BMJ Open

'Not clinically effective but cost-effective' - paradoxical conclusions in randomised controlled trials with 'doubly null' results: a cross-sectional study

James Raftery ,[1] HC Williams,[2] Aileen Clarke,[3] Jim Thornton,[2] John Norrie,[4] Helen Snooks,[5] Ken Stein[6]

To cite: Raftery J, Williams HC, Clarke A, *et al*. 'Not clinically effective but cost-effective' - paradoxical conclusions in randomised controlled trials with 'doubly null' results: a cross-sectional study. *BMJ Open* 2020;**10**:e029596. doi:10.1136/bmjopen-2019-029596

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-029596).

Received 01 February 2019
Revised 11 November 2019
Accepted 19 November 2019

For numbered affiliations see end of article.

Correspondence to
Professor James Raftery;
j.p.raftery@soton.ac.uk

ABSTRACT

Objectives Randomised controlled trials in healthcare increasingly include economic evaluations. Some show small differences which are not statistically significant. Yet these sometimes come to paradoxical conclusions such as: 'the intervention is not clinically effective' but 'is probably cost-effective'. This study aims to quantify the extent of non-significant results and the types of conclusions drawn from them.

Design Cross-sectional retrospective analysis of randomised trials published by the UK's National Institute for Health Research (NIHR) Health Technology Assessment programme. We defined as 'doubly null' those trials that found non-statistically significant differences in both primary outcome and cost per patient. Paradoxical was defined as concluding in favour of an intervention, usually compared with placebo or usual care. No human participants were involved. Our sample was 226 randomised trial projects published by the Health Technology Assessment programme 2004 to 2017. All are available free online.

Results The 226 projects contained 193 trials with a full economic evaluation. Of these 76 (39%) had at least one 'doubly null' comparison. These 76 trials contained 94 comparisons. In these 30 (32%) drew economic conclusions in favour of an intervention. Overall report conclusions split roughly equally between those favouring the intervention (14), and those favouring either the control (7) or uncertainty (9).

Discussion Trials with 'doubly null' results and paradoxical conclusions are not uncommon. The differences observed in cost and quality-adjustedlife year were small and non-statistically significant. Almost all these trials were also published in leading peer-reviewed journals. Although some guidelines for reporting economic results require cost-effectiveness estimates regardless of statistical significance, the interpretability of paradoxical results has nowhere been addressed.

Conclusions Reconsideration is required of the interpretation of cost-effectiveness analyses in randomised controlled trials with 'doubly null' results, particularly when economics favours a novel intervention.

Strengths and limitations of this study

- A strength of this study is the identification of a problem to do with results on cost-effectiveness from randomised trials which is fairly common but has not otherwise been reported or examined.
- A limitation was that the sample was confined to that of trials funded and published by the UK's Health Technology Assessment (HTA) programme.
- The study was strengthened by using full reports of each study in the HTA monograph series, which requires conclusions and recommendations for research.
- The generalisability of this study was enhanced by the fact that most of the trials reviewed were published in major peer-reviewed medical journals.

BACKGROUND

Randomised trials are widely seen as providing the most robust evidence of the effectiveness of healthcare interventions. For that reason, they are mandatory for drug licensing. Many trials show null results, that is a non-statistically significant difference in the primary outcome. Several studies have identified 'spin' whereby authors nonetheless draw possibly unwarranted conclusions.[1–3]

Given the importance of costs, trials increasingly include estimation of cost-effectiveness. Even when both outcome and cost show only small, non-significant differences, some trials make strong claims. Apparently paradoxical conclusions are sometimes drawn along the lines that 'the intervention was not clinically effective' but 'is probably cost-effective'.

The definition of 'spin' relies largely on statistical significance, a concept which some have seen as outdated.[4] Economic evaluation in randomised trials has abandoned statistical significance in favour of decision analysis.[5]

This new approach has become the norm, as reflected in guidelines for reporting such as CHEERS (Consolidated Health Economics Evaluation Reporting statement).[6] Most trials however continue to be designed to test hypotheses using statistical significance[7] not least due to drug licensing requirements.[8]

Claims that a new intervention is cost-effective usually imply that it should be adopted. Claims to the contrary, in favour of placebo or usual care, by contrast imply no change. Our interest has mainly to do with the former due to the implications for healthcare.

We defined as 'doubly null' those trials that found no statistically significant difference in both primary outcome and cost per patient. Paradoxical was defined as going on to conclude in favour of an intervention, whether compared with placebo, usual care or to another intervention.

We aimed to establish how often 'doubly null' results occur and how commonly paradoxical conclusions were drawn. Our sample was randomised trials funded and published by the UK National Health Service (NHS). The NHS funds research through National Institute for Health Research (NIHR). Its largest programme, the Health Technology Assessment (HTA) programme mainly funds systematic reviews and randomised trials. Those trials usually include economic evaluation. The programme publishes the entire findings of each trial in its monograph series. These combine reports of clinical efficacy, cost and cost-effectiveness in a format that requires an overall conclusion.[9]

METHODS

The criteria for trials with 'doubly null' results were non-statistically significant differences in both (a) the primary outcome, and (b) mean cost per patient. Significance was defined at the 0.05 level (or 95% CIs).

Our sampling frame was all 226 projects which included a randomised trial published by the NIHR HTA programme 2004 to 2017. We then excluded three types

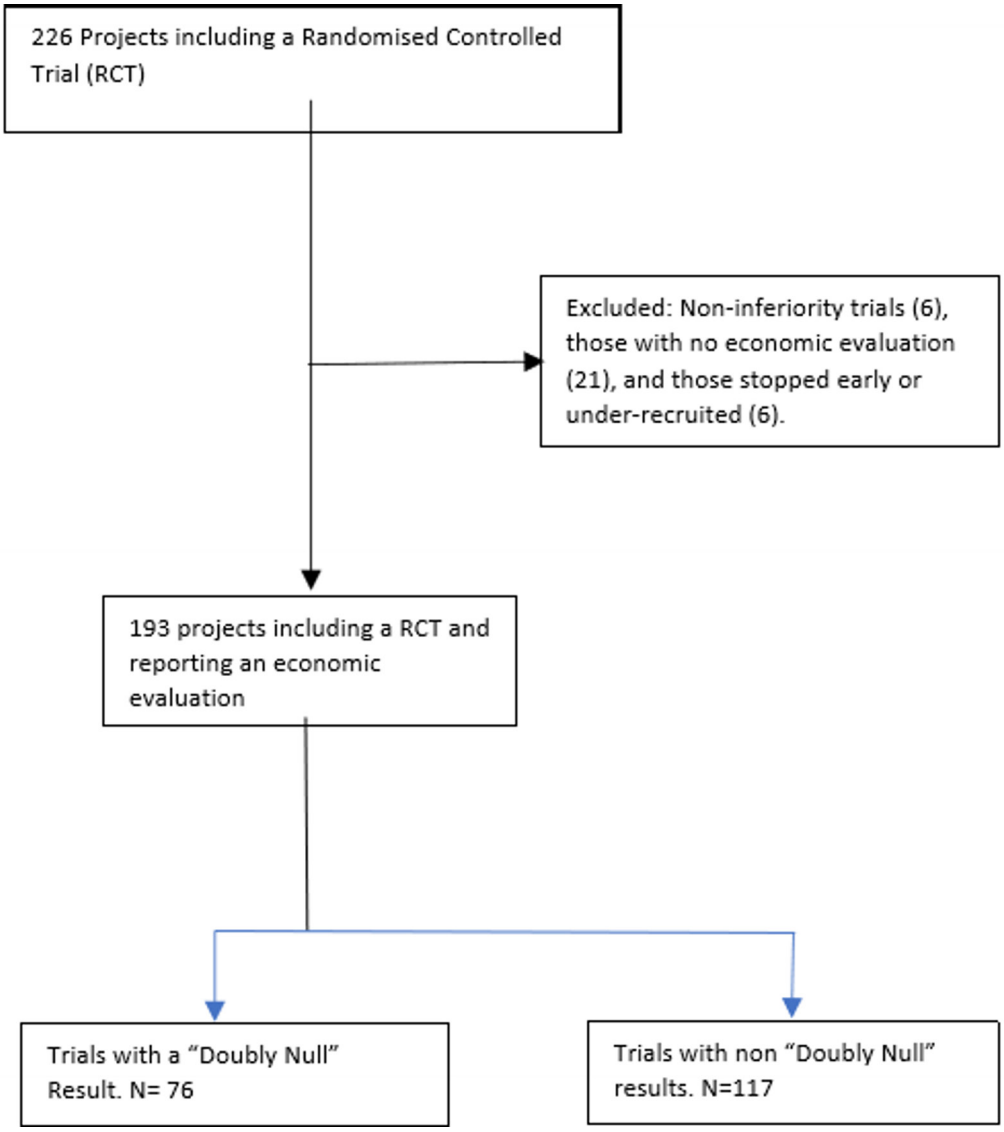

Figure 1 Flowchart: projects including randomised trials identifying those which included a 'doubly null' result.

of trials: those testing non-inferiority, those which did not report an economic evaluation and those which were stopped prematurely.

This study did not have an original protocol. It evolved from an initial cross-sectional review of 'doubly null' results to consideration of the implications of paradoxical conclusions. This led to preparation of focussed summaries of each trial with such conclusions. All the reports reviewed are available free at https://www.journalslibrary.nihr.ac.uk/hta/#/. The volume and issue numbers in online supplementary appendix 1 identify all those included and those with 'doubly null' results.

Monograph titles were used to identify those containing a randomised trial. Each trial was checked to identify those with 'doubly null' results. For each of these, the abstract was extracted along with additional text to do with cost-effectiveness. Only primary comparisons were included, as defined by each project. Overall conclusions were classified based on the conclusion section of the Abstract. Quotes from those whose economic conclusion favoured an intervention are provided in online supplementary appendix 2.

Eligibility and data extraction were carried out by JR with independent checks by two researchers (SB, ABJ) Disagreements were resolved by discussion with reference to the extracts, with additional text extracted if necessary. Extracts were stored in Word and analysed in Excel.

Patient and public involvement

There was no patient or public involvement in this project.

RESULTS

Our exclusion criteria reduced the number of trial projects from 226 to 193. Twenty-one projects did not include economic evaluation, six were closed early and six were non-inferiority trials (figure 1).

Of the 193 trials which included economic evaluation, 76 (39%) had 'doubly null' results for at least one comparison. These are identified in online supplementary appendix 1. This proportion varied widely by year (table 1).

The economic analyses of these results (table 2) found an intervention to have an acceptable incremental cost-effectiveness ratio in 30 instances (30/94, 32%).

Project reports with economics analyses supporting an intervention came to overall conclusions as stated in their abstracts that split roughly equally between favouring the intervention (14 comparisons), and those favouring either the control (7 comparisons) or uncertainty (9 comparisons) (figure 2).

Differences in all effect sizes were small. Differences in the primary outcome were all less than any prespecified minimally important difference. None of the quality-adjustedlife year (QALY) differences were statistically significant. Both cost and QALY differences were very small.

The probability of an intervention being cost-effective was only loosely associated with the conclusion drawn. Those favouring the intervention had on average 89% probability of being cost-effective, compared with 79% for those favouring a control (see online supplementary appendix 2). These averages however spanned wide ranges and are based on relatively few studies. Some projects found small differences convincing while others saw them as indicating lack of evidence. Other factors such as secondary outcomes, borderline significance and patient convenience were sometimes mentioned as having influenced the overall conclusions.

DISCUSSION

The potential for drawing conclusions from 'doubly null' results arises due to different methodological perspectives that can be described as hypothesis testing versus decision analysis. Rejection of a hypothesis of superiority precludes any conclusion other than the null hypothesis. Decision analysts by contrast estimate an acceptable incremental cost per QALY regardless of the size or the statistical significance of differences in cost and outcome.[5] CIs are replaced by the probability that an intervention is cost-effective relative to some threshold such as that associated with National Institute for Health and Care Excellence (NICE).

Limitations on our analysis include reliance on a possibly atypical sample. However, decision analysis as

Table 1 Projects containing randomised trials published by National Institute for Health Research Health Technology Assessment programme 2004 to 2017, by those with 'doubly null' results and conclusions drawn

Volume	Year	RCTs (n)	Doubly null (n)	%	Comparisons
8	2004	8	1	12.5	1
9	2005	13	6	46.2	8
10	2006	9	3	33.3	4
11	2007	8	1	12.5	1
12	2008	3	1	33.3	1
13	2009	16	7	43.8	10
14	2010	11	5	45.5	6
15	2011	8	0	0.0	0
16	2012	8	7	87.5	8
17	2013	11	7	63.6	10
18	2014	17	10	58.8	15
19	2015	28	13	46.4	14
20	2016	27	11	40.7	11
21	2017	26	4	15.4	5
		193	76	39.4	94

The 76 trials contained 94 comparisons (table 1). This was due to some trials having several arms.
RCTs, Randomised controlled trials.

Table 2 "Doubly null" comparisons: those with economics pro an intervention by overall conclusion

		Comparisons		Overall conclusions pro:		
Volume	Year	Comparisons	Economics pro an intervention	Control	Intervention	Mixed/uncertain
8	2004	1	1		1	
9	2005	8	5	1	2	2
10	2006	4	1			1
11	2007	1	0			
12	2008	1	0			
13	2009	10	1			1
14	2010	6	2		2	
15	2011	0	0			
16	2012	8	1			1
17	2013	10	2	2		
18	2014	15	6		5	1
19	2015	14	3		1	2
20	2016	11	6	4	2	
21	2017	5	2		1	1
		94	30	7	14	9
Economics pro intervention (%)			32%			
Overall conclusion (%)				23% (7/30)	47% (14/30)	30% (9/30)

opposed to hypothesis testing has become the recommended approach in guidelines for economic evaluation. The widely used CHEERS guideline[6] for reporting economic results requires that uncertainty be characterised and that:

> Because failure to reject the hypothesis about the equality of two therapies is not the same as finding that outcomes of two therapies are identical, cost-effectiveness analysis should still be performed if the clinical study fails to demonstrate a statistically significant difference in clinical end points.[6]

Journals that have adopted CHEERS may not have understood that they were thereby requiring decision analysis which would in turn lead to paradoxical conclusions of the sort described here.

Another limitation was that we included some comparisons of one intervention against another as opposed to a placebo or usual care. This increases the number of comparisons that favour an intervention. Against this, most trials compared interventions with placebo or usual care.

Many of the inter intervention comparisons included were all to do with surgery, where interventions were commonly used procedures as chosen by the participating surgeon. In such comparisons, a conclusion in favour of either was deemed paradoxical.

Most of the trials reviewed were also published elsewhere. Of the 24 trials with 'doubly null' results published in HTA monographs in 2015 and 2016 all but one were also published in reputable medical peer-reviewed journals. However, since clinical and economical results could be reported in different articles, the extent of paradoxical conclusions may not have been apparent.

Our results contribute to the debate over the value of a hypothesis testing not least by showing that non-significant differences are almost always very small. Although authorities such as the American Statistical Association[4] have urged abandonment of statistical significance, some prominent medical statisticians have argued that by dropping prespecified significance, interpretation could become completely arbitrary.[8]

The monograph reports with 'doubly null' results that came to conclusions favouring an intervention rarely discussed the plausibility of those claims. Research recommendations, which are required in HTA monographs, never proposed researching the plausibility of the seemingly paradoxical conclusions.

While trials with non-statistically significant results can contribute to knowledge through inclusion in meta-analysis, no similar aggregation is possible with estimates of cost-effectiveness. To be useful, trial reports should include full data on costs and effects along with joint uncertainty.

The practice of economic evaluation alongside every randomised trial deserves reconsideration. Rather than being based on a single trial, economic evaluation should generally be based on the totality of evidence established by systematic reviews and meta-analysis. NICE and similar decision-making bodies specify such methods. Only exceptionally (first or biggest trial) can a single trial provide grounds for implementation.

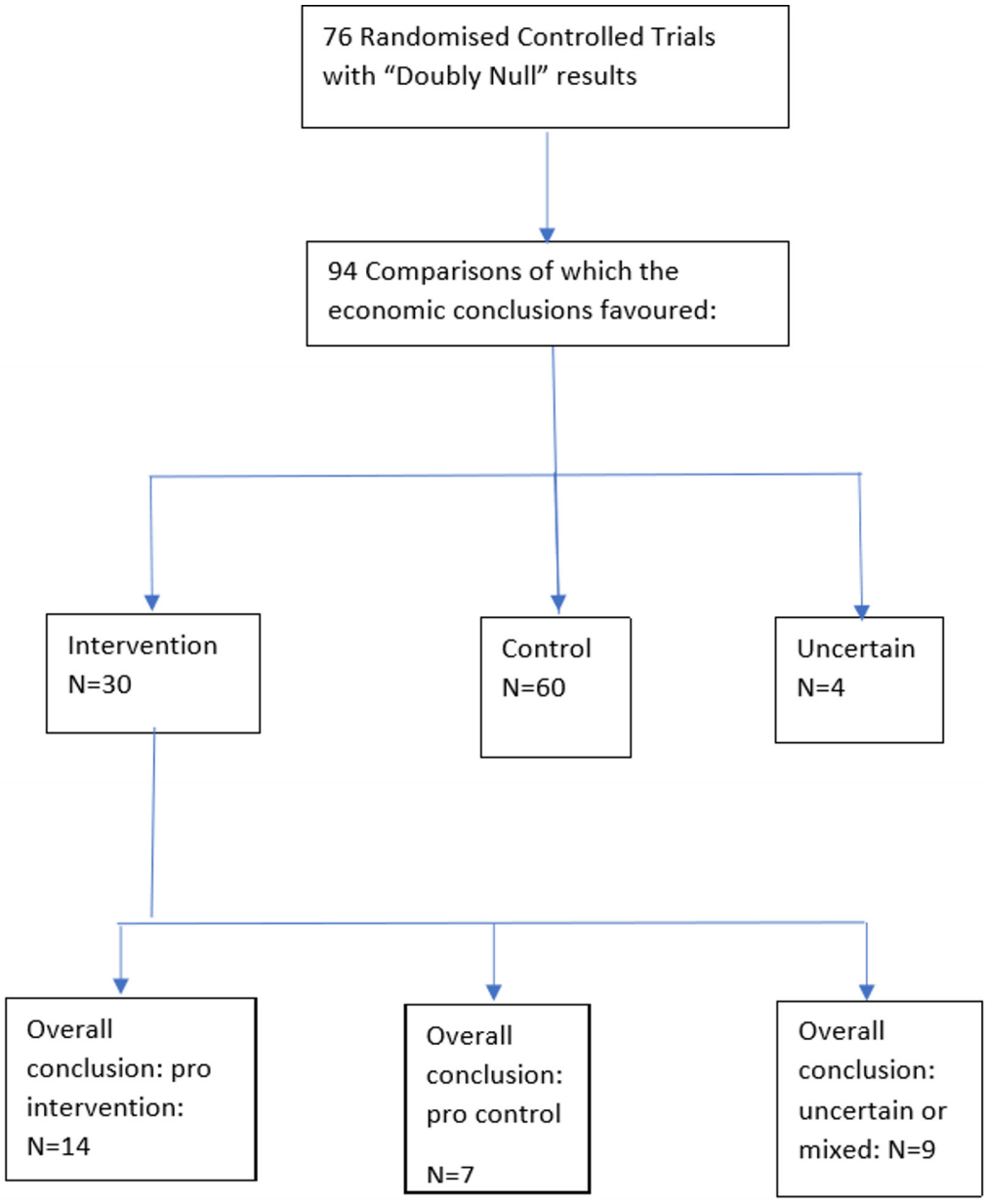

Figure 2 Types of overall conclusions drawn from comparisons which were "doubly null".

CONCLUSIONS

'Doubly null' results and paradoxical conclusions are increasingly drawn from randomised trials of healthcare interventions. In our cohort, almost 40% of trials had 'doubly null' results . From those whose economic evaluation favoured the intervention, projects' overall conclusions split roughly equally between those favouring the intervention, and those favouring the control or uncertainty.

The possibility of paradoxical conclusions arises from different paradigms which are unlikely to change. Trials seem likely to continue to be powered on target differences.[7] Decision analysis is required in widely used guidelines for reporting economic evaluations.

Editors of journals reporting randomised trials should be more alert to the issue of paradoxical conclusions. Besides noting them, the plausibility of the conclusions need to be questioned. Further research may sometimes be required.

Author affiliations
[1]Faculty of Medicine, Southampton University, Southampton, UK
[2]University of Nottingham, Nottingham, UK
[3]Division of Health Sciences, University of Warwick, Coventry, UK
[4]Edinburgh Clinical Trials Unit, University of Edinburgh No. 9, Bioquarter, Edinburgh, UK
[5]Medicine, Swansea University, Swansea, UK
[6]PenTAG, University of Exeter Medical School, Exeter, UK

Acknowledgements Independent checks of eligibility and data extraction were carried out by Sheetal Bhurke (SB) and Amanda Blatch-Jones (ABJ) both researchers employed by NETSCC, University of Southampton.

Contributors JR initiated the project, carried out the research and wrote successive drafts. HW helped to initiate the project and made helpful comments on draft manuscripts. AC made helpful comments on draft manuscripts. JT made helpful comments on draft manuscripts. JN made helpful comments on draft

manuscripts. HS made helpful comments on draft manuscripts. KS helped to initiate the project and commented on draft manuscripts.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests All authors work part-time for the NIHR HTA programme as well as being independent academics. All but HW and AC are current members of the editorial board of the HTA monograph series. HW is director of the HTA programme. AC is a past member of the editorial board of the HTA monograph series. JR is employed part-time as a researcher by NETSCC, the secretariat which manages the NIHR HTA programme. Neither NIHR nor the HTA programme had any role in interpreting the results nor in preparing this paper for publication. NETSCC has taken responsibility for payment of the publication fee.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

ORCID iD

James Raftery http://orcid.org/0000-0003-1094-8578

REFERENCES

1 Khan MS, Lateef N, Siddiqi TJ, *et al*. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: a systematic review. *JAMA Netw Open* 2019;2:e192622.
2 Tello M, Zaiem F, Tolcher MC, *et al*. Do not throw the baby out with the bath water: a guide for using non-significant results in practice. *Evid Based Med* 2016;21:161–2.
3 Boutron I, Dutton S, Ravaud P, *et al*. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303:2058–64.
4 Wasserstein RL, Lazar NA. The ASA Statement on *p* -Values: Context, Process, and Purpose. *Am Stat* 2016;70:129–33.
5 Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ* 1999;18:341–64.
6 Husereau D, Drummond M, Petrou S, *et al*. Health Economic Evaluation Publication Guidelines - CHEERS Good Reporting Practices Task Force CHEERS guideline Consolidated Health Economic Evaluation Reporting Standards. (CHEERS). *BMJ* 2013;346.
7 Cook JA, Julious SA, Sones W, *et al*. DELTA^2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ* 2018;363:k3750.
8 Cook JA, Fergusson DA, Ford I, *et al*. There is still a place for significance testing in clinical trials. *Clin Trials* 2019;16:223–4.
9 Health Technology Assessment. Available: https://www.nihr.ac.uk/funding-and-support/funding-for-research-studies/funding-programmes/health-technology-assessment/ [Accessed 15 Jun 2019].