Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment

Myura Nagendran,[1] Tiago V Pereira,[2] Grace Kiew,[3] Douglas G Altman,[4] Mahiben Maruthappu,[5] John P A Ioannidis,[6] Peter McCulloch[7]

[1]Division of Anaesthetics, Pain Medicine and Intensive Care, Imperial College London, UK
[2]Health Technology Assessment Unit, Institute of Education and Sciences, Hospital Alemão Oswaldo Cruz, Sao Paulo, Brazil
[3]Gonville and Caius College, University of Cambridge, UK
[4]Centre for Statistics in Medicine, Oxford, UK
[5]Department of Epidemiology and Public Health, University College London, UK
[6]Departments of Medicine, of Health Research and Policy, and of Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, USA
[7]Nuffield Department of Surgical Science, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK

Correspondence to: P McCulloch
peter.mcculloch@nds.ox.ac.uk

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2016;355:i5432
http://dx.doi.org/10.1136/bmj.i5432

Accepted: 29 September 2016

ABSTRACT

OBJECTIVE

To examine whether a very large effect (VLE; defined as a relative risk of ≤0.2 or ≥5) in a randomised trial could be an empirical marker that subsequent trials are unnecessary.

DESIGN

Meta-epidemiological assessment of existing published data on randomised trials.

DATA SOURCES

Cochrane Database of Systematic Reviews (2010, issue 7) with data on subsequent large trials updated to 2015, issue 12.

ELIGIBILITY CRITERIA

All binary outcome forest plots were selected, which contained an index randomised trial with a VLE that was nominally statistically significant (P<0.05), included a subsequent large randomised trial (≥200 events and ≥200 non-events) for validation of the effect, assessed a primary outcome of the review, and was not a subgroup or sensitivity analysis.

RESULTS

Of 3082 reviews yielding 85 002 forest plots, only 44 (0.05%) satisfied the inclusion criteria. Index trials were generally small, with a median sample of 99 (median 14 events). Few index trials were rated at low risk of bias (9 of 44; 20%). The relative risk was closer to the null in the subsequent large trials in 43 of 44 cases. Subsequent large trial data failed to find a statistically significant (P<0.05) effect in the same direction in 19 cases (43%, 95% confidence interval 29% to 58%). Even when the subsequent large trials did find a significant effect in the same direction, the additional primary outcomes in most of these trials would have to be considered before deciding in favour of using the intervention. Subsequent large trial data found a statistically significant effect in the same direction in 19 of 21 cases when the index trial also had a value of P<0.001.

CONCLUSIONS

The frequency of VLEs followed by a large trial is vanishingly small, and where they occur they do not appear to be a reliable marker for a benefit that is reproducible and directly actionable. An empirical rule using a VLE in a randomised controlled trial as a marker that further trials are unnecessary would be neither practical nor useful. Caution should be taken when interpreting small studies with very large treatment effects.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Most healthcare interventions provide modest benefits, but randomised trials occasionally report very large improvements over existing treatments or inactive controls; this often leads to speculation that further trials might be unnecessary

The use of very large treatment effects as an empirical marker could highlight where resources might be wasted on unnecessary follow-up trials

However, large effect estimates are usually downgraded in subsequent trials, and the profile of their appearance and shift suggests regression to the mean as the cause

WHAT THIS STUDY ADDS

There does not appear to be an effect size large enough to be confident that future large (reliable) trials will always show a significant effect rather than one that could be due to chance

Most very large effect estimates come from small trials with large confidence intervals that should be interpreted with caution

These findings are highly relevant to fields such as surgery, where the average size of trials is usually much smaller than for drug trials, for logistical reasons

Introduction

Randomised controlled trials are perceived as the gold standard for settling interventional questions and maintain a dominant position in the hierarchy of medical evidence.[1] Under ideal circumstances, their data can provide essential information on efficacy and harms to clinicians and act as a powerful guide for policy makers. However, the value of conducting trials can be limited by both logistical factors that inhibit recruitment and recognised deficiencies in reporting (bias, selective publication, and lack of transparency).[2] A further crucial aspect of conducting such trials is the ethical requirement for clinical equipoise between treatments. Reaching a consensus agreement within the medical community on whether such equipoise exists in a given situation can often be difficult.[3]

Some clinicians might find equipoise more difficult than others,[4] and where initial reports have generated enthusiasm in the clinical community, the argument that the superiority of the new treatment is "obvious" and that a further trial would therefore be "unethical" is frequently advanced. This has led to serious problems in areas such as surgery where it has proved difficult or impossible to conduct randomised controlled trials of new techniques and devices because of strong beliefs based on weak evidence of large benefits. Therefore, the question of when an effect is so obvious that it does not require further testing has real practical importance.

There are some situations in which treatment effects are so large that bias, while perhaps having some impact on the overall effect size, is unlikely to affect the

large clinical and statistical significance of the result.[5] Although most healthcare interventions tend to provide only modest benefits,[6] there might be a subset where a very large effect (VLE) is seen.[7] If a set of conditions could be defined where it could be demonstrated that VLE sizes made it highly unlikely that the superiority of the treatment would be refuted by further trials, such trials would be wasteful of resources as well as potentially unethical.[8]

Therefore, we set out to identify trials showing a VLE (relative risk of ≤0.2 or ≥5) that were followed by a further large trial (≥200 events and ≥200 non-events) within the Cochrane Database of Systematic Reviews, and to evaluate this relative risk threshold as an empirical marker indicating that further trials are unlikely to be useful or necessary.

Methods

Definition of a VLE, index trial, and large trial

For consistency, we focused only on binary outcomes in randomised trials. We based our definition for a VLE effect on that used in previous empirical work on assessing large treatment effects.[7] Pereira and colleagues formed a definition based on the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) scale for relative risks in non-randomised data. Within this scale, relative risks of two to five are defined as large, and those greater than five as very large.[9] The relative risk was preferred over the odds ratio because the odds ratio may be substantially larger when outcomes are very common. Accepting that point estimates of effect might not provide useful information when the confidence intervals are wide and the effect is not nominally significant, we included only trials with a relative risk of five or more (or ≤0.20) that had a nominally statistically significant effect based on a Fisher exact test (P<0.05).

The index trial was defined as a trial with a nominally statistically significant VLE that was then followed by at least one large trial. A large trial was defined a priori as one with at least 200 events and 200 non-events. This choice is arbitrary, but it selects for trials that have a very large power to detect not only a relative risk exceeding five, but also a much smaller relative risk. For example, with 200 events and 200 non-events and 1:1 ratio of participants in the compared arms, the power is 90% to detect a relative risk of 1.44 at an α of 0.05.

Empirical data

We used the Cochrane Database of Systematic Reviews (2010, issue 7) as in a previous study.[7] However, for our final dataset of included forest plots, we also manually checked whether there were any newer versions of the Cochrane review published since 2010; where such newer versions contained newer trials, our database was updated with these extra trials using the Cochrane Database of Systematic Reviews up to 2015, issue 12.

Inclusion and exclusion criteria

Quantitative summary data on treatment comparisons and outcomes are presented in forest plots. Inclusion was assessed at two stages: an automated computerised algorithm and then a manual human scrutinisation process. We further assessed forest plots that satisfied the following initial inclusion criteria by the automatic algorithm: two or more studies, VLE in one trial, the VLE had a P<0.05 by Fisher's exact test, and the trial with a VLE was followed by at least one large trial. If two or more trials were published in the same year and it was not feasible to identify which was published first, we randomly picked up one as the index trial. We included VLEs regardless of either the choice of intervention or treatment comparison.

Additional inclusion criteria during the manual scrutinisation process were:

- The VLE was explicitly defined as a primary outcome of the review in which it appeared
- The forest plot was not a sensitivity analysis
- The forest plot was not a subgroup analysis
- If two forest plots satisfying the first three criteria had overlapping trials, only the plot with the largest number of trials was included.

We excluded forest plots using outcomes measured on continuous scales and those not including the year of publication of each trial (because it would not be possible to determine if the trial was followed by a large trial). We also excluded reviews with issues preventing adequate data extraction in their structure (that is, information that could not be parsed or with inconsistent data hierarchy), methodological reviews, and protocols.

Data extraction

The primary data extraction of eligible forest plots was performed using an automated algorithm approach. Full details are described elsewhere.[7] Briefly, raw data from each of the 3545 available reviews within the 2010 issue 7 of the Cochrane Database Systematic Reviews are stored under a hierarchical structure. Python computer scripts were applied to these data to parse and extract the required information from each review. This approach has previously been validated by hand using 200 randomly selected forest plots with 100% agreement.[7] Updating of the eligible topics using 2015 issue 12 of the Cochrane Database Systematic Reviews was performed manually.

For each eligible forest plot, we automatically extracted the following characteristics: Cochrane Database Systematic Reviews identification, title, comparison, outcome, subgroup, total number of trials, year of publication of index trial, and relative risk of index trial. Two authors (MN and GK) then independently conducted the manual scrutinisation process of potentially eligible forest plots. In cases of disagreement, consensus was obtained by discussion with a third author.

There were some cases where the list of outcomes within the review was not explicitly split by the Cochrane review authors into primary and secondary. Where this occurred, the data extractors for this study made a judgment to include the forest plot if they thought that the outcome was highly likely to represent an outcome of critical or primary importance, given the

stated objective of the review. Any case where this occurred was automatically referred to the third author for arbitration. Only cases with agreement from all three authors were accepted. Further characteristics extracted in this final subset of forest plots included relative risk sizes of all subsequent large trials, numbers of events and non-events, and time lag between index and large trials.

We used two approaches to assess whether an index trial VLE was upheld or refuted by subsequent large trials. Firstly, we deemed a VLE refuted if at least one subsequent large trial presented a statistically significant effect in the opposite direction or a non-significant result. Secondly, if more than one large trial followed the index trial, we performed a fixed effect meta-analysis of all large trials to assess whether this effect estimate refuted the VLE (that is, a statistically significant effect in the opposite direction or a non-significant result).

Risk of bias assessment by the Cochrane reviewers was manually extracted for all index trials.[10] Specifically, number of bias domains rated at low risk and total number of bias domains assessed were extracted. An index trial was classified as being at low risk of bias if all domains were rated at low risk.

Statistical analysis

Descriptive statistics are expressed as medians with interquartile ranges or absolute counts and percentages. The magnitude of effect was captured by the relative risk metric, but the absolute risk difference is also presented for comparability. Because most of the index trials were small with few events and non-events, we calculated 95% confidence intervals for the relative risks via an exact approach.[11] This method has been shown to provide confidence intervals with better coverage probability than asymptotic methods when samples sizes are small.[11] For the absolute risk difference, we calculated 95% confidence intervals using the Woolf method.[12] For larger trials, we computed P values and 95% confidence intervals using asymptotic approaches.

Comparisons between independent groups were performed with Fisher's exact, Mann-Whitney U, and Kruskal-Wallis tests, as appropriate. Data analyses were performed using Stata (version 12.1, Stata Corp) and R 3.1.0 (R Core Team, 2014, www.R-project.org/). All P values were two tailed with nominal statistical significance claimed for $P<0.05$.

Patient involvement

Patients were not involved in any aspect of the study design, conduct, or in the development of the research question or outcome measures. This study was a meta-epidemiological assessment of existing published research and therefore there was no active patient recruitment for data collection.

Results

Selection of forest plots for analysis

Of 3545 reviews within the Cochrane Database Systematic Reviews up to issue 7 in 2010, 3082 reviews provided 85 002 forest plots for investigation. Of these forest plots, 294 (0.35%) satisfied the computerised selection algorithm for containing at least one trial with a nominally statistically significant VLE (index trial) followed by at least one further trial with at least 200 events and 200 non-events (a large trial). Figure 1 summarises the flow of forest plots through the selection process.

From these 294 plots, in-depth scrutiny was performed (fig 1) to exclude non-eligible ones. Before arbitration, the initial κ score between the two authors was 0.85 (95% confidence interval 0.77 to 0.92). After discussions between the two authors and arbitration with the third author, consensus was reached on inclusion of 44 plots for final inclusion (0.05% of the 85 002 plots assessed by the computerised algorithm).

Baseline characteristics of eligible forest plots

Table 1 presents baseline characteristics of the index trial, forest plot, and subsequent large trials. The relative risks displayed in table 1 have been consistently coined so that all are above one (that is, a relative risk of 0.2 becomes 5). The median relative risk was 7.95 (interquartile range 5.5-12.8; range 5.0-48.6). Obstetrics and gynaecology was the most well represented specialty with 10 topics. Index trials were generally small with a median of 14 events and 91 non-events. 21 topics had an updated version of the Cochrane Database Systematic Review after 2010. The updates contributed one new trial each to two topics, and four new trials each to two topics.

Few index trials were rated at low risk of bias (9/44; 20%) and very few forest plots assessed mortality (7/44; 16%). The median proportion of events contributed by an index trial to its forest plot was 1.4% (interquartile range 0.6-3.0). Most forest plots had an I^2 statistic suggesting

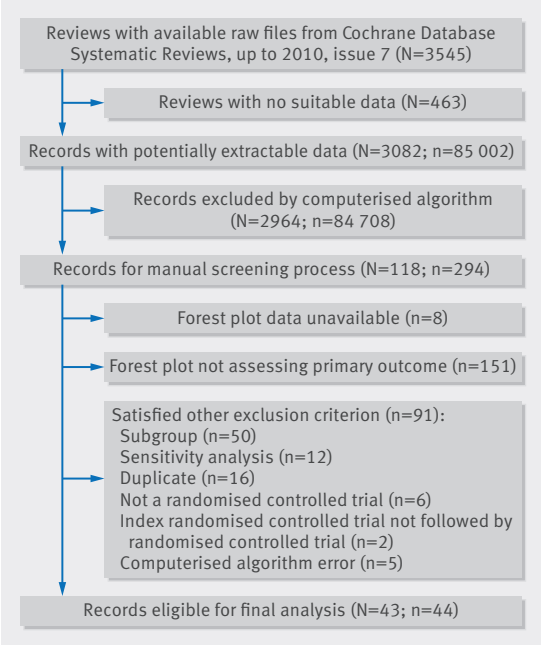

Fig 1 | Flow of records through the selection process. N=Cochrane reviews, n=forest plots.

Table 1 | Summary characteristics of included forest plots, both overall and by upheld/refuted status. Data are median (interquartile range) unless otherwise stated

Characteristic	Forest plots		
	Overall (n=44)	Upheld (n=25)	Refuted (n=19)*
Index trial, relative risk	7.95 (5.53-12.78)	7.58 (5.40-10)	8.49 (5.61-13)
Year of index trial	1994 (1987-2000)	1994 (1987-2000)	1993 (1986-99)
Index trials with mortality outcome (No (%))	7 (16)	1 (2)	6 (14)
Index trials at low risk of bias (No (%))	9 (20)	6 (14)	3 (7)
Sample size of index trial	99 (57-204)	160 (80-319)	66 (52-103)
No of events in index trial	14 (9-27)	21 (13-36)	9 (7-13)
No of non-events in index trial	91 (44-180)	131 (44-306)	51 (43-97)
No of studies in forest plot	17 (9-22)	11 (7-17)	19 (16-24)
Percentage of events contributed by index trial	1.4 (0.6-3.0)	2.0 (1.0-4.6)	0.8 (0.5-1.7)
Percentage of non-events contributed by index trial	1.6 (0.8-4.4)	2.3 (0.9-7.0)	1.1 (0.7-2.7)
I^2 percentage heterogeneity in forest plots	49 (20-66)	55 (32-83)	40 (14-57)
No of large trials	1 (1-2)	1 (1-1)	2 (1-3)
No of events in largest trial	320 (243-485)	312 (240-492)	368 (250-446)
No of non-events in largest trial	1020 (428-3971)	944 (431-3865)	1385 (408-4501)
Percentage of events contributed by largest trial	37 (19-54)	39 (22-54)	35 (14-47)
Percentage of non-events contributed by largest trial	33 (18-45)	36 (14-60)	29 (18-43)
No of years between index trial and largest trial	6 (3-12)	6 (3-13)	6 (4-12)

*Refuted by at least one subsequent large trial.

moderate heterogeneity (median 49% (interquartile range 20-66)). The median number of studies was 17 (interquartile range 9-22). There was a median of six years (interquartile range 3-12) between the index trial and the largest trial, and the largest trial had a median of 320 events (interquartile range 243-485). The median proportion of events contributed by the largest trials to the forest plot was 37% (interquartile range 19-54).

Comparison of index trials to subsequent large trials

At least one subsequent large trial refuted the index trial VLE in 19 of 44 cases (43%, 95% confidence interval 29% to 58%). The relative risk was closer to the null in the subsequent large trials in 43 of 44 cases. Of the 44 forest plots, 27 had only one subsequent large trial, nine plots had two subsequent large trials, and eight plots had three or more subsequent large trials. In the 17 plots with at least two subsequent large trials, the fixed effect meta-analysis of all subsequent large trials upheld the index result in six cases and refuted it in 11. In the 17 cases where both approaches (that is, at least one subsequent large trial refuting VLE versus fixed effect meta-analysis of all subsequent large trial data) could be directly compared, there was agreement in 16 cases (11 both refuted, five both upheld). One case was refuted by at least one large trial but upheld by the meta-analysis of all large trials. Index trials that were upheld by a large trial had a higher median number of events than those that were refuted (21 *v* 9, P<0.01). Of the 19 plots where an index trial VLE was refuted, there were two cases in which the large trial data presented a statistically significant effect in the opposite direction to the index trial.

Figure 2 plots the index trial VLE size against the coined (that is, all >1) relative risk in subsequent large trial data. Even with a stricter cutoff value in relative risk of at least 10 (or ≤0.1), only six of 13 index trial VLEs were upheld. Figure 3 plots the index trial P value against the size of coined (that is, all >1) relative risk of large trial data. Most refuted cases occurred when the index trials had a P value between 0.05 to 0.001. If the index trial had a P value of less than 0.001, the effect was upheld in 19 of the 21 cases by subsequent large trials. Table 2 demonstrates the positive predictive value with our data for a range of different cutoff values of relative risks and P values. Confidence intervals for the estimates were extremely wide owing to the small number of cases.

Upheld index trial VLEs

Information on the 25 plots in which the index trial VLE was upheld by subsequent large trial data is displayed in table 3 (for both relative risk and absolute risk difference). All but three of the 25 interventions were compared with an inactive control rather than another active treatment. The vast majority of forest plots also pertained to primary outcomes that are unlikely to be the only primary outcome of interest that might dictate whether the intervention is adopted (that is, specific adverse events or surrogate laboratory measures as

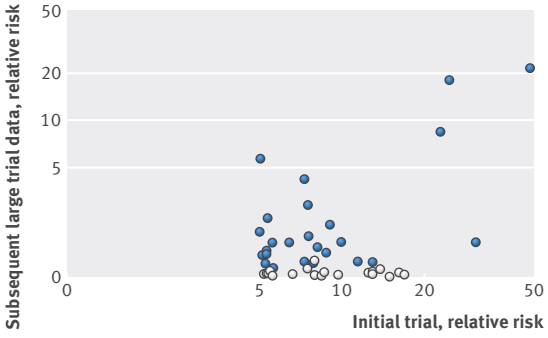

Fig 2 | Relative risk for index trial versus subsequent large trial data. VLE=very large effect; light dots=refuted index trial VLEs according to the a priori definition; dark dots=upheld index trial VLEs. Relative risks have been consistently coined so that all are above one (that is, a relative risk of 0.2 becomes 5)

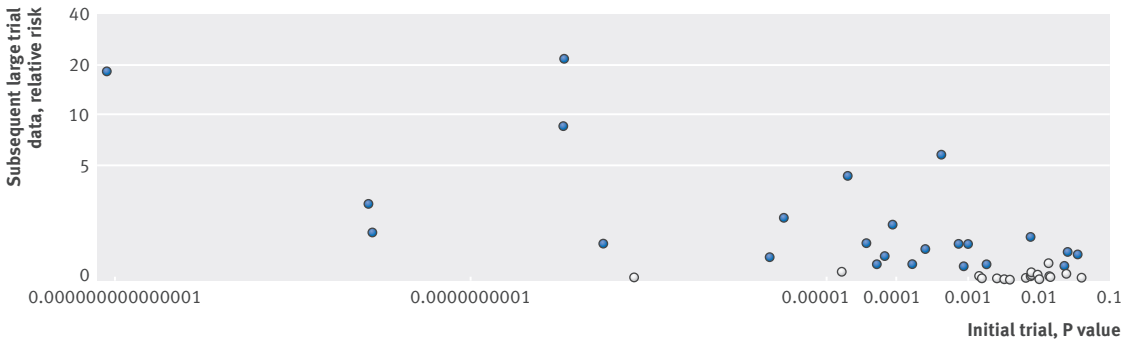

Fig 3 | Index trial P value versus subsequent large trial data effect size. VLE=very large effect; light dots=refuted index trial VLEs based on the a priori definition; dark dots=upheld index trial VLEs. Relative risks have been consistently coined so that all are above one (that is, a relative risk of 0.2 becomes 5)

opposed to hard clinical endpoints). There was large variability of the absolute risk differences across the index trials (range 0.01-0.89) and across the upholding subsequent large trials (range 0.00-0.94). Even among the 25 upheld topics, only eight had an absolute risk difference exceeding 10%.

In only four of the 25 cases was the confirmatory large trial effect also very large. There was only one subsequent large trial in each case. In three cases, a treatment of known effectiveness for the outcome measure was compared with placebo (hepatitis B antibody seroconversion with hepatitis B vaccine; rise in haemoglobin with iron in pregnant women; improvement in postoperative pain with rofecoxib).[13-15] In the other case, the comparator intervention was practically the outcome, giving a control value of 100% (misoprostol *v* surgery in women with miscarriage; outcome: surgical evacuation of the fetus).[16] Given the choice of controls and outcomes used, these results are unsurprising.

Of the index trial VLEs that were upheld, only one pertained to mortality. This plot assessed the effect of early nitrate anti-hypertensive treatment on all cause mortality up to day two in patients with an acute cardiovascular event. There was, however, no statistically significant benefit to early nitrate treatment at the co-primary outcome time points of days 3-10 and day 30 onwards. Hence, the forest plot providing evidence on mortality up to day two did not translate to any tangible, lasting clinical benefit.

Table 2 | Positive predictive values with various cutoff values for relative risks and P values for index trials

Relative risk	P value	No of forest plots	Index trial VLE upheld	Index trial VLE refuted	Positive predictive value (%, 95% CI)
≥5	<0.05	44	25	19	57 (41 to 72)
≥5	<0.01	35	22	13	63 (45 to 79)
≥5	<0.001	21	19	2	90 (70 to 99)
≥5	<0.0001	15	13	2	87 (60 to 98)
≥5	<0.00001	9	8	1	89 (52 to 100)
≥10	<0.05	14	7	7	50 (23 to 77)
≥15	<0.05	7	4	3	57 (18 to 90)
≥20	<0.05	4	4	0	100 (40 to 100)
≥30	<0.05	2	2	0	100 (16 to 100)
≥40	<0.05	1	1	0	100 (3 to 100)

VLE=very large effect.

Additional tables are available in the online supplementary appendix. Table S1 contains the references to Cochrane Reviews for table 3, table S2 contains the equivalent data of table 3 but for refuted studies, and table S3 details how many large trials followed each index trial VLE and the number of large trials that refuted the index trial VLE.

Discussion

Principal findings

In this study, there were only rare instances where an initial very large treatment effect in a trial from a primary outcome forest plot was followed by a large trial (0.05% of more than 85 000 binary outcome forest plots within the Cochrane Database). Most VLEs occurred in small studies with very few events. Just over half of the VLEs were subsequently upheld as nominally statistically significant by a subsequent large trial, although typically the effect estimate was heavily attenuated. Even when the effect was upheld, the specific primary outcome was often one of many primary outcomes that would have to be considered before adopting the intervention. Furthermore, it would be important to also consider the absolute risk reduction in deciding whether a treatment was to be used.

Our main objective in this study was to evaluate the usefulness of a VLE in a randomised controlled trial as an empirical marker that further trials were unlikely to be necessary. Theoretically, this kind of empirical marker could highlight where resources might be wasted on unnecessary follow-up trials. The scale of waste within the research process has been well acknowledged.[8 17] Unfortunately, our results show that a simple rule of thumb based on relative risk size in randomised controlled trials appears impractical, given the low frequency of VLEs and the positive predictive value of the rule. In nearly half of cases, the rule we chose would have given an incorrect reassurance, but the rarity of VLEs would strictly limit its usefulness in any case.

Comparison with other studies

The previous empirical evaluation of very large treatment effects in the literature, by Pereira and colleagues,[7] demonstrated that most of these effects

Table 3 | Intervention and comparator information from Cochrane reviews for upheld index trial VLEs

			Index trial		Large trial data	
Population	Intervention	Outcome	Relative risk (95% CI)	Absolute risk difference (95% CI)	Relative risk (95% CI)	Absolute risk difference (95% CI)
Patients at the stage of pre- exposure or post-exposure of hepatitis A (infectious hepatitis)	No intervention, placebo, or control *v* immunoglobulins	Number with hepatitis A or infectious hepatitis at 6-12 months	5.03 (3.18 to 7.94)	0.01 (0.00 to 0.01)	1.94 (1.74 to 2.17)	0.00 (0.00 to 0.01)
Pregnant women	Daily iron *v* placebo/no intervention	Haemoconcentration (that is, Hb rise) during second or third trimester	5.08 (1.87 to 14.64)	0.15 (0.07 to 0.24)	5.66 (4.26 to 7.52)	0.57 (0.52 to 0.63)
Children with asthma at risk of asthma related admission to the emergency department	Control *v* education (any type)	Emergency department visits	5.17 (2.35 to 11.76)	0.31 (0.19 to 0.43)	1.37 (1.12-1.68)	0.15 (0.06 to 0.24)
Adult smokers	Motivational interviewing *v* brief advice/usual care	Abstinence at maximal follow-up	5.28 (1.75 to 17.92)	0.15 (0.07 to 0.23)	1.21 (1.05 to 1.39)	0.03 (0.01 to 0.05)
Adults needing surgery	Control *v* aprotinin (anti-fibrionolytic)	Number exposed to allogeneic blood	5.33 (1.10 to 90.83)	0.54 (0.16 to 0.93)	1.47 (1.31-1.66)	0.17 (0.12 to 0.23)
Patients with cancer and neutropenia	Intervention *v* quinolone antibiotic	Febrile episodes	5.33 (2.04 to 17.49)	0.62 (0.38 to 0.85)	1.39 (1.11-1.74)	0.05 (0.02 to 0.09)
Adults with rheumatoid arthritis	Abatacept *v* placebo	ACR 50% improvement	5.40 (2.32 to 14.04)	0.17 (0.11 to 0.22)	2.37 (1.72 to 3.26)	0.23 (0.16 to 0.30)
Adults with type 2 diabetes mellitus	Rosiglitazone *v* control	Oedema	5.61 (1.79 to 18.36)	0.09 (0.04 to 0.14)	1.65 (1.34 to 2.04)	0.06 (0.03 to 0.08)
Children living in malaria endemic regions	Placebo *v* intermittent preventive treatment with antimalarial	Clinical malaria	5.64 (4.10 to 7.76)	0.34 (0.29 to 0.39)	1.13 (1.01 to 1.26)	0.07 (0.01 to 0.13)
Adults receiving intensive care	No prophylaxis *v* topical plus systemic antibiotic	Respiratory tract infections	6.46 (2.86 to 15.87)	0.56 (0.40 to 0.72)	1.66 (1.36 to 2.02)	0.23 (0.14 to 0.31)
Adults undergoing heart surgery	Control *v* all treatments	Postoperative atrial fibrillation	7.32 (2.08 to 35.41)	0.34 (0.19 to 0.50)	1.25 (1.05 to 1.48)	0.08 (0.02 to 0.14)
Adults with migraine pain of moderate to severe intensity	Rizatriptan *v* placebo	Pain-free response at 2 h	7.32 (2.45 to 25.79)	0.22 (0.12 to 0.32)	4.23 (2.96 to 6.03)	0.32 (0.26 to 0.38)
Adult pregnant women	Placebo/control *v* any antibiotic	Failure of test of cure of bacterial vaginosis	7.58 (3.47 to 18.72)	0.75 (0.60 to 0.89)	2.87 (2.54 to 3.24)	0.37 (0.34 to 0.41)
Children at risk of whooping cough	Whole cell vaccine *v* acellular vaccine	Primary series non-completion due to adverse events	7.59 (1.63 to 42.27)	0.05 (0.00 to 0.11)	1.82 (1.43 to 2.33)	0.00 (0.00 to 0.00)
Patients with acute cardiovascular event	Control *v* nitrates	All cause mortality at 2 days	7.90 (1.27 to 99.80)	0.04 (0.01 to 0.08)	1.22 (1.11 to 1.35)	0.00 (0.00 to 0.01)
Adult smokers	Nicotine replacement therapy (patch) *v* placebo/control	Smoking cessation at >6 months' follow-up	8.21 (2.18 to 39.39)	0.18 (0.08 to 0.28)	1.54 (1.34 to 1.77)	0.07 (0.05 to 0.09)
Kidney transplant recipients	Placebo or no treatment *v* IL2Ra	Biopsy proven acute rejection at 1 year	8.80 (1.17 to ∞)	0.36 (0.09 to 0.63)	1.43 (1.14 to 1.79)	0.08 (0.03 to 0.13)
Children	Placebo *v* rotarix (RV1) vaccine	Rotavirus diarrhoea of any severity up to 1 year	9.08 (2.41 to 42.30)	0.15 (0.07 to 0.23)	2.15 (1.75 to 2.63)	0.06 (0.05 to 0.08)
Pregnant women in spontaneous preterm labour	Placebo *v* all betamimetics	Birth within 48 h of treatment	10.00 (1.99 to 183.52)	0.60 (0.33 to 0.87)	1.66 (1.30 to 2.12)	0.14 (0.08 to 0.21)
Adults needing first line anti-hypertensive treatment	Placebo *v* high dose thiazide	Total cardiovascular events	11.47 (1.97 to 204.42)	0.14 (0.05 to 0.23)	1.25 (1.03 to 1.51)	0.01 (0.00 to 0.01)
Patients undergoing general anaesthesia, regional anaesthesia, or sedation	Placebo *v* ondansetron	Nausea	13.00 (2.57 to 224.97)	0.60 (0.37 to 0.83)	1.24 (1.02 to 1.52)	0.10 (0.01 to 0.19)
Adults with moderate to severe postoperative pain	Rofecoxib *v* placebo	>50% pain relief over 4-6 h	23.00 (4.42 to 393.80)	0.69 (0.52 to 0.85)	8.40 (5.48 to 12.87)	0.55 (0.50 to 0.61)
Women being treated for spontaneous miscarriage	Surgery *v* misoprostol	Surgical evacuation	24.79 (9.79 to 67.48)	0.89 (0.82 to 0.96)	18.00 (10.39 to 31.21)	0.94 (0.90 to 0.97)
Adults with type 2 diabetes mellitus	Pioglitazone *v* control	Oedema	30.98 (3.93 to ∞)	0.04 (0.02 to 0.06)	1.67 (1.47 to 1.88)	0.09 (0.07 to 0.11)
Patients with chronic renal failure who are seronegative for HBsAg	Plasma vaccine *v* placebo	Seroconversion to anti-HBs	48.64 (6.47 to ∞)	0.36 (0.25 to 0.47)	21.43 (11.83 to 38.84)	0.35 (0.31 to 0.38)

ACR=American College of Rheumatology score; Hb=haemoglobin; HBs=hepatitis B virus (surface) antibodies; HBsAg=hepatitis B virus surface antigen; IL2Ra= interleukin-2 receptor alpha chain; VLE=very large effect.

doi: 10.1136/bmj.i5432 | *BMJ* 2016;355:i5432 | the bmj

represented regression to the mean, with subsequent trials usually reporting smaller effects. We used the same data source and a similar initial extraction approach, with a focus on the feasibility of an empirical rule for the reliability of VLEs. So far, there has been no empirical evaluation of such a rule, although Glasziou and colleagues discussed the circumstances under which observational evidence might be accepted when the signal (effect size) to noise (bias) ratio is large.[5] They suggested that relative risks beyond 10 are highly likely to reflect real treatment effects, even if confounding factors associated with the treatment may have contributed to the size of the observed associations. More stringent criteria, such as a relative risk cutoff value of at least 10 (or ≤0.1), led to an even poorer positive predictive value of only 50% (seven of 14 cases) in our data (table 2). While another selection criterion—the presence of P<0.001 in the index trial—improved the positive predictive value substantially, there were still cases where the subsequent large trial did not uphold the index trial's finding.

Conclusions and policy implications

Our findings show that even a relative risk of five is a rare event, and mostly occur in small trials with large confidence intervals. Because index trials with VLEs for primary outcomes are so rare, attempts to improve the positive predictive value by making the criteria more stringent would effectively rule out nearly all trials (eg, only four trials that we assessed had a relative risk of ≥20 and ≤0.05). Even when these criteria are satisfied, issues of heterogeneity in treatment effect could still mean that the results apply only to a narrow population and therefore need further trials in different patient groups or circumstances.[18]

Methodological problems in interpreting the results of small studies have been well documented.[19 20] Reversals in the medical literature, even for randomised controlled trials, are common.[21 22] Therefore, it might actually be dangerous to consider a case open and shut after a single trial with a VLE. A more important practical lesson from this study could be that the place of small randomised controlled trials needs re-evaluation. If even very large treatment effects in small trials are unreliable evidence of significant benefit, perhaps we should avoid conducting small trials (unless explicitly justified for any case specific reason—eg, rare diseases) and aim instead to conduct studies that are larger and properly powered to detect modest effects. This has serious implications for complex interventions such as surgery, where large randomised controlled trials are known to be more difficult to deliver.[23]

Strengths and limitations of study

Using the large number of forest plots available within the Cochrane Database as a source of data was a major strength of our work given the rarity of VLEs. Furthermore, our systematic approach to obtaining a set of independent VLEs and assessing them under a range of possible cutoff values for relative risks and P values also lends further credence to our conclusion that an empirical rule using a VLE would be neither practical nor useful.

However, our findings must be considered in light of several limitations. Firstly, our definition of a VLE, while based on previous empirical work,[7] necessarily imposes an arbitrary cutoff value on a continuum. Our stringent rule left very few eligible topics compared with the vast number of topics handled by Cochrane. One might speculate whether a more lenient rule would change our inferences. However, if anything, smaller effects are likely to be even less commonly upheld than the VLE that we studied.

Secondly, we considered effects in the context of the primary outcome of the Cochrane review in which they appeared rather than the primary outcome of the trial itself, mainly on logistical grounds. Thirdly, clinical and statistical significance are not synonymous. There might be statistically significant upheld effects that are attenuated in size to a point where they lose clinical significance, and vice versa.

Fourthly, it was difficult to accurately ascertain whether an effect pertained to a subgroup or sensitivity analysis where such analyses were not explicitly defined in the Cochrane review. We attempted to ensure objectivity by using a review process involving two independent authors and discussion with a third author in cases of ambiguity. Fifthly, while the Cochrane Database Systematic Reviews represent a considerable body of trial meta-analyses, it nonetheless provides imperfect coverage of the entire body of randomised trial evidence. However, there is no obvious reason to believe that non-covered topics are likely to be substantially different about VLE prevalence and validation.

Finally, the decision to perform a subsequent large trial when a VLE has been seen in one trial is not a random process. Trials with VLEs might be less likely to have subsequent large trials done on the same question, if the early trials are considered to be well done and their findings are deemed conclusive. If so, our data underestimate the proportion of VLEs that are true. However, subsequent large trials might be less likely to be performed if the original trial results are thought to lack credibility or be unreliable. If so, our data overestimate the proportion of VLEs that are true. Given that the early trials showing VLEs are almost ubiquitously very small ones, it is more likely that our data overestimate the proportion of VLEs.

Summary

Our study suggests that the frequency of VLEs followed by a large trial is vanishingly small in the Cochrane Database of Systematic Reviews, and where they occur they do not appear to be a reliable marker for a reproducible and clinically actionable benefit. An empirical rule using a VLE as a marker that further trials are unnecessary would be neither practical nor useful. Caution should be taken when interpreting small studies with very large treatment effects.

Contributors: PM and JPAI conceived the study. MN, TVP, GK, and MM extracted and sorted data for the study. MN and TP performed the analysis. MN wrote the first draft of the manuscript. All authors

contributed to critical revision of the manuscript for important intellectual content and approved the final version. MN and PM are the guarantors.

Funding: No specific funding was provided for this study.

Competing interests: All authors have completed the ICMJE uniform disclosure at www.icmje.org/coi_disclosure.pdf and declare no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: No ethical approval required as a meta-epidemiological study.

Data sharing: Raw data and analysis available on request from the authors.

The lead authors affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

1 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887-92. doi:10.1056/NEJM200006223422507.
2 Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869. doi:10.1136/bmj.c869.
3 Kurzrock R, Stewart DJ. Equipoise abandoned? Randomization and clinical trials. *Ann Oncol* 2013;24:2471-4. doi:10.1093/annonc/mdt358.
4 McCulloch P, Kaul A, Wagstaff GF, Wheatcroft J. Tolerance of uncertainty, extroversion, neuroticism and attitudes to randomized controlled trials among surgeons and physicians. *Br J Surg* 2005;92:1293-7. doi:10.1002/bjs.4930.
5 Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007;334:349-51. doi:10.1136/bmj.39070.527986.68.
6 Djulbegovic B, Kumar A, Glasziou P, Miladinovic B, Chalmers I. Medical research: Trial unpredictability yields predictable therapy gains. *Nature* 2013;500:395-6. doi:10.1038/500395a.
7 Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 2012;308:1676-84. doi:10.1001/jama.2012.13444.
8 Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet* 2009;374:86-9. doi:10.1016/S0140-6736(09)60329-9.
9 Guyatt GH, Oxman AD, Sultan S, et al. GRADE Working Group. GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol* 2011;64:1311-6. doi:10.1016/j.jclinepi.2011.06.004.
10 Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0.*The Cochrane Collaboration, 2011.
11 Reiczigel J. Abonyi-Tóth Z, Singer J. An exact confidence set for two binomial proportions and exact unconditional confidence intervals for the difference and ratio of proportions. *Comput Stat Data Anal* 2008;52:5046-53doi:10.1016/j.csda.2008.04.032.
12 Woolf B. On estimating the relation between blood group and disease. *Ann Hum Genet* 1955;19:251-3. doi:10.1111/j.1469-1809.1955.tb01348.x.
13 Bulley S, Derry S, Moore RA, McQuay HJ. Single dose oral rofecoxib for acute postoperative pain in adults. *Cochrane Database Syst Rev* 2009;4:CD004604.
14 Peña-Rosas JP, Viteri FE. Effects and safety of preventive oral iron or iron+folic acid supplementation for women during pregnancy. *Cochrane Database Syst Rev* 2009;4:CD004736.
15 Schroth RJ, Hitchon CA, Uhanova J, et al. Hepatitis B vaccination for patients with chronic renal failure. *Cochrane Database Syst Rev* 2004;3:CD003775.
16 Neilson JP, Gyte GM, Hickey M, Vazquez JC, Dou L. Medical treatments for incomplete miscarriage (less than 24 weeks). *Cochrane Database Syst Rev* 2010;1:CD007223.
17 Ioannidis JP, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014;383:166-75. doi:10.1016/S0140-6736(13)62227-8.
18 Iwashyna TJ, Burke JF, Sussman JB, Prescott HC, Hayward RA, Angus DC. Implications of Heterogeneity of Treatment Effect for Reporting and Analysis of Randomized Trials in Critical Care. *Am J Respir Crit Care Med* 2015;192:1045-51. doi:10.1164/rccm.201411-2125CP.
19 IntHout J, Ioannidis JP, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. *J Clin Epidemiol* 2015;68:860-9. doi:10.1016/j.jclinepi.2015.03.017.
20 Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;294:218-28. doi:10.1001/jama.294.2.218.
21 Prasad V, Cifu A, Ioannidis JP. Reversals of established medical practices: evidence to abandon ship. *JAMA* 2012;307:37-8. doi:10.1001/jama.2011.1960.
22 Prasad V, Vandross A, Toomey C, et al. A decade of reversal: an analysis of 146 contradicted medical practices. *Mayo Clin Proc* 2013;88:790-8. doi:10.1016/j.mayocp.2013.05.012.
23 Rerkasem K, Rothwell PM. Meta-analysis of small randomized controlled trials in surgery may be unreliable. *Br J Surg* 2010;97:466-9. doi:10.1002/bjs.6988.

Appendix: Online supplemental data