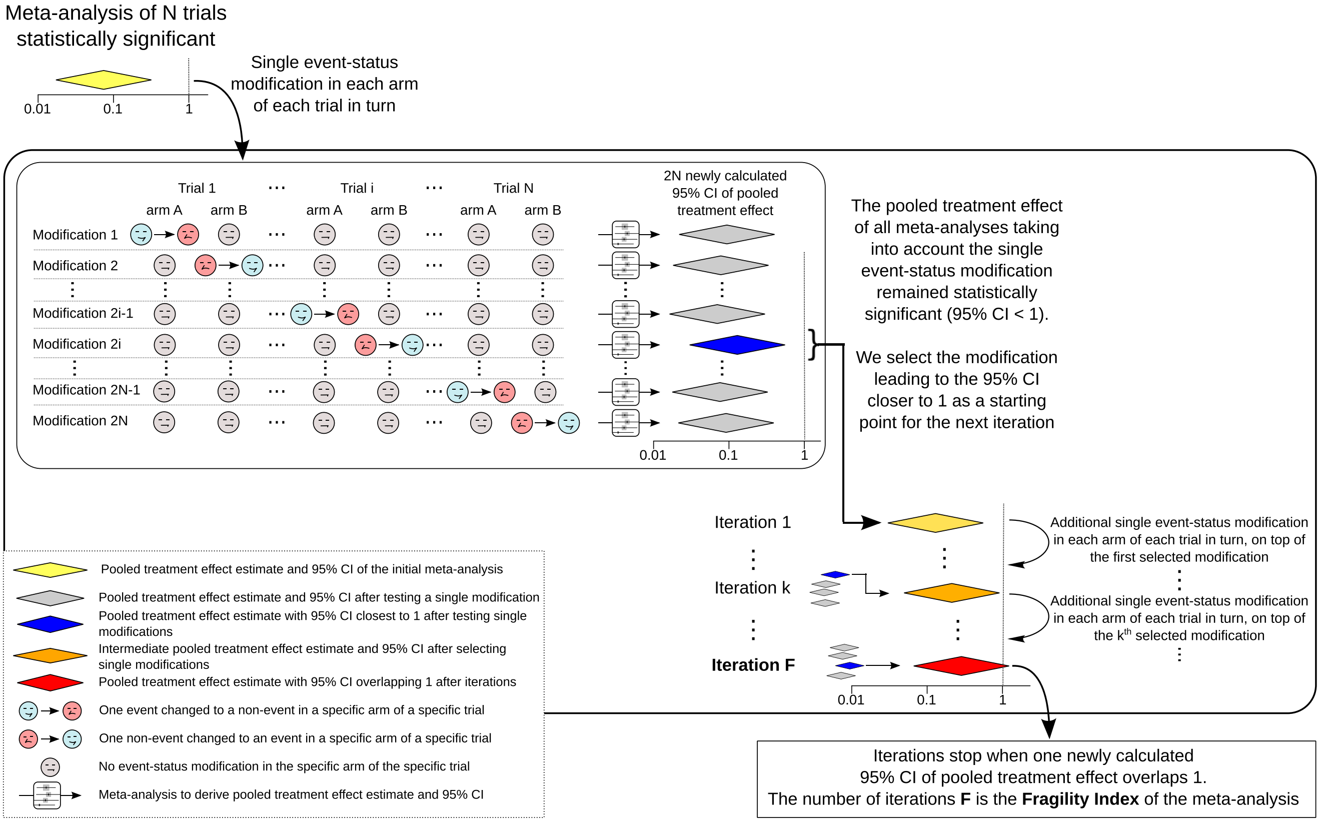

Meta-analysis of N trials statistically significant
0.01
0.1
1
Single event-status modification in each arm of each trial in turn
Trial 1
Trial i
Trial N
arm A
arm B
2N newly calculated 95% CI of pooled treatment effect
Modification 1
Modification 2
Modification 2i-1
Modification 2i
Modification 2N-1
Modification 2N
The pooled treatment effect of all meta-analyses taking into account the single event-status modification remained statistically significant (95% CI < 1).
We select the modification leading to the 95% CI closer to 1 as a starting point for the next iteration
Iteration 1
Iteration k
Iteration F
Additional single event-status modification in each arm of each trial in turn, on top of the first selected modification
Additional single event-status modification in each arm of each trial in turn, on top of the k^th selected modification
Iterations stop when one newly calculated 95% CI of pooled treatment effect overlaps 1.
The number of iterations F is the Fragility Index of the meta-analysis
Pooled treatment effect estimate and 95% CI of the initial meta-analysis
Pooled treatment effect estimate and 95% CI after testing a single modification
Pooled treatment effect estimate with 95% CI closest to 1 after testing single modifications
Intermediate pooled treatment effect estimate and 95% CI after selecting single modifications
Pooled treatment effect estimate with 95% CI overlapping 1 after iterations
One event changed to a non-event in a specific arm of a specific trial
One non-event changed to an event in a specific arm of a specific trial
No event-status modification in the specific arm of the specific trial
Meta-analysis to derive pooled treatment effect estimate and 95% CI

Fig. 2. Iterative method for estimating the fragility index of a statistically significant meta-analysis. The method is based on an iterative re-evaluation of the statistical significance of the pooled treatment effect of modified meta-analyses, iteratively derived from the original meta-analysis by performing single event-status modifications in each arm of each trial in turn. The same method used in the original meta-analysis for estimating pooled treatment effects (Mantel-Haenszel, Peto, or inverse variance, as well as fixed or random effects) was used in each iteration to evaluate the statistical significance of the modified meta-analyses.

for which a unique event status modification (i.e., changing a nonevent to an event in arm A or an event to a nonevent in arm B) changed the statistical significance of the meta-analysis. If all the newly calculated 95% CIs for the pooled RR remained <1, we selected the specific trial and specific event-status modification that led to the 95% CI for the pooled RR being closer to 1 as a starting point for the next iteration. Then, we repeated the process by performing a new single event-status modification in each arm of each trial in turn, on top of the first selected modification. Similarly, if one of these *2N* event status modifications led to a newly calculated 95% CIs for the pooled RR overlapping 1, the fragility index of the meta-analysis was then equal to 2. This process was iterated until one event-status modifications led to a newly calculated 95% CI for the pooled RR overlapping 1. The number of iterations needed to find a combination of event-status modifications in specific arms and trials leading to a modified meta-analysis with 95% CI for the pooled RR overlapping 1 was defined as the fragility index of the meta-analysis.

The same iterative method was used to evaluate the fragility index for statistically nonsignificant meta-analyses. For a statistically nonsignificant meta-analysis, the algorithm considers the two cases of when event-status modifications for specific arms of specific trials could lead to a pooled treatment effect statistically significantly in favor of treatment A and in favor of treatment B, respectively. The minimum number of modifications for both was defined as the fragility index of the statistically nonsignificant meta-analysis.

The method developed is not based on an exhaustive evaluation of all possible combinations of event status modifications in all trials and all arms until finding the minimal combination changing the statistical significance of the pooled treatment effect. Such a method would need the testing of an exponentially growing amount of modification scenarios as the number of trials in the meta-analysis grows, which can become computationally too intensive in most of the cases. Our method is rather based on an iterative process designed to optimally choose the sequence of event status modifications being the most likely to rapidly change the statistical significance of the pooled treatment effect, in a reasonable number of steps (Fig. S1).

This general method is applicable whatever the measure used to evaluate the treatment effect (e.g., RR, odds ratio or risk difference) or method used in the meta-analysis to pool the treatment effects across trials, such as how trials without events were handled, or the method used for weighting trials (e.g., fixed or random effects).

2.3. Data sources and selection

We obtained all Cochrane Systematic Reviews published between March 2011 and September 2014, which were described in a previous work [5]. For each review, we considered only the first meta-analysis reported (i.e., the first comparison and first outcome of the review). We included only meta-analyses of RCTs (i.e., analyses not including observational studies) with a binary outcome and having an estimated pooled treatment effect in the review. If the pooled treatment effect was estimated at a subgroup level only (without an overall estimate), we retained the first subgroup only. In addition, we excluded meta-analyses if the sum of the sample sizes of the included studies was less than 30 and if no patient or all patients of all studies had an event.

For each meta-analysis selected, we extracted for each trial included the number of events and number of patients in each group of patients. We also extracted the measure used for evaluating the treatment effect (RR, odds ratio, risk difference), and the methods used for pooling treatment effects (Mantel-Haenszel, Peto, or inverse variance, as well as fixed or random effects) to evaluate the statistical significance of each meta-analysis by the same method used in the original systematic review.

2.4. Data analysis

We evaluated the fragility index for statistically significant and nonsignificant meta-analyses by using for each iteration of the evaluation process the same measure and method used in the original systematic review to recalculate the new 95% CIs of the pooled treatment effect after single event-status modifications in trials, that is, if the original systematic review used RR to measure treatment effects, and Mantel-Haenszel with random effects to estimate the pooled treatment effect and its 95% CI, we consistently used the same effect measure and method to pool the treatment effects at each step of the iterative method to evaluate the fragility index.

For each meta-analysis, we assessed the number of trials included as well as the total sample size and number of events, corresponding to the sum of the sample sizes and number of events in the included trials, respectively. If the total number of events was greater than half the total sample size, we considered nonevents as events so as to always have a total number of events less than 50% of the total sample size. In fact, because in our definition of the fragility index, events could be changed to nonevents and nonevents to events, both events and nonevents play symmetric roles.

We computed the fragility index divided by the total sample size for statistically significant, statistically nonsignificant, and all meta-analyses, corresponding to the minimal proportion of patients from one or more specific trials for whom changing the event status would change the statistical significance of the meta-analysis. We also evaluated the ratio between the fragility index and the total number of events, corresponding to the minimum proportion of events to be modified in one or more specific trials to change the statistical significance of the meta-analysis.

For additional descriptive analyses, we evaluated for each initial meta-analysis the *P*-value, and compared it with the fragility index. In addition, we compared the fragility index with the point estimate of the effect size for meta-analyses using RR as measure and by interchanging treatment A and treatment B to have an estimated RR $>$ 1.

2.5. Changing event status only in the trial with the largest or the smallest sample size

To further investigate the fragility index of a meta-analysis as evaluated in our work, we evaluated if it was possible to change the statistical significance of the pooled treatment effect after modifying event status only in the largest (respectively, the smallest) trial. If that was possible, we compared the fragility index with the minimal number of event-status modifications needed in the largest (respectively, the smallest) trial to change the statistical significance of the pooled treatment effect. When several trials had the same sample size and were together the largest (or the smallest) trials, we conducted modifications of event status separately for each of those trials, and considered the overall minimal number of event-status modifications across those trials.

2.6. Research reproducibility

All data analyses and figure preparation involved use of R 3.3.1 (R Development Core Team, Vienna, Austria) and *meta* package [6]. A web interface for calculating the fragility index of meta-analyses is available at http://clinicalepidemio.fr/fragility_ma/.

3. Results

From 2,796 Cochrane Systematic Reviews, 906 meta-analyses met our inclusion criteria (Fig. S2). These meta-analyses included data from 6,625 trials (median four trials [Q1−Q3: 3−8]). The median total sample size was 756 (301−2,014) and median total number of events 127 (47−341) (Table 1). Overall, 400 (44.2%) meta-analyses had statistically significant results, and 506 (55.8%) had statistically nonsignificant results.

3.1. Fragility index of statistically significant meta-analyses

The median fragility index of statistically significant meta-analyses was 12 (Q1−Q3: 4−33; range 1−858). In total, 29.0% of statistically significant meta-analyses had a fragility index of 5 or less, and 6.2% had a fragility index

Table 1. Characteristics of meta-analyses, statistically significant, statistically nonsignificant, and overall

Characteristic	Statistically significant *N* = 400	Statistically nonsignificant *N* = 506	Overall *N* = 906
Number of trials			
Median (Q1–Q3)	5 (2–7)	4 (2–7)	4 (3–8)
Min–Max	2–138	2–78	2–138
Total sample size			
Median (Q1–Q3)	917 (363–2,457)	635 (259–1,717)	756 (301–2,014)
Min–Max	30–732,400	30–341,300	30–732,400
Total number of events			
Median (Q1–Q3)	187 (78–492)	76 (26–238)	127 (47–341)
Min–Max	9–18,810	1–32,960	1–32,960
Effect measure			
Risk ratio	318	387	705
Odds ratio	79	113	192
Risk difference	3	6	9
Method used			
Mantel-Haenszel	369	455	824
Peto	17	36	53
Inverse variance	14	15	29
Fixed/random effects	235/165	325/181	560/346

of 1 (Fig. 3). In terms of total sample size, the fragility index did not increase when the total sample size was less than 500 (Fig. 4). The proportion of statistically significant meta-analyses with fragility index of 5 or less was 55.6%, 40.5%, and 50.1% for total sample sizes less than 100, 100 to 200, and 200 to 500, respectively. In addition, for all ranges of total sample sizes, we found statistically significant meta-analyses with fragility index equal to 1. The

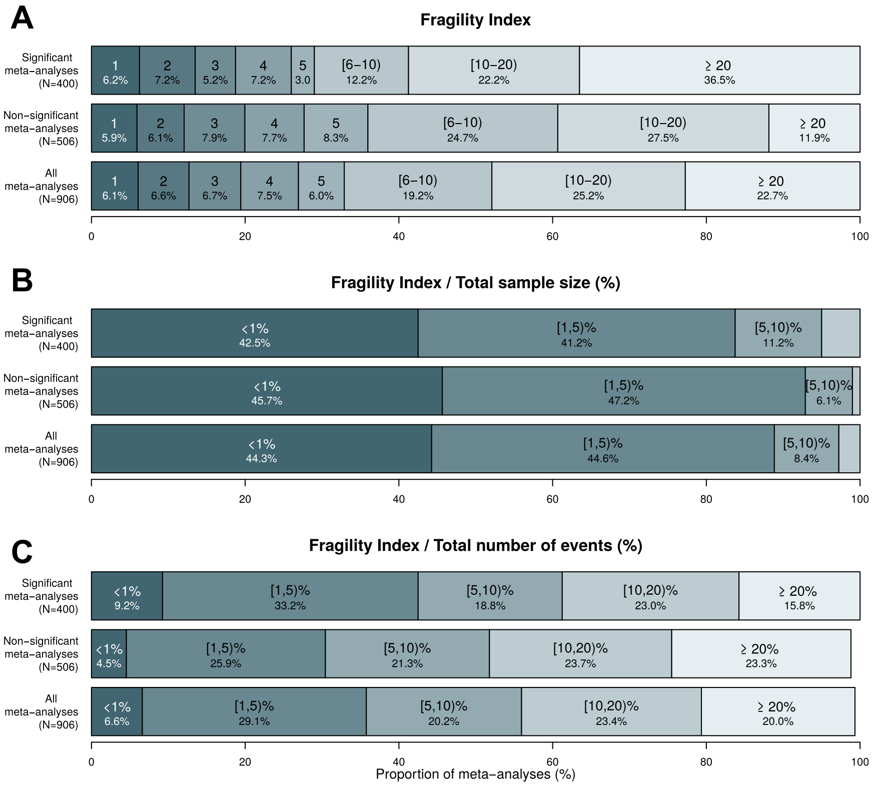


Fig. 3. Distribution of meta-analyses by fragility index (A), ratio between the fragility index and total sample size (B), and ratio between the fragility index and total number of events (C), for statistically significant and nonsignificant meta-analyses, and overall meta-analyses. Total sample size and total number of events corresponds to the sum of the sample sizes and number of events in the included trials of the meta-analyses, respectively.

fragility index increased according to total number of events and was consistently higher than five when the total number of events was more than 1,000.

For 42.5% of the statistically significant meta-analyses, the pooled treatment effect would have changed to nonsignificant if the event status was modified for less than 1% of the total participants in one or more specific trials (Fig. 3B), and for 9.2% and 42.5%, the statistical significance would have changed if less than 1% and 5% of the total events, respectively, were modified in one or more specific trials (Fig. 3C).

Statistically significant meta-analyses with P-value $>$ 0.045 had systematically a fragility index of 1 or 2 (Fig. S4). In our sample, the median fragility index increased as P-values decreased. Nevertheless, having a P-value $<$ 0.005 did not excluded the possibility of having a fragility index of 1 [7]. The distribution of the fragility index of statistically significant meta-analyses was similar for all ranges of effect sizes (Fig. S5).

3.2. Fragility index of statistically nonsignificant meta-analyses

The median fragility index of statistically nonsignificant meta-analyses was 7 (Q1–Q3: 4–14, range 1–102). More than one-third (36.0%) of statistically nonsignificant meta-analyses had a fragility index of 5 or less, and 5.9% had a fragility index of 1 (Fig. 3A). We identified 10 statistically nonsignificant meta-analyses with total sample size more than 1,000 for which a single event-status modification was needed to change to statistically significant. Nevertheless, the fragility index for statistically nonsignificant meta-analyses increased according to the total number of events (Fig. 4). In particular, nonsignificant meta-analyses with more than 1,000 events would need more than six event-status modifications in one or more specific trials to become statistically significant.

For 45.7% of the statistically nonsignificant meta-analyses, the pooled treatment effect would have changed to significant if the event status was modified for less than 1% of the total participants in one or more specific trials (Fig. 3B), and for 4.5% and 30.4%, the statistical significance would have changed if less than 1% and 5% of the total events, respectively, were modified in one or more specific trials (Fig. 3C).

Statistically nonsignificant meta-analyses with P-value $<$ 0.055 had systematically a fragility index of 1 or 2 (Fig. S4). In our sample, the median fragility index increased as P-values increased. Nevertheless, having a P-value $\geq$ 0.5 did not excluded the possibility of having a fragility index of 1 [8]. The median fragility index of statistically nonsignificant meta-analyses decreased when the effect size increased (Fig. S5).

3.3. Fragility index for all meta-analyses

When pooling statistically significant and nonsignificant meta-analyses, the median fragility index was 9 (Q1–Q3: 4–18). The statistical significance of one-third (32.9%) of meta-analyses depended on the event status of five or fewer participants from one or more specific trials, and the statistical significance of 6.1% of meta-analyses depended on the event status of only 1 participant (Fig. 3A). For almost half (44.3%) of the meta-analyses, the statistical significance depended on the event status of less than 1% of the participants from one or more specific trials (Fig. 3B). For 6.6% and 36.7% of the meta-analyses, the statistical significance would have changed if less than 1% and 5% of the events, respectively, were modified in one or more specific trials (Fig. 3C). The distribution of the fragility index overall and for statistically significant and nonsignificant meta-analyses were similar (Fig. S3).

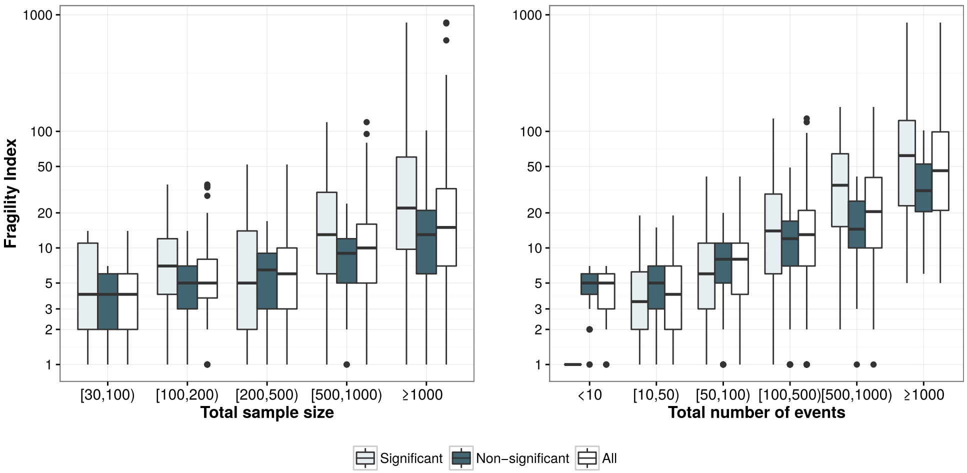


Fig. 4. Fragility index by total sample size and total number of events for statistically significant and nonsignificant meta-analyses, and overall meta-analyses. Total sample size and total number of events corresponds to the sum of the sample sizes and number of events in the included trials of the meta-analyses, respectively.

3.4. Changing event status only in the trial with the largest or the smallest sample size

The statistical significance of the pooled treatment effect could not be changed after conducting all possible event-status modifications in the largest trial and the smallest trial only for 28.6% and 57.4% of the meta-analyses, respectively. When modifying event status in the largest trial and the smallest trial only changed the statistical significance of the pooled treatment effect, the minimal number of modifications needed to change the statistical significance was inferior to the fragility index for 1.7% and 0.9% of the meta-analyses, respectively (Fig. S6).

4. Discussion

From the primary analysis of 906 Cochrane Systematic Reviews, we showed that the statistical significance of both statistically significant and nonsignificant meta-analyses were fragile to a few event-status modifications for patients in specific trials. Indeed, for almost one-third (32.9%) of all meta-analyses, the statistical significance could be changed after modifying the event status for five or fewer patients in one or more specific trials, and the statistical significance of 6.1% of the meta-analyses depended on the event status of only one patient from a specific trial. For almost half (44.3%) of the meta-analyses, the statistical significance depended on the event status of less than 1% of participants. These proportions were similar for statistically significant and nonsignificant meta-analyses.

Our results show that the statistical significance of meta-analyses, statistically significant or not, are fragile in the sense that they depend on the outcome of a few patients. The fragility index of meta-analyses may mechanically depend on a variety of factors such as the number of trials included, the number of participants or events per trial, the heterogeneity of treatment effects across trials, the method used for combining results, or the unknown true treatment effect. Other clinical research−driven factors may affect the fragility of meta-analyses, such as the quality of trials, publication bias, errors in data collection, subjectivity of outcomes, blinding of data collectors and analysts, or fraud [9,10]. Nevertheless, dissociating these sociological factors from purely mechanical factors is difficult when studying the fragility index. In addition, the fragility index may vary across disciplines: recent studies have reported a median fragility index of 16 and 2 for samples of statistically significant RCTs of diabetes and ophthalmology, respectively [11,12].

Our work has several strengths. First, we expanded the definition of the fragility index to meta-analyses so as to give an intuitive measure of confidence in their conclusions. The fragility index for statistically significant RCTs might be shifted toward low values when finite resources for clinical research force RCTs to be sized adequately to demonstrate an expected treatment effect with the minimum resources [13]. This situation should not be the case for meta-analyses because patient recruitment in RCTs is not initially planned to feed meta-analyses. Nevertheless, the fragility index for meta-analyses could be considered for planning future RCTs to increase the confidence in the conclusions of a meta-analysis [14]. For interpreting the results of a statistically significant meta-analysis, the GRADE working group has suggested comparing the total sample size of the meta-analysis to the size of an equivalent adequately sized trial for expected efficacy [15]. A priori assumptions of an effect size may be inaccurate, and may influence the fact that a meta-analysis meets the criterion [16,17]. Conversely, the fragility index for meta-analyses provides an intuitive measure for confidence in their conclusions that does not depend on an a priori effect-size hypothesis. Second, we conceptually extended the notion of the fragility index to statistically nonsignificant results. Indeed, statistically significant and nonsignificant results with low fragility index are equivalently close to the edge of changing their statistical significance with a few event-status modifications. Third, we evaluated the fragility index with a large number of meta-analyses, using only primary meta-analyses from the Cochrane Systematic Reviews, which are known to produce high-quality evidence impacting clinical care.

Our work has some limitations. First, the fragility index for meta-analyses as defined in our work considers that the minimal number of event-status modifications that could change the statistical significance of a meta-analysis may need to occur in specific trials well chosen by our iterative method. For instance, if the fragility index for a meta-analysis was 5, we could not say that the statistical significance of the meta-analysis depended on the outcome of any group of five patients from any trial, but we can say that for at least one combination of five patients from specific trials, modifying their outcome from an event to a nonevent and/or a nonevent to an event would change the statistical significance of the meta-analysis. Nevertheless, if our iterative method stopped at the fifth iteration, we cannot exclude either that another combination of four event-status modifications changing the statistical significance exists, unless we evaluate the 95% CI for all possible combinations of the four modifications. For instance, when modifying event status only in the trial with the largest (or the smallest, respectively) sample size, we found a number of event-status modifications changing the statistical significance of the pooled treatment effect inferior to the fragility index for 1.7% (and 0.9%, respectively) of the meta-analyses (Fig. S6). However, testing all the possible combinations is computationally intensive with increased number of modifications and number of trials included in a meta-analysis. Our method may then slightly overestimate the minimal number of specific modifications that would change the statistical significance of a meta-analysis, which is conservative. However, our method does not overestimate the fragility index when it is 1 or 2. Second, we did not take into account that the statistical significance of a meta-analysis may depend on the measure used (e.g., RR or odds ratio), or the method used for estimating the pooled treatment effect. For instance, by changing from fixed to random effects, a statistically significant meta-analysis may become nonsignificant without any event-status modification. In our work, we used the same method for combining study results as used in the original systematic review for each meta-analysis, and the fragility index was then evaluated relative to that method in each iteration of our iterative method. Finally, the fragility index is based on the binary criterion for statistical significance: "does the 95% CI of the treatment effect includes 1?" We acknowledge that the conclusions of meta-analyses should not be drawn only based on the statistical significance of their pooled treatment effect. Indeed, the use of arbitrary thresholds for defining statistical significance such as P-value < 0.05 have major drawbacks when conducting comparative effectiveness research [18]. However, researchers and publishers might still be inclined to report and publish only statistically significant results [19,20]. Our work illustrates an additional drawback of the use of such arbitrary thresholds for defining statistical significance. The fragility index appears as an intuitive means for clinicians to underline the fragility of the statistically significance of meta-analyses, and we suggest its reporting as part of the results of meta-analyses to shed another light on them.

5. Conclusion

Here, with a large number of meta-analyses from Cochrane Systematic Reviews, we showed that the statistical significance of meta-analyses often depend on the outcome of only few patients. The fragility index of meta-analyses can help interpret the conclusions of meta-analyses.

CRediT authorship contribution statement

Ignacio Atal: Conceptualization, Methodology, Formal analysis, Data curation. **Raphaël Porcher:** Conceptualization, Methodology, Data curation. **Isabelle Boutron:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Philippe Ravaud:** Conceptualization, Methodology, Data curation.

Acknowledgments

The authors thank Anna Chaimani for helpful discussions and critiques during the preparation of the manuscript, and Laura Smales for language revision of manuscript.

This work did not receive any specific grant.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.03.012.

References

[1] Murad MH, Montori VM, Ioannidis JPA, Jaeschke R, Devereaux PJ, Prasad K, et al. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. JAMA 2014;312:171–9.

[2] Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. BMJ 2013;346:f2304.

[3] Dechartres A, Altman DG, Trinquart L, Boutron I, Ravaud P. Association between analytic strategy and estimates of treatment outcomes in meta-analyses. JAMA 2014;312:623–30.

[4] Walsh M, Srinathan SK, McAuley DF, Mrkobrada M, Levine O, Ribic C, et al. The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index. J Clin Epidemiol 2014;67:622–8.

[5] Dechartres A, Trinquart L, Atal I, Moher D, Dickersin K, Boutron I, et al. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. BMJ 2017;357:j2490.

[6] Guido S. Meta: an R package for meta-analysis. R News 2007;7(3):40–5.

[7] Young C, von Dadelszen P, Alfirevic Z. Instruments for chorionic villus sampling for prenatal diagnosis. Cochrane Database Syst Rev 2013;CD000114.

[8] Njei B, Kongnyuy EJ, Kumar S, Okwen MP, Sankar MJ, Mbuagbaw L. Optimal timing for antiretroviral therapy initiation in patients with HIV infection and concurrent cryptococcal meningitis. Cochrane Database Syst Rev 2013;CD009012.

[9] Dechartres A, Ravaud P, Atal I, Riveros C, Boutron I. Association between trial registration and treatment effect estimates: a meta-epidemiological study. BMC Med 2016;14(1):100.

[10] George SL, Buyse M. Data fraud in clinical trials. Clin Investig 2015;5(2):161–73.

[11] Shen C, Shamsudeen I, Farrokhyar F, Sabri K. Fragility of results in ophthalmology randomized controlled trials: a systematic review. Ophthalmology 2018;125:642–8.

[12] Chase Kruse B, Matt Vassar B. Unbreakable? An analysis of the fragility of randomized trials that support diabetes treatment guidelines. Diabetes Res Clin Pract 2017;134:91–105.

[13] Carter RE, McKie PM, Storlie CB. The fragility index: a P-value in sheep's clothing? Eur Heart J 2017;38:346–8.

[14] Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. Stat Med 2007;26:2479–500.

[15] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. J Clin Epidemiol 2011;64:1283–93.

[16] Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ 2009;338:b1732.

[17] Garcia-Alamino JM, Bankhead C, Heneghan C, Pidduck N, Perera R. Impact of heterogeneity and effect size on the estimation of the optimal information size: analysis of recently published meta-analyses. BMJ Open 2017;7(11):e015888.

[18] Sterne JAC, Cox DR, Smith GD. Sifting the evidence—what's wrong with significance tests?Another comment on the role of statistical methods. BMJ 2001;322:226–31.

[19] Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. Proc Natl Acad Sci U S A 2018;115:2613–9.

[20] Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting P values in the biomedical literature, 1990-2015. JAMA 2016;315:1141–8.