A re-analysis of 150 women's health trials to investigate how the Bayesian approach may offer a solution to the misinterpretation of statistical findings

Hemming, K.; Melo, P.; Luo, R.; Taljaard, M.; Coomarasamy, A.

DOI:
10.1111/1471-0528.17570

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Hemming, K, Melo, P, Luo, R, Taljaard, M & Coomarasamy, A 2023, 'A re-analysis of 150 women's health trials to investigate how the Bayesian approach may offer a solution to the misinterpretation of statistical findings', *BJOG: An International Journal of Obstetrics & Gynaecology*. https://doi.org/10.1111/1471-0528.17570

Link to publication on Research at Birmingham portal

Accepted: 26 May 2023

DOI: 10.1111/1471-0528.17570

RESEARCH ARTICLE

A re-analysis of 150 women's health trials to investigate how the Bayesian approach may offer a solution to the misinterpretation of statistical findings

K. Hemming[1] | P. Melo[2] | R. Luo[3,4] | M. Taljaard[5,6] | A. Coomarasamy[2]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK

[2]Tommy's National Centre for Miscarriage Research, Institute of Metabolism and Systems Research, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, UK

[3]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada

[4]OMNI Research Group, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[5]Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada

[6]School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Ontario, Canada

Correspondence
K. Hemming, Institute of Applied Health Research, University of Birmingham, Public Health Building, University of Birmingham, Birmingham B15 2TT, UK.
Email: k.hemming@bham.ac.uk

Funding information
UK NIHR Collaborations for Leadership in Applied Health Research and Care West Midlands initiative

Abstract

Objective: To investigate whether a Bayesian interpretation might help prevent misinterpretation of statistical findings and support authors to differentiate evidence of no effect from statistical uncertainty.

Design: A Bayesian re-analysis to determine posterior probabilities of clinically important effects (e.g., a large effect is set at a 4 percentage point difference and a trivial effect to be within a 0.5 percentage point difference). Posterior probabilities greater than 95% are considered as strong statistical evidence, and less than 95% as inconclusive.

Sample: 150 major women's health trials with binary outcomes.

Main Outcome Measures: Posterior probabilities of large, moderate, small and trivial effects.

Results: Under frequentist methods, 48 (32%) were statistically significant (p-value ≤ 0.05) and 102 (68%) statistically non-significant. The frequentist and Bayesian point estimates and confidence intervals showed strong concordance. Of the statistically non-significant trials (n = 102), the Bayesian approach classified the majority (94, 92%) as inconclusive, neither able to confirm or refute effectiveness. A small number of statistically non-significant findings (8, 8%) were classified as having strong statistical evidence of an effect.

Conclusions: Whilst almost all trials report confidence intervals, in practice most statistical findings are interpreted on the basis of statistical significance, mostly concluding evidence of no effect. Findings here suggest the majority are likely uncertain. A Bayesian approach could help differentiate evidence of no effect from statistical uncertainty.

KEYWORDS
Bayesian interpretation, randomised controlled trials, statistical significance

1 | BACKGROUND

Randomised trials are the backbone of evidence-based medicine, and over the past decades the quality of their implementation has improved and risk of bias decreased.[1] Each day the reports of more than 75 randomised trials are published and this number is increasing year on year.[2] For a well-conducted, well-reported randomised trial, correct interpretation of its findings is essential to ensure that only truly effective interventions are adopted, truly ineffective interventions are disregarded and, where uncertainty exists, this is acknowledged and recognised.[3]

Interpretation of trial findings depends on context, risk of bias, other scientific evidence and, importantly, the primary or other key outcome results.[4,5] Frequentist statistical approaches are most commonly used in practice, which, for example in the case of a binary outcome, includes reporting the proportions with the outcome of interest in each arm, a corresponding absolute or relative difference and its 95% confidence interval.[6] Reporting guidelines, such as the CONSORT statement, include confidence intervals as a minimum reporting requirement.[7] Unfortunately, many researchers ultimately interpret the primary and other key outcomes based on whether the confidence interval includes the null – and are thus implicitly reverting to interpretation based on statistical significance.[8,9]

The following two case studies illustrate the problem. The INFANT trial ($n = 46\,614$) evaluated the use of computerised interpretation of cardiotocographs on the occurrence of adverse neonatal outcomes.[10] The primary outcome occurred in 171/23 351 (0.73%) in the treatment arm versus 172/23 263 (0.74%) in the control. The difference was not statistically significant, with a risk ratio of 1.01, 95% confidence interval (CI) 0.82–1.25. In the conclusion, this result was interpreted as 'continuous electronic fetal monitoring in labour does not improve clinical outcomes'. This might be construed as misinterpretation of statistical significance.[6] However, inspection of the finding on the risk difference scale (percentage points risk difference = −0.00, 95% CI −0.16 to 0.15) reveals any difference in outcomes is almost certainly smaller than half a percentage point (upper bound of 95% confidence interval 0.15 percentage points) in adverse neonatal outcomes. Yet, implicit in this interpretation is that this small reduction is not clinically important. Thus, not reporting effect sizes on a clinically interpretable scale and not explicitly interpreting the range of effect sizes supported by the confidence interval, the mechanism by which this conclusion was reached is not transparent. This is problematic, because it lends itself to a perpetuation of misinterpretation in other smaller trials, as well as making assumptions about sizes of effects that are clinically important without making this explicit.

To illustrate how non-statistically significant results are often misinterpreted, we consider a second case study. This trial compared titrated-dose oral misoprostol (intervention) with static-dose to increase the likelihood of a vaginal birth. The risk ratio was 0.98 (95% CI 0.77–1.24) based on 47/73 events (64%) in the treatment arm and 48/73 (66%) events in the control arm.[11] Similar to the first case study, this primary outcome result was interpreted as evidence of 'similarity'. Yet, in this trial the difference in percentage points was –1.36 (95% CI –16.83 to 14.09). This confidence interval indicates considerable uncertainty, providing evidence that this intervention might either increase or reduce this outcome by a considerable amount. Thus, in this example, the interpretation of the primary outcome as showing evidence of 'similarity' is highly misleading – a more accurate interpretation is that unfortunately the study is too small to tell us anything conclusive.

These case studies illustrate how non-statistically significant outcomes can be misinterpreted as evidence of no effect; and, moreover, even when results are sufficiently precise to rule out clinically important effects, trialists still persist in interpreting key outcomes based on statistical significance.[12-14]

2 | OBJECTIVES

To illustrate how a Bayesian approach might help mitigate some of the problems around the misinterpretation of statistical findings, we undertook a Bayesian re-analysis of a contemporary sample of women's health randomised trials with binary primary outcomes. We first illustrate how the Bayesian and frequentist analyses show strong concordance. We then formulate a mechanism for how a Bayesian interpretation can be implemented by introducing the concept of the strength of statistical evidence and clinically important effect sizes. We illustrate the approach for an example set of large, moderate, small and trivial effect sizes (as well as unanticipated harm), and varying degrees of strength of statistical evidence. We contrast the interpretation of the Bayesian analysis with that from a frequentist interpretation.

3 | METHODS

3.1 | Search strategy

We identified individually randomised, two-arm superiority trials (1:1 randomisation ratio) with a binary primary outcome, whose primary report of findings was published in one of seven English language high-impact general medical and specialty journals, between January 2015 and December 2020: *New England Journal of Medicine*, *Lancet*, *JAMA* (*the Journal of the American Medical Association*), *BMJ*, *BJOG* (*British Journal of Obstetrics and Gynaecology*), *American Journal of Obstetrics & Gynaecology* and *Obstetrics & Gynaecology*. We included trials evaluating pharmacological and non-pharmacological interventions targeted at women to improve fertility, maternal or fetal, or perinatal outcomes. We made no restrictions on the type of comparator or setting, but excluded non-inferiority and equivalence trials. We made a post-hoc decision to exclude any trials with zero events (or 100% with the event), in either of the study arms, and studies where the primary outcome was unclear. The searches were conducted in EMBASE and MEDLINE on the Ovid platform, restricting the journal name (to one of the seven included journals) and limiting the search to randomised controlled trials published between 2015 and 2020. The list of identified studies was imported into Covidence. An initial title and abstract screen were performed, followed by a full text screen. All screening was conducted independently and in duplicate (PM and RL), with discrepancies resolved by discussion or, where needed, arbitration by a third author (KH or MT). The protocol for the review is registered on PROSPERO (PROSPERO 2021 CRD42021236171).

3.2 | Data extraction

Where available, we extracted absolute event numbers (i.e. numerators and denominators in each arm) for the primary analysis of the primary outcome; where authors only reported denominators and percentages, these were extracted instead. We also extracted the journal; intervention type, classified as pharmacological, procedural (e.g. a surgical technique or type of dressing), non-pharmacological (e.g. psychotherapy, or lifestyle changes), diagnostic or a mixture; and the primary outcome type (classified as adverse fetal outcome, adverse maternal outcome, live birth or other). We also classified each trial as to whether higher or lower event rates were desirable (e.g. reduction of adverse fetal outcomes or increased detection of adverse fetal outcome). Two authors (PM and RL) independently extracted data in duplicate and resolved any discrepancies by discussion.

3.3 | Data analysis

We used the extracted or derived number of events and total sample size for each arm to create individual level data for each trial. The contrast of interest is that of the absolute or relative difference between the proportion with the outcome in the treatment arm versus control arm. For trials in which lower event rates are desirable, negative values suggest benefit of the intervention; for trials in which higher event rates are desirable, we reversed the calculation. Thus, all absolute differences less than 0 (or relative differences less than 1) were indicative of treatment benefit.

For each trial these data were then used to estimate the risk ratio and risk difference, 95% confidence intervals (CI) and *P-values* under a frequentist approach. This analysis was implemented in STATA 17 using the *cc* function (STATA 17).[15] Any cases of non-convergence were noted. For the Bayesian analysis we estimated risk ratios, and risk differences using binomial regression with a log link, and binomial regression with an identity link, respectively. We used a vague prior throughout (normal distribution with mean zero and standard deviation 10000) to model the risk difference or log risk ratio. We report point estimates and associated 95% credible intervals (CrI). This analysis was implemented in STATA 17 using the *bayes* function with default options (Metropolis–Hastings algorithm using 12500 iterations removing the first 2500 burn-in iterations, no thinning and starting points based on iterative reweighted least-squares estimates). Again, any cases of non-convergence were noted.

We then determined, based on the Bayesian model, the posterior probabilities of a large, moderate and small beneficial effect, and evidence of at most a trivial effect. For illustration only, we defined a large beneficial effect to be a risk difference greater than (–) 4 percentage points (pp); a moderate beneficial effect as a risk difference greater than (–) 1 percentage points; a small beneficial effect as a risk difference greater than (–) 0.5 percentage points; and a trivial effect to be within 0.5 percentage points difference (either way) from the null (Figure 1). We defined an unanticipated harmful effect as a difference of at least 0.5 pp in the unanticipated direction (i.e. harm). In addition, we calculated the posterior probability of a risk difference greater than 0 percentage points ('any beneficial effect'). To estimate these posterior probabilities, we used the bayestest interval command, which uses the simulated posterior distribution of model parameters estimated using the bayes command.

These cut-points for large, moderate and small effects are used for illustration only, and we suggest in practice these be

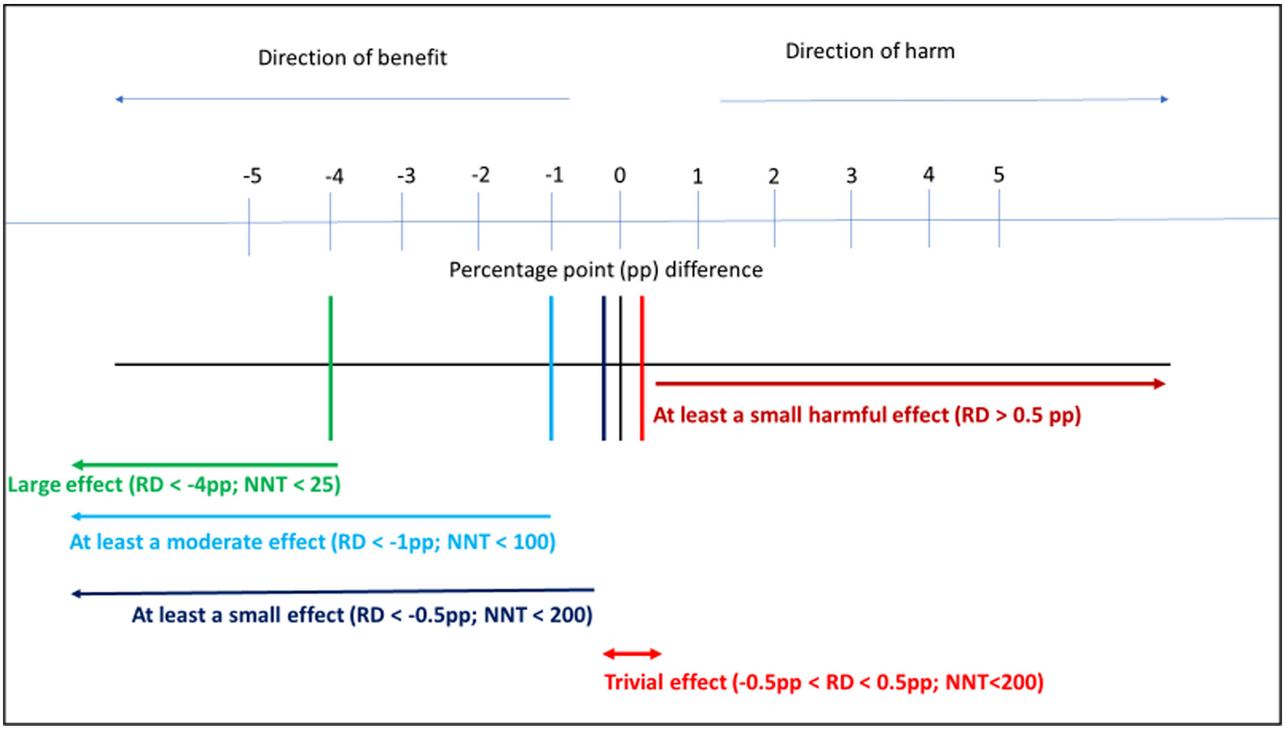

Direction of benefit
Direction of harm
-5
-4
-3
-2
-1
0
1
2
3
4
5
Percentage point (pp) difference
At least a small harmful effect (RD > 0.5 pp)
Large effect (RD < -4pp; NNT < 25)
At least a moderate effect (RD < -1pp; NNT < 100)
At least a small effect (RD < -0.5pp; NNT < 200)
Trivial effect (-0.5pp < RD < 0.5pp; NNT<200)

FIGURE 1 Proposed classification of large, small, moderate and trivial effect sizes.

grounded by effect sizes of clinical importance in the particular trial context. Working on scales that are known to be more interpretable can help to this end; consideration of effect sizes of other common interventions might also help. For example, the values we have chosen are equivalent to numbers needed to treat (NNT) of 25, 100 and 200, respectively. The use of aspirin for stroke prevention has an NNT in the region of 300 over 10 years[16]; the use of aspirin after stroke has an NNT in the region of 150 over 3 years to prevent a non-fatal heart attack;[17] whereas the use of dexamethasone in COVID-19 has an NNT in the region of 40 (RECOVERY, 2021).[18] Thus, although our choice is to some extent arbitrary and not context-specific, these values are unlikely to be very dissimilar to those chosen in practice. By evaluating 'trivial effects' we implicitly consider evidence of no benefit.

We then quantified the strength of the statistical evidence of this range of effect sizes. We suggest posterior probabilities >95% might be considered as strong statistical evidence, posterior probabilities between 90% and 94% are classified as moderate statistical evidence, and anything <90% is classified as weak statistical evidence. In a sensitivity analysis we set 97.5% as the cut-point for strong statistical evidence, 95% for moderate statistical evidence and anything <95% for weak statistical evidence. Conventionally posterior probabilities are reported without any such categorisation,[19-22] although others have also proposed categorising, for example using >80%, 90% or 95% posterior probabilities as convincing evidence.[23,24] Finally, we classified the overall statistical evidence as strong if there was strong statistical evidence of either at least a small effect (which includes moderate and larger effects), a trivial effect or an unanticipated harmful effect.

4 | RESULTS

4.1 | Characteristics of included trials

The search was performed on 4 March 2021 (Figure 2); the characteristics of the 150 trials, published between 2015 and 2020, and assessed to be eligible are summarised in Table 1. The studies were roughly evenly distributed across the seven journals, albeit with proportionately fewer published in both JAMA (12, 8%) and the BMJ (9, 6%). Most were testing a pharmacological intervention (59, 39%), a procedural intervention (48, 32%) or a non-pharmacological intervention (27, 18%). The most common outcome type was either adverse fetal (27, 18%) or adverse maternal outcomes (46, 31%). The average prevalence of the outcome (in the control arm) was 22% (interquartile range [IQR] 10–41%). The majority (98, 65%) of the trials were trying to reduce the primary outcome (e.g. reduction in adverse fetal outcome) and, in a smaller number (52, 35%), the objective was to increase the primary outcome (e.g. increase the live birth rate). For those 52 trials with an objective to increase the primary outcome, the comparisons that follow relate to control-intervention rather than intervention-control. The median number of participants randomised (total across both arms) was 503 (IQR 238–1092).

4.2 | Frequentist and Bayesian results

Of the 150 trials, approximately a third (48, 32%) were statistically significant according to our frequentist reanalysis (Table 2). Across all 150 trials under the Bayesian reanalysis, the average percentage point difference was −1.73 (IQR −7.18 to 0.77) and the average risk ratio was 0.92 (IQR 0.73–1.07) (Table 2). When estimating the risk ratio and risk difference, the occurrence of non-convergence was low (4% and 0%, respectively). The average posterior probability of any beneficial effect (risk difference ≤0) was 0.79 (IQR 0.40–0.99). The frequentist and Bayesian approaches all led to similar inferences, as indicated by the similarity of the point estimates and confidence intervals/credible intervals, and this was the case for both absolute and relative measures of effect (Figure S1, Table 2). Thus, there is a strong one-to-one alignment between the two analytical approaches.

4.3 | Classification of strength of statistical evidence

Among the 102 non-statistically significant trials, eight (8%) had strong statistical evidence whereas 94 (92%) of the studies yielded moderate or weak (posterior probability <95%) statistical evidence (Table 3). Of the eight trials classified as having strong statistical evidence, three (3%) had strong statistical evidence (posterior probability ≥95%) of at least a small benefit (NNT <200). None of these had strong statistical evidence of large benefit (NNT <25) or moderate benefit (NNT <100). A further two (2%) were classified as having strong statistical evidence of a trivial effect (percentage point difference within 0.5 of 0) and three (3%) were classified as having strong statistical evidence (posterior probability ≥95%) of harm (percentage point risk difference ≥0.5 pp).

Of those trials where the primary outcome was statistically significant, 47 (98%) were classified as having strong statistical evidence (Table 3). Moreover, many (40, 83%) were classified as having strong statistical evidence of at least a moderate effect, and many (24, 50%) were classified as having strong statistical evidence of a large effect. In addition, five (10%) of statistically significant trials were classified as having strong statistical evidence of an unanticipated harmful effect, although these would also have been interpreted as evidence of harm under the frequentist approach.

Over all 150 trials there was strong statistical evidence (posterior probability ≥95%) in around one-third of the studies (55, 37%). When we increased the stringency of the statistical evidence so that strong statistical evidence was classified as posterior probabilities >97.5%, the certainty of all conclusions decreased: the proportion of trials classified as having strong statistical evidence decreased from 37% to 29% (Table S1, Figure 3). In only two of the trials was there

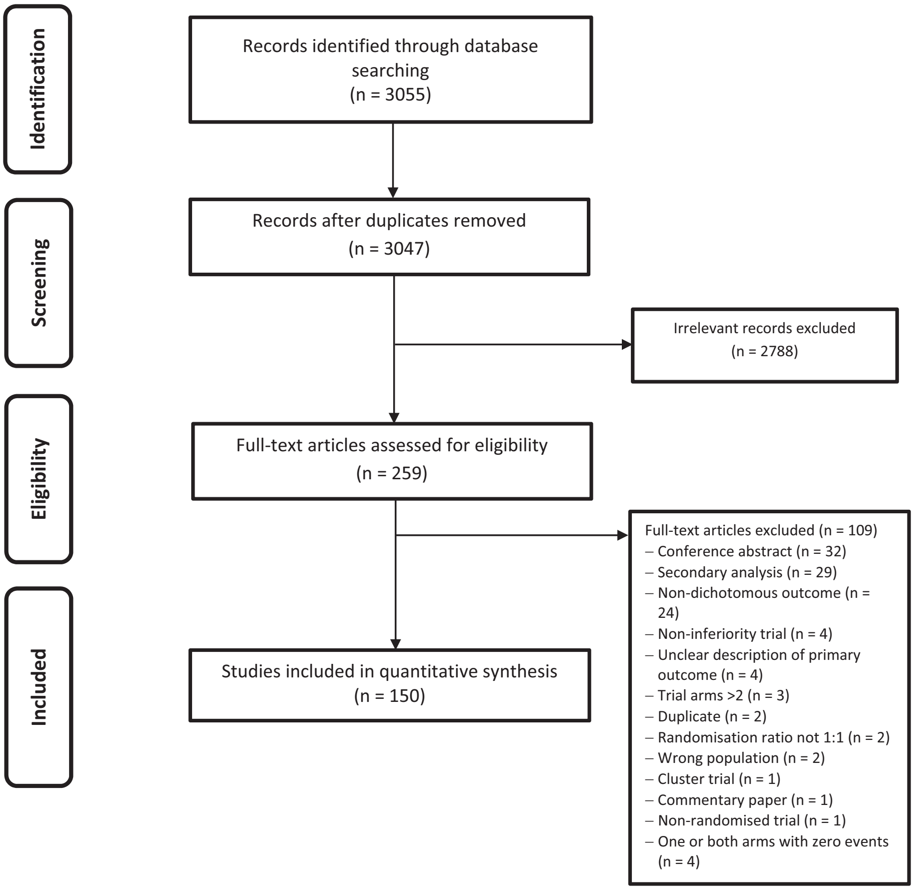

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 3055)
Records after duplicates removed (n = 3047)
Irrelevant records excluded (n = 2788)
Full-text articles assessed for eligibility (n = 259)
Full-text articles excluded (n = 109)
– Conference abstract (n = 32)
– Secondary analysis (n = 29)
– Non-dichotomous outcome (n = 24)
– Non-inferiority trial (n = 4)
– Unclear description of primary outcome (n = 4)
– Trial arms >2 (n = 3)
– Duplicate (n = 2)
– Randomisation ratio not 1:1 (n = 2)
– Wrong population (n = 2)
– Cluster trial (n = 1)
– Commentary paper (n = 1)
– Non-randomised trial (n = 1)
– One or both arms with zero events (n = 4)
Studies included in quantitative synthesis (n = 150)

FIGURE 2 PRISMA flow chart for included randomised control trials.

strong statistical evidence of a trivial effect. Of the 102 non-statistically significant trials, in only two (2%) was there strong statistical evidence (both for a trivial effect).

5 | DISCUSSION

5.1 | Summary of findings

Almost all trials fail to interpret values supported by their confidence intervals.[12-14] A Bayesian approach to interpretation might help distinguish those studies for which there is evidence the intervention does not work, from those for which the studies were probably too small and the resulting findings inconclusive. We first provide reassurance that the two analytical techniques have a strong one-to-one correspondence. Secondly, we illustrate how a Bayesian interpretation of the 102 statistically non-significant trials in this sample, can differentiate those for which there is statistical evidence of no effect (a small minority) from those for which there is considerable statistical uncertainty (the majority).

The approach requires the specification of effect sizes that are clinically important. Although this is not expected to be an easy task, the Bayesian approach is transparent about this, whereas the frequentist approach mostly ignores this is a necessary condition for interpreting confidence intervals.[25] On the other hand, the frequentist strict interpretation of *P-values* ensures that the floodgates do not open for declaring any intervention as effective – whereas the Bayesian approach proposed here might be viewed as opening the gate a little more to prevent it being shut on interventions that might well be effective.

5.2 | Research in context

5.2.1 | Detecting uncertainty

Although the 95% confidence intervals and 95% credible intervals were highly consistent across the two approaches, the Bayesian approach to interpretation allowed identification of trials with wide confidence intervals which supported both benefit and harm, and were thus inconclusive. This applied to around two-thirds of the studies in this sample. Our results underscore that many studies are under-powered to detect small but still meaningful effects (the average total sample size was in the region of 500). Returning to the second case study, the comparison of titrated-dose oral misoprostol (intervention) with static-dose oral misoprostol (control), where the reported risk ratio for the event of vaginal delivery was 0.98 (95% CI 0.77–1.24) based on 47/73 events in the treatment arm and 48/73 events in the control arm.[11] Here

TABLE 1 Characteristics of included studies.

Characteristic	All
Journal	*n* = 150
New England Journal of Medicine	24 (16.0%)
JAMA	12 (8.0%)
Lancet	23 (15.3%)
BMJ	9 (6.0%)
BJOG	25 (16.7%)
Obstetrics and Gynaecology	30 (20.0%)
American Journal of Obstetrics and Gynaecology	27 (18.0%)
Intervention type	
Pharmacological	59 (39.3%)
Procedural	48 (32.0%)
Non-pharmacological/non-procedural	27 (18.0%)
Mixed	10 (6.7%)
Diagnostic	6 (4.0%)
Outcome type	
Adverse fetal outcome	27 (18.0%)
Adverse maternal outcome	46 (30.7%)
Live birth	13 (8.7%)
Other	64 (42.7%)
Anticipated direction	
Increase	52 (34.7%)
Decrease	98 (65.3%)
Average prevalence[a]	
Percentage with outcome, median (IQR)	22 (10–41)
Study size (across both arms)	
Number randomised, median (IQR)	503 (238–1092)
Number randomised (range)	12–46 614

IQR, interquartile range.

[a] Average prevalence in the control arm.

the strength of statistical evidence is <60% for all effect sizes, thus the Bayesian interpretation here is that the findings of this study are uncertain.

5.2.2 | Detecting small effects

In a handful of studies, we were able to identify evidence of a trivial impact (that is, an effect size so small as to almost certainly not be of clinical importance), which we defined as a number needed to treat >200, but which in practice can be smaller or larger depending on the nature of the specific setting, intervention and outcome. The INFANT trial that included nearly 50 000 participants, with the primary outcome occurring in 0.7%, that was not statistically significant (adjusted risk ratio 1.01, 95% CI 0.82–1.25), was a candidate study for being able to demonstrate no impact.[10] For this study the posterior probability of a trivial effect (number needed to treat >200) was 100%. Although trials indeed need to be very large definitely to rule out small effects, this example nicely illustrates how the Bayesian approach can help with a definitive interpretation of a non-statistically significant outcome.

5.2.3 | Unanticipated harmful effects

Although our focus was on posterior probabilities of beneficial effects, it is possible that an intervention which is hypothesized to bring about benefit, can actually have a harmful effect. We do not necessarily suggest that evaluating posterior probabilities of harmful effects should be routine, as posterior probabilities of benefit in such settings would be low. Nonetheless, we did identify that a minority of trials had strong statistical evidence of effects in the unanticipated direction. For example, in one trial delaying infertility treatment

TABLE 2 Examination of consistency of inferences between Bayesian and frequentist approaches.

	Frequentist (*n* = 150)	Bayesian (*n* = 150)
Risk ratio		
Non-convergence,[a] *n* (%)	0 (0%)	6 (4%)
Statistically significant,* *n* (%)	48 (32%)	NA
P-value (or posterior probability of any effect)	0.21 (0.02–0.64)	0.76 (0.38–0.98)
Point estimate, median (IQR)	0.90 (IQR 0.72–1.04)	0.92 (IQR 0.73–1.07)
Risk difference		
Non-convergence,[a] *n* (%)	0 (0%)	0 (0%)
Statistically significant,* *n* (%)	48 (32%)	NA
P-value (or posterior probability of any effect)	0.21 (0.02–0.64)	0.79 (0.40–0.99)
Point estimate (median, IQR)	−1.74 (IQR −7.33 to 0.70)	−1.73 (IQR −7.18 to 0.77)

[a] Subsequent summaries are presented over results that converged.

*$P \leq 0.05$.

TABLE 3 Classification of trials based on strength of statistical evidence of important beneficial effect sizes.

	Statistically non-significant trials (*n* = 102)	Statistically significant trials (*n* = 48)	All (*n* = 150)
Overall strength of statistical evidence[a]			
Strong statistical evidence	8 (7.8%)	47 (100%)	55 (36.7%)
Moderate or weak evidence	94 (92.2%)	1 (2.0%)	95 (63.3%)
At least a small beneficial effect (RD < −0.5 pp; NNT < 200)			
Strong statistical evidence	3 (2.9%)	42 (87.5%)	45 (30.0%)
Moderate statistical evidence	8 (7.8%)	0 (0.0%)	8 (5.3%)
Weak statistical evidence	91 (89.2%)	6 (12.5%)	97 (64.7%)
Large beneficial effect (RD < −4 pp; NNT < 25)			
Strong statistical evidence	0 (0.0%)	24 (50.0%)	24 (16.0%)
Moderate statistical evidence	0 (0.0%)	6 (12.5%)	6 (4.0%)
Weak statistical evidence	102 (100.0%)	18 (37.5%)	120 (80.0%)
Moderate beneficial effect (RD < −1 pp; NNT < 100)			
Strong statistical evidence	0 (0.0%)	40 (83.3%)	40 (26.7%)
Moderate statistical evidence	6 (5.9%)	1 (2.1%)	7 (4.7%)
Weak statistical evidence	96 (94.1%)	7 (14.6%)	103 (68.7%)
Trivial effect (RD > −0.1 pp and RD < 0.1 pp; NNT > 200)			
Strong statistical evidence	2 (2.0%)	0 (0.0%)	2 (1.3%)
Moderate statistical evidence	1 (1.0%)	0 (0.0%)	1 (0.7%)
Weak statistical evidence	99 (97.1%)	48 (100.0%)	147 (98.0%)
Unanticipated harmful effect (RD > 0.1 pp)			
Strong statistical evidence	3 (2.9%)	5 (10.4%)	8 (5.3%)
Moderate statistical evidence	3 (2.9%)	0 (0.0%)	3 (2.0%)
Weak statistical evidence	96 (94.1%)	43 (89.6%)	139 (92.7%)
Any beneficial effect (RD < 0 pp)			
Strong statistical evidence	8 (7.8%)	43 (89.6%)	51 (34.0%)
Moderate statistical evidence	9 (8.8%)	0 (0.0%)	9 (6.0%)
Weak statistical evidence	85 (83.3%)	5 (10.4%)	90 (60.0%)

Note: Strong statistical evidence: posterior probability ≥95%; moderate statistical evidence: posterior probability between 90% and 94%; weak statistical evidence posterior probability <94%.

pp, percentage points.

[a]Overall statistical evidence classified as strong if strong statistical evidence of either at least a small effect or a trivial effect or an unanticipated harmful effect. Italics are non-mutually exclusive categories.

to after a 6-month lifestyle-intervention programme in obese women, statistically significantly reduced, rather than increased, the proportion of women having a vaginal birth within 24 months.[26] In practice we suggest that if trialists did observe a potential harmful effect, it could be useful to examine the probability of large, moderate or small harmful effects.

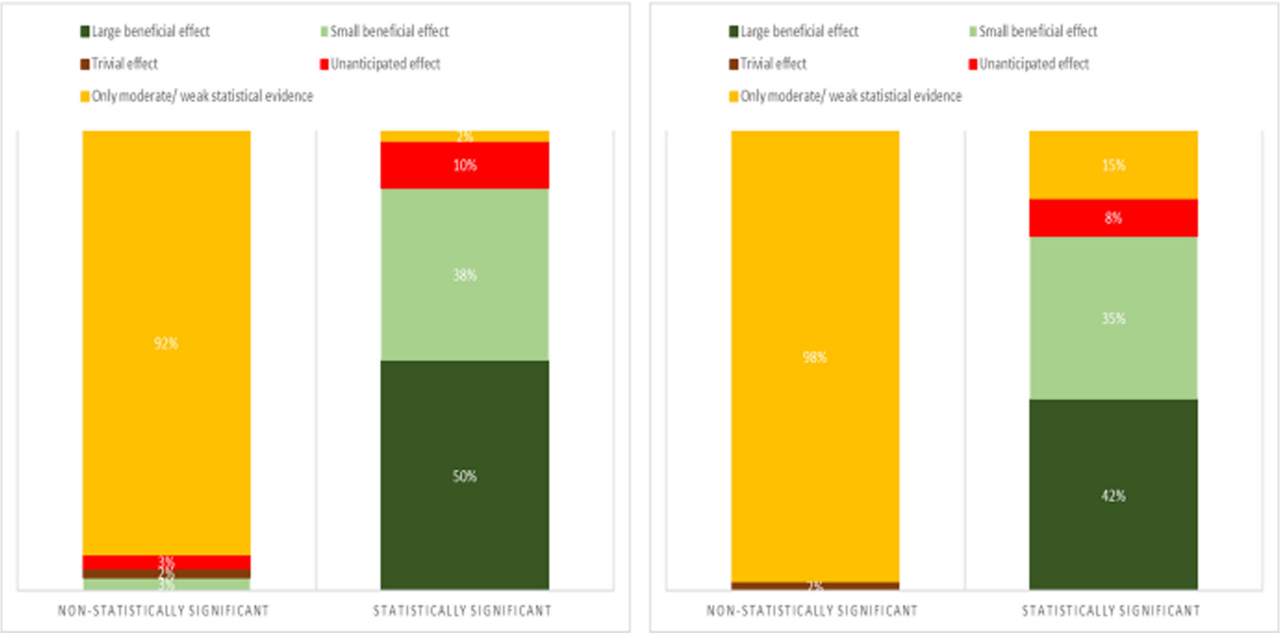
5.3 | Limitations

5.3.1 | Statistical versus scientific evidence

Our classification of the strength of statistical evidence was concerned with the inference based on the primary outcome. In practice, researchers must consider much wider influences – for example, the scientific rigour of the trial, the context, costs and potential harms of the treatment.[4,5] We have not considered these factors but have instead tried to provide researchers with effective tools properly to interpret key outcomes. Only after key outcomes have been interpreted can investigators properly consider the wider implications of whether the intervention should be used. Thus, although we have illustrated this technique on a sample of real trials, we do not attempt to make inferences about specific interventions, and for these reasons we have not undertaken a risk of bias assessment and do not recommend our results be used to inform treatment decisions.

5.3.2 | Retaining reproducibility

We used an arbitrary classification for the strength of the statistical evidence. When we increased the stringency of the statistical evidence by classifying posterior probabilities



(A) classification of strong statistical evidence at 95% **(B)** classification of strong statistical evidence at 97.5%

FIGURE 3 Classification of trials into clinical important effect sizes: at (A) 95% and (B) 97.5% for strong statistical evidence.

>97.5% as strong statistical evidence, the number of studies for which it was possible to conclude something definitive decreased. Lowering thresholds for strength of statistical evidence might lead to increases in non-reproducible results. Relatedly, a similar approach could be undertaken using *P* cut-points. However, the frequentist approach is tightly woven within a paradigm that strongly controls type-1 error (claiming there is an effect when it is a chance finding) and, as a consequence, opens the floodgates for type-2 errors (claiming there is no effect where one exists). Thus, whatever approach adopted, care must be taken to ensure both types of errors are controlled. There are of course other ways to control type-1 errors, such as pre-specification of primary outcomes and anticipated effects, as well as showing reproducibility in other settings. As with any classification system, the pros and cons of misclassification depend on context.[27] For example, very stringent evidence might be required before the acceptance of an invasive surgical procedure, but perhaps less convincing evidence might be acceptable before recommending a low-cost, low-harm, non-invasive therapy.[4,5]

5.3.3 | Classification of size of effects

We have used somewhat arbitrary classifications for clinically important and trivial effect sizes.[28] We thus suggest that with appropriate contextual knowledge, clinically important effect sizes should be defined at the planning stage.[23] Creating an explicit necessity to specify clinically important effect sizes up front, should prompt decision makers to think about this important question at the planning stage rather than the interpretation stage. Although we only consider binary outcomes, the methods proposed can readily be extended to continuous outcomes, where the concepts of clinically important differences are often better established.[29]

5.3.4 | Accessibility and implementation

Frequentist inference is by far the predominant method of inference (Gupta 2012).[13,30,31] Unlike the frequentist approach, a Bayesian analysis requires specification of prior distributions and this might be a perceived barrier to its use.[22] In this application we used standard informative priors illustrating how the approach can be used without dependency on 'priors', which might induce concerns of lack of reproducibility.[32] The finding that the Bayesian and frequentist point estimates, confidence/credible intervals and *P*s/posterior probabilities showed strong concordance gives confidence that inferences are not dependent on the chosen prior.[19,20,21,22,33] Furthermore, the Bayesian approach is pitched here as an aid to interpretation and not as a technique that will radically change the numerical results; thus it might even have a place alongside a conventional frequentist analysis. However, the approach might also be used in conjunction with an informative prior, fully embracing the Bayesian philosophy, and this might be particularly important in rare diseases or interventions in difficult to recruit populations.

5.3.5 | Generalisability

Our review was limited to trials in the area of women's health, but the proposal and its implications should be generalisable to other clinical areas with binary outcomes, albeit perhaps with some reconsideration of what constitutes clinically important effect sizes. In addition, our review was limited to trials in high impact journals, which might suggest that the true proportions of trials with statistically significant findings (~one-third) or with strong statistical evidence (~one-third) in the wider medical literature might be lower than in our review. As others have suggested, when considered from a perspective of clinically important effects, there is no real difference in superiority, non-inferiority and equivalence trials.[25] We thus suggest this approach could be used for the interpretation of non-inferiority as well as superiority trials.[34]

6 | CONCLUSION

The key findings of most randomised trials are interpreted on the basis of statistical significance – leading to many interventions being declared as ineffective when the findings are statistically uncertain (type-2 error). This is a well-known problem. In part, this problem of misinterpretation arises because a strict frequentist interpretation of statistical significance prioritises not misclassifying treatments as effective when they are not (type-1 error). In so doing, this perpetuates the problem of treatments being declared as ineffective when they are actually uncertain. A Bayesian interpretation of findings, alongside reporting of confidence intervals and effect sizes, may help strike a balance between minimising both types of errors.

AUTHOR CONTRIBUTIONS

KH, MT and AC led the development of the idea. KH analysed the data and wrote the first draft of the article. PM and RL undertook the search and data abstraction. PM led the development of the associated protocol for the review. All authors made an intellectual contribution to the development of the ideas and commented on draft versions of the paper.

ACKNOWLEDGEMENTS

None.

FUNDING INFORMATION

This research was partly funded by the UK NIHR Collaborations for Leadership in Applied Health Research and Care West Midlands initiative. Karla Hemming is funded by a NIHR Senior Research Fellowship SRF-2017-10-002. This research is independent to the funder.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY STATEMENT

Data are available on request.

ETHICS APPROVAL

No ethical approval was obtained for this study, which is a review and therefore no ethical review is needed.

REFERENCES

1. Vinkers CH, Lamberink HJ, Tijdink JK, Heus P, Bouter L, Glasziou P, et al. The methodological quality of 176,620 randomized controlled trials published between 1966 and 2018 reveals a positive trend but also an urgent need for improvement. PLoS Biol. 2021;19(4):e3001162. https://doi.org/10.1371/journal.pbio.3001162
2. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010;7(9):e1000326. https://doi.org/10.1371/journal.pmed.1000326
3. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Obstet Gynecol. 2009;114(6):1341–5. https://doi.org/10.1097/AOG.0b013e3181c3020d
4. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. Am Stat. 2019;73(sup1):235–45.
5. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Med Res Methodol. 2020;20:244.
6. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485.
7. Schulz KF, Altman DG, Moher D, The CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Ann Intern Med. 2010;152(11):726–32.
8. Gewandter JS, McDermott MP, Kitt RA, Chaudari J, Koch JG, Evans SR, et al. Interpretation of CIs in clinical trials with non-significant results: systematic review and recommendations. BMJ Open. 2017;7(7):e017288. https://doi.org/10.1136/bmjopen-2017-017288
9. Hemming K, Taljaard M. Why proper understanding of confidence intervals and statistical significance is important. Med J Aust. 2021;214(3):116–8.e1.
10. INFANT Collaborative Group. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. Lancet. 2017;389(10080):1719–29.
11. Rouzi AA, Alsahly N, Alamoudi R, Almansouri N, Alsinani N, Alkafy S, et al. Randomized clinical trial between hourly titrated and 2 hourly static oral misoprostol solution for induction of labor. Am J Obstet Gynecol. 2017;216(4):405.e1–6. https://doi.org/10.1016/j.ajog.2016.11.1054
12. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305–7. https://doi.org/10.1038/d41586-019-00857-9
13. Gates S, Ealing E. Reporting and interpretation of results from clinical trials that did not claim a treatment difference: survey of four general medical journals. BMJ Open. 2019;9(9):e024785.
14. Hemming K, Taljaard M. A review of high impact journals found that misinterpretation of non-statistically significant results from randomised trials was common. J Clin Epidemiol. 2022;145:112–20.
15. StataCorp. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC; 2021.
16. Bibbins-Domingo K. Aspirin use for the primary prevention of cardiovascular disease and colorectal cancer: U.S. Preventative Service Task Force recommendation statement. Ann Intern Med. 2016;164:836–45.
17. Antithrombotic Trialists Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. BMJ. 2002;324(7329):71–86.
18. RECOVERY Collaborative Group, Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, et al. Dexamethasone in hospitalized patients with Covid-19. N Engl J Med. 2021;384(8):693–704. https://doi.org/10.1056/NEJMoa2021436
19. Hamilton FW, Lee T, Arnold DT, Lilford R, Hemming K. Is convalescent plasma futile in COVID-19? A Bayesian re-analysis of the RECOVERY randomized controlled trial. Int J Infect Dis. 2021;109:114–7.
20. Hoek JM, Field SM, de Vries YA, Linde M, Pittelkow M-M, Muradchanian J, et al. Rethinking remdesivir for COVID-19: a Bayesian reanalysis of trial findings. PLoS One. 2021;16(7):e0255093.
21. Zampieri FG, Damiani LP, Bakker J, Ospina-Tascón GA, Castro R, Cavalcanti AB, et al. Effects of a resuscitation strategy targeting

peripheral perfusion status versus serum lactate levels among patients with septic Shock. A Bayesian reanalysis of the ANDROMEDA-SHOCK trial. Am J Respir Crit Care Med. 2020;201(4):423–9.

22. Zampieri FG, Casey JD, Shankar-Hari M, Harrell FE Jr, Harhay MO. Using Bayesian methods to augment the interpretation of critical care trials. An overview of theory and example reanalysis of the alveolar recruitment for acute respiratory distress syndrome trial. Am J Respir Crit Care Med. 2021;203(5):543–52.
23. Harrell F. Language for communicating frequentist results about treatment effects. 2021 [accessed 2021 July 15]. https://discourse.datamethods.org/t/language-for-communicating-frequentist-results-about-treatment-effects/934
24. Yarnell CJ, Abrams D, Baldwin MR, Brodie D, Fan E, Ferguson ND, et al. Clinical trials in critical care: can a Bayesian approach enhance clinical and scientific decision making? Lancet Respir Med. 2021;9(2):207–16.
25. Dunn DT, Copas AJ, Brocklehurst P. Superiority and non-inferiority: two sides of the same coin? Trials. 2018;19:499. https://doi.org/10.1186/s13063-018-2885-z
26. Mutsaerts MA, van Oers AM, Groen H, Burggraaff JM, Kuchenbecker WK, Perquin DA, et al. Randomized trial of a lifestyle program in obese infertile women. N Engl J Med. 2016;374(20):1942–53.
27. Harrell. 2021. [cited 2023 June 12]. https://hbiostat.org/proj/covid19/
28. Watson SI, Chen YF, Nguyen-Van-Tam JS, Myles PR, Venkatesan S, Zambon M, et al. Evidence synthesis and decision modelling to support complex decisions: stockpiling neuraminidase inhibitors for pandemic influenza usage. F1000Res. 2016;5:2293.
29. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA. 2014;312(13):1342–3.
30. ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH Harmonized Tripartite Guideline. Stat Med. 1999;18:1905–42.
31. Gupta SK. Use of Bayesian statistics in drug development: advantages and challenges. Int J Appl Basic Med Res. 2012;2(1):3–6. https://doi.org/10.4103/2229-516X.96789
32. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. Boca Raton, FL: CRC Press; 2013.
33. Goligher EC, Tomlinson G, Hajage D, Wijeysundera DN, Fan E, Jüni P, et al. Extracorporeal membrane oxygenation for severe acute respiratory distress syndrome and posterior probability of mortality benefit in a post hoc Bayesian analysis of a randomized clinical trial. JAMA. 2018;320:2251–9.
34. van Ravenzwaaij D, Monden R, Tendeiro JN, Ioannidis JPA. Bayes factors for superiority, non-inferiority, and equivalence designs. BMC Med Res Methodol. 2019;19(1):71.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hemming K, Melo P, Luo R, Taljaard M, Coomarasamy A. A re-analysis of 150 women's health trials to investigate how the Bayesian approach may offer a solution to the misinterpretation of statistical findings. BJOG. 2023;00:1–10. https://doi.org/10.1111/1471-0528.17570