Meta-analysis of cardiovascular superiority trials published in the *New England Journal of Medicine* to elucidate the concept of superiority margin

Nanda Gamad, Nusrat Shafiq, Samir Malhotra

Department of Pharmacology,
Postgraduate Institute of
Medical Education and
Research, Chandigarh, India

Correspondence to

Professor Samir Malhotra,
Postgraduate Institute of
Medical Education and
Research, Chandigarh 160012,
India; smal.pgi@gmail.com

Received 9 March 2019

Revised 21 August 2019

Accepted 21 February 2020

Published Online First 10 March
2020

ABSTRACT

Objective To show that overpowered trials claim statistical significance detouring clinical relevance and warrant the need of superiority margin to avoid such misinterpretation.

Design Selective review of articles published in the *New England Journal of Medicine* between 1 January 2015 and 31 December 2018 and meta-analysis following Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist.

Eligibility criteria for selecting studies and methods Published superiority trials evaluating cardiovascular diseases and diabetes mellitus with positive efficacy outcome were eligible. Fixed effects meta-analysis was performed using RevMan V.5.3 to calculate overall effect estimate, pooled HR and it was compared with mean clinically significant difference.

Results Thirteen eligible trials with 164 721 participants provided the quantitative data for this review. Largely, the primary efficacy endpoint in these trials was the composite of cardiovascular death, non-fatal myocardial infarction, unstable angina requiring rehospitalisation, coronary revascularisation and fatal or non-fatal stroke. The pooled HR was 0.86 (95% CI 0.84 to 0.89, $I^2=45\%$) which was lower than the mean clinically significant difference of 0.196 (19.6%, range: 0.09375–0.35) of these studies. There was a wide 95% CI in these studies from 0.56 to 0.99. The upper margin of CI in most of the studies was close to the line of no difference. Absolute risk reduction was small (1.19% to 2.3%) translating to a high median number needed to treat of 63 (range: 43 to 84) over a follow-up duration of 2.95 years.

Conclusions The results of this meta-analysis indicate that overpowered trials give statistically significant results undermining clinical relevance. To avoid such misuse of current statistical tools, there is a need to derive superiority margin. We hope to generate debate on considering clinically significant difference, used to calculate sample size, as superiority margin.

INTRODUCTION

Superiority trials test if one drug is better than the other. This information is mostly used by busy clinicians who want quick information and are in search of new treatment for the unmet needs of their patients.¹ The design of such trials facilitates determination of the smallest difference in the primary endpoint between the study arms. Traditional null hypothesis statistical testing is used to decide the statistical significance of two-sided tests using p value and CI. Progress towards such informative trial

design and analysis has been slow and this method of testing carries a high chance of misinterpretation and misuse.^{1–3} The smallest difference to be detected is usually made a priori and used to calculate sample size. Currently, there is no consensus or scientific formulae to choose the smallest difference that is clinically meaningful and important for public health decisions.⁴ Over the years smallest difference is taken for granted for a realistic difference that can be identified with the available sample size,⁵ which raises two issues: smaller than an ideal sample size increases type II error while larger than required sample size increases the power of the study ('overpowered trials') and decreases the magnitude of the smallest difference required to produce statistical significance.

Criticisms have been raised by us^{3 6 7} and other authors^{2 4 8} regarding the conduct and interpretation of superiority trials. We have raised our concerns on the results of trials such as EMPA-REG³ and CANTOS⁷ the HRs of which are 0.86 and 0.85, respectively. Its 95% CIs are 0.74 to 0.99 and 0.74 to 0.98 and the clinically meaningful differences used to calculate its sample size are 0.77 and 0.80 (20% risk difference translates to HR of 0.80), respectively. Since the upper limits of 95% CI of these studies is close to line of no difference, there is a small improvement over the control arms. Had the authors considered clinically meaningful difference while analysing their study results, both the studies would give statistically non-significant results. Such interpretation of results has the potential of misuse of data as the drugs may not be sufficiently beneficial to obtain regulatory approval. Few methodologies have been put forth to evaluate the statistical significance such as lowering p value to 0.005,⁹ Cohen's rule of thumb¹⁰ and number needed to treat (NNT),¹¹ but none of them work as well as non-inferiority (NI) margin for NI trials in justifying both statistical and clinical significance.

In this meta-analysis, we consolidate our suggestion about superiority margin for superiority trials akin to NI margin for NI trials, with some evidence. We have analysed superiority trials with significant outcomes and have assessed the clinical relevance of their results. The analysis made in this study may impact on the considerations to be taken while conducting and also while interpreting the results of superiority trials.

© Author(s) (or their employer(s)) 2021. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Gamad N, Shafiq N, Malhotra S. *Postgrad Med J* 2021;**97**:227–233.

METHODS

This is a selective review and meta-analysis of all the superiority trials published in the *New England Journal of Medicine* (NEJM) between 1 January 2015 and 31 December, 2018.

Search strategy

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses consensus statement for the design, implementation and reporting of this study.¹² All trials published in NEJM between 2015 and 2018 were hand-searched for the type of study design. Articles were selected based on our inclusion criteria.

Study selection, inclusion and exclusion criteria

We included all blinded or open-label superiority trials with pharmacological intervention conducted in adults that had positive primary outcome with a continuous or categorical variable. We set a 4 year search limit of studies published in NEJM because we thought this was sufficient to show the trend of conducting and reporting of superiority trials practiced worldwide by eminent researchers. We chose to include studies conducted in populations with cardiovascular diseases and diabetes mellitus because these disease conditions encompass large sample size trials, owing to high prevalence and chronicity, and also because several drugs in this section receive Food and Drug Administration (FDA) approval every year claiming superiority over the existing drugs. Studies conducted in paediatric population, or in populations with venous thromboembolism, bleeding, cardiomyopathy and shock, were excluded. We also excluded trials involving any non-pharmacological intervention such as (but not limited to) lifestyle modification, percutaneous coronary intervention, coronary artery bypass grafting and pacemaker implantation. Studies which expressed their primary endpoint in mean difference,^{13–23} OR²⁴ and risk ratio^{25–26} and also studies which had non-significant primary efficacy outcome were excluded. This eligibility criteria largely yielded a uniform primary outcome, that is, the composite of cardiovascular death, non-fatal myocardial infarction (MI), unstable angina requiring rehospitalisation, coronary revascularisation, fatal or non-fatal stroke,^{24–39} except for one study⁴⁰ which had safety endpoint as its primary outcome. So although it met our eligibility criteria, we excluded this study to maintain uniformity and generalisability of our results.

Information sources

Individual journal volumes and issues were hand-searched for superiority trials. All the selected articles and their supplementary materials were searched for clinically meaningful difference or effect size used to calculate the sample size.

Data extraction

NG did the search and independently screened the titles and abstracts yielded by the search against the inclusion criteria. Full reports for all titles and abstracts that appeared to meet the inclusion criteria or where there was any uncertainty were obtained. NG and SM then screened the full-text reports and decided whether these met the inclusion criteria. We resolved disagreements through discussion with our third author NS. Reasons for excluding trials were recorded. The review authors extracted the data on study characteristics such as (1) trial design, population and sample size; (2) intervention details; (3) duration of treatment and median follow-up; (4) outcome measures; and (5) study results for the primary efficacy outcome.

Data collection process

Data abstracted included endpoints for decision-making, that is, HR in the primary efficacy endpoint along with their CI, p value and NI margin or clinically significant difference, if present.

Risk of bias within and across studies

Two authors (NG and SM) independently assessed the risk of bias of each included trial. If risk of bias was not clear from the main manuscript, we obtained records from the articles' supplementary materials. Cochrane's risk of bias tool was used to assess risk of bias among included studies. We assessed overall risk based on the description of randomisation methods, allocation concealment, blinding of participants, investigators and statistician and attrition and reporting biases. As per tool input, biases were recorded as low risk, high risk or unclear (either lack of information or uncertainty over the potential for bias). Any disagreements in the risk of bias were discussed and resolved by consensus of all the authors.

Data analysis

Results for all primary efficacy outcomes are expressed as HR with 95% CIs, calculated from end of treatment values. For data expressed in person years, we calculated incident rate for the total duration of the study. Overall effect estimate of the meta-analysis was calculated using RevMan V.5.3 with fixed effects model of heterogeneity and Observed-Expected (O-E) and variance statistic as per Cochrane Handbook for Systematic Reviews of Interventions⁴¹ and Tierney *et al.*⁴² Two studies^{34–35} had two coprimary efficacy outcomes. In the fixed effects model, HR of only the first coprimary outcome is used, as HR and its CI for second coprimary outcome were almost the same. For the same reason, we used only the single-dose analysis values from the study by Bonaca *et al.*,³⁹ for the final assessment. Sensitivity analysis including these values did not differ significantly from the reported result. Further, descriptive statistics were used to describe sample size, duration of the studies and overall clinically significant difference. Absolute risk reduction (ARR) and NNT were also calculated from the end of treatment values.

RESULTS

Search results

During the search limit period, a total of 840 original articles were published in NEJM. The selection process of the articles is outlined in figure 1. Studies reporting results of subgroup analysis or combined analysis were excluded as duplicates. Filtering the articles based on disease-specific criteria yielded 171 potentially relevant articles. Further, 113 studies were excluded as they were either not a superiority trial or involved any non-pharmacological intervention. Full text of the remaining studies were retrieved and assessed. In total, 16 superiority trials were eligible for the final analysis. Out of these, three trials^{26–28} had to be excluded as they reported results as OR or risk ratio (RR). Since OR and RR evade time-dependent factor, in contrast to HR, it was considered inappropriate to club trials reporting OR/RR with HR. However, sensitivity analysis including these three trials was conducted, it yielded similar pooled estimate but with increased heterogeneity (result not shown).

Study characteristics

Overall, 164 721 participants were included in the 13 trials that provided the quantitative data for this review (for individual study details, see table 1). Largely, the primary efficacy endpoint in these trials was composite of cardiovascular death, non-fatal MI,

Figure 1 PRISMA flow diagram for selecting studies. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

unstable angina requiring rehospitalisation, coronary revascularisation and fatal or non-fatal stroke. Two studies had two coprimaries: composite of death from cardiovascular causes, non-fatal MI or non-fatal stroke (first coprimaries outcome) and the composite of these events plus resuscitated cardiac arrest, heart failure or revascularisation (second coprimaries outcome). HRs for each coprimaries outcomes were similar and hence for pooled analysis, HR and CI of first coprimaries outcome were used. Sensitivity analysis including both the primary outcomes did not vary the final effect estimate considerably.

Out of 13, only 9 (69.2%) studies mentioned the clinically significant difference used to calculate their sample size. The range of these clinically significant differences was between 9.375 and 35 with an average value of 19.6. Studies which used annual rate of primary event or NI margin to calculate their sample size were excluded from calculating average of differences. Further, the primary hypothesis of four studies was to test NI with the use of NI margin and conclude superiority when the upper boundary of 95% CI was <1.0 .^{30 32 33 36} Overall, lower boundary of 95% CI of the included studies ranged from 0.56 to 0.89, while upper boundary of 95% CI ranged from 0.86 to 0.99. Incidence of the primary event was described as HR in all the 13 studies.

Synthesis of results

The overall effect estimate (pooled HR) was 0.86 with 95% CI of 0.84 to 0.89, $p < 0.00001$. The test for heterogeneity, I^2 in our meta-analysis was 45%, which indicates that there was moderate heterogeneity or inconsistency in the studies included, which conflicts with the fixed effect assumption. According to fixed effects model assumption, each study included in the meta-analysis is trying to estimate the same quantity of intervention effect⁴¹ (figure 2).

On an average, the sample size in each included study of our meta-analysis was 12 671 (range 3297–27 395). These participants experienced an ARR of 1.19% to 2.3% in the primary outcome which translated to a median NNT of 63 (range:

43–84) over a follow-up duration of 2.95 years. This follow-up duration was taken from the calculated median NNT (of 63) that corresponded to studies by Schwartz *et al*²⁵ and the Zinman *et al*,³⁶ who had a median follow-up duration of 2.8 years and 3.1 years, respectively (table 1), yielding a median duration of 2.95 years (midpoint of above studies) for the said NNT.

Risk of bias

Full details of the risk of bias are outlined in figure 3. All the included studies described their randomisation methods and were considered as low risk. Information about allocation concealment was unclear for two trials, rest of them used either interactive voice response or web response or generated unique number to conceal allocation. All the included trials were double-blinded except for two, in which one was single-blinded³⁰ and the other one was open label.³⁷ Only in one trial³⁹ even biostatisticians were blinded until analyses. Seven studies clearly mentioned that statisticians were not blinded, while it was unclear in rest of the studies. We did not look for incomplete outcome data and selective reporting as our study objective was to look for primary outcome of these studies.

DISCUSSION

The main objective of superiority trial is to determine the smallest difference in the primary outcome between the study arms.⁴³ Therefore, they are designed in such a way that they have enough power to detect those differences. However, in this meta-analysis, we debate on assigning meaning to the smallest difference making it clinically significant and more specific. Currently, the clinically significant difference is used to calculate the sample size but it is not considered at the time of analysis of findings.

In this study, the effect estimate of the included studies is 0.86 with 95% CI 0.84 to 0.89. It means that there is ~14% decrease in risk of having the primary outcome, if the patient consumes experimental drug. This 14% may not be clinically significant as the estimated mean of clinically meaningful difference of these studies is 19.6%. Nine (69.2%) out of 13 included studies did not meet their own respective marks for clinical significance, despite obtaining statistical significance (clinically meaningful difference was not mentioned for other four studies). Either their pooled HR was beyond their own clinically meaningful difference^{25 29 31 34 35 38 39} or their 95% CI^{27 37} included this value, which means had the authors considered clinically meaningful difference during their analysis, the null hypothesis would be accepted and superiority of experimental drug would be rejected. This leads us to think whether a statistically significant result is clinically relevant. In one of the included trials,³⁴ the investigators intended to evaluate 35% of lower risk with experimental drug.

Further, the effect estimate translates to a median NNT of 63 which means 63 patients need to be treated for a median duration of 2.95 years to prevent one extra outcome from happening. It is debatable whether such high NNT for the said time period is acceptable for a new, rather expensive medication with potential and unknown adverse effect profile, even more so because trial-based NNT often underestimates NNT in clinical practice.⁴⁴

There are also critics highlighting that with the existing design of superiority trials, even the borderline effect will become statistically significant if we test with large sample size.⁴⁵ In our study, the average sample size was ~12 600. In 11 of 13 included studies, the upper bound 95% CI crossed 0.90 and in two of 11 such studies, the upper bound 95% CI was 0.99,^{36 38} which

Table 1 Summary of study characteristics and calculation of ARR and NNT

Study	No of participants (% female)	Age (years;mean)	Primary outcome	Dose of the intervention group	Clinically meaningful difference	NI margin	Incident rate of primary event (test vs control)	Duration of the study (years; mean or median)	Absolute risk reduction (%)	Number needed to treat
Schwartz <i>et al</i> , 2018 (ODYSSEY OUTCOMES) ²⁵	18 924 (25.3)	58.6	Composite of death from coronary heart disease, non-fatal MI, fatal or non-fatal ischaemic stroke or unstable angina requiring hospitalisation.	Alirocumab: 75 mg, s.c, every 2 weeks	15%	NA	903 (9.5%) of 9462 versus 1052 (11.1%) of 9462 HR: 0.85 (95% CI: 0.78 to 0.93) p<0.001	2.8	1.6	63
Eikelboom <i>et al</i> , 2017 (COMPASS) ²⁸	27 395 (22)	68.2	Composite of cardiovascular death, stroke or MI.	Rivaroxaban 2.5 mg and aspirin 100 mg/day, oral	20%	NA	379 (4.1%) of 9152 patients versus 496 (5.4%) of 9126 patients; HR: 0.76 (95% CI: 0.66 to 0.86) p<0.001	1.92	1.3	77
Ridker <i>et al</i> , 2017 (CANTOS) ²⁹	10 061 (25.6)	61.1	First occurrence of non-fatal MI, non-fatal stroke or cardiovascular death.	Canakinumab: 50 mg, 150 mg, 300 mg, s.c, every 3 months	20%	NA	150 mg group (statistical significant group) 320 of 2284 (3.86 events per 100 person-years) versus 535 of 3344 (4.50 events per 100 person-years) HR: 0.85 (95% CI: 0.74 to 0.98) p=0.021	3.7	2	50
Neal <i>et al</i> , 2017 (CANVAS Programme) ³⁰	10 142 (35.8)	63.3	Composite of death from cardiovascular causes, non-fatal MI or non-fatal stroke.	Canagliflozin: 100 mg and 300 mg/day, oral	NA	1.3	Integrated experimental group, n=5735 versus n=4347 in control group 26.9 versus 31.5 participants per 1000 patient-years HR: 0.86 (95% CI: 0.75 to 0.97) p=0.02 for superiority	3.6	2.3	43
Sabatine <i>et al</i> , 2017 (FOURIER) ³¹	4465 (49.3)	58	Major cardiovascular events, defined as the composite of cardiovascular death, MI, stroke, hospitalisation for unstable angina or coronary revascularisation.	Evolocumab: 140 mg every 2 weeks or 420 mg monthly, s.c	15%	NA	1344 (9.8%) of 13 784 versus 1563 (11.3%) of 13 780 HR: 0.85 (95% CI: 0.79 to 0.92) p<0.001	0.93	1.5	67
Marso <i>et al</i> , 2016 (SUSTAIN-6) ³²	3297 (60.7)	64.6	First occurrence of cardiovascular death, non-fatal MI or non-fatal stroke.	Semaqlutide: 0.5 or 1 mg every week, s.c	Annual event rate of 1.98% in each group	1.8	108 (6.6%) of 1648 versus 146 (8.9%) of 1649 HR: 0.74 (95% CI: 0.58 to 0.95) p=0.02 for superiority	1.99	2.3	43
Marso <i>et al</i> , 2016 (LEADER) ³³	9340 (64.25)	64.3	First occurrence of death from cardiovascular causes, non-fatal MI or non-fatal stroke.	Liraglutide: 1.8 mg/day or the maximum tolerated dose, s.c	Annual event rate of 1.8% in each group	1.3	608 (13.0%) of 4668 versus 694 (14.9%) of 4672 HR: 0.87 (95% CI: 0.78 to 0.97) p=0.01 for superiority	3.8	1.9	53
Yusuf <i>et al</i> , 2016 (HOPE-3 trial) ³⁴	12 705 (46.25)	65.7	Composite of death from cardiovascular causes, non-fatal MI or non-fatal stroke (first coprimary outcome) and the composite of these events plus resuscitated cardiac arrest, heart failure or revascularisation (second coprimary outcome)	Rosuvastatin: 10 mg/day Cardesartan: 16 mg/day HCTZ: 12.5 mg/day; oral	35%	NA	First coprimary outcome: 113 (3.6%) of 3180 in the combined therapy group versus 157 (5.0%) of 3168 in the dual-placebo group HR: 0.71 (95% CI: 0.56 to 0.90) p=0.005 Second coprimary outcome: 136 (4.3%) of 3180 patients versus 187 (5.9%) of 3168 HR: 0.72 (95% CI: 0.57 to 0.89) p=0.003	5.6	1.5	67

Continued

Table 1 Continued

Study	No of participants (% female)	Age (years;mean)	Primary outcome	Dose of the intervention group	Clinically meaningful difference	NI margin	Incident rate of primary event (test vs control)	Duration of the study (years: mean or median)	Absolute risk reduction (%)	Number needed to treat
Yusuf <i>et al</i> , 2016 (HOPE-3 trial) ³⁵	12 705 (46.25)	65.7	Composite of death from cardiovascular causes, non-fatal MI or non-fatal stroke (first coprimary outcome) and the composite of these events plus resuscitated cardiac arrest, heart failure, or revascularisation (second coprimary outcome).	Rosuvastatin: 10 mg/day, oral	22.5%	NA	First coprimary outcome: 235 (3.7%) of 6361 versus 304 (4.8%) of 6344 HR: 0.76 (95% CI: 0.64 to 0.91) p=0.002 Second coprimary outcome: 277 (4.4%) of 6361 versus 363 (5.7%) of 6344 HR: 0.75 (95% CI: 0.64 to 0.88) p<0.001	5.6	1.2	83
Zimman <i>et al</i> , 2015 (EMPA-REG OUTCOME) ³⁶	7020 (28.4)	63.2	Composite of death from cardiovascular causes, non-fatal MI or non-fatal stroke, as analysed in the pooled empagliflozin group versus the placebo group.	Empagliflozin: 10 mg or 25 mg/day, oral	NA	1.3	490 (10.5%) of 4687 versus 282 (12.1%) of 2333 HR: 0.86 (95% CI: 0.74 to 0.99) p=0.04 for superiority	3.1	1.6	63
Wright <i>et al</i> , 2015 (The SPRINT Research Group) ³⁷	9361 (35.6)	67.9	Composite outcome of MI, other acute coronary syndromes, stroke, heart failure or death from cardiovascular causes.	NA	20%	NA	1.65% per year versus 2.19% per year; HR: 0.75 (0.64 to 0.89) p<0.001	3.26	1.7	59
Cannon <i>et al</i> , 2015 (IMPROVE-IT) ³⁸	18 144 (24.3)	63.6	Composite of cardiovascular death, non-fatal MI, unstable angina requiring rehospitalisation, coronary revascularisation (≥30 days after randomisation) or non-fatal stroke.	Ezetimibe: 10 mg/day and simvastatin: 40 mg; oral	9.375%	NA	2572 (32.7%) of 9067 (34.7%) versus 2742 of 9077 HR: 0.936 (95% CI: 0.89 to 0.99) p=0.016	6	2	50
Bonaca <i>et al</i> , 2015 (PEGASUS-TIMI) ³⁹	21 162 (23.5)	65.3	Composite of cardiovascular death, MI or stroke.	Ticagrelor: 90 or 60 mg two times per day; oral	20% lower RR in 90 mg 19% lower RR in 60 mg	NA	90 mg 493 (7.85%) of 7050 versus 578 (9.04%) of 7067 HR=0.85 (95% CI: 0.75 to 0.96) p=0.008 60 mg 487 (7.77%) of 7045 versus 578 (9.04%) of 7067 HR: 0.84 (95% CI: 0.74 to 0.95) p=0.004	2.75	1.19	84

ARR, absolute risk reduction; MI, myocardial infarction; NI, non-inferiority; NNT, number needed to treat; s.c, subcutaneous.

Original research

Figure 2 Meta-analysis of primary efficacy endpoint of included trials using fixed effect model.

is very close to line of no difference. We believe that the large sample size has given enough power to rule out null hypothesis in these studies. Such ‘overpowered or oversized trials’ have been criticised earlier by many researchers.^{2 46 47}

It is a known fact that CIs offer predictive information about treatment effects that can be expected posttrial use. Narrow confidence intervals suggest little variability within the population and assure that even a wider interval, which we are bound to get once the drug is marketed, may provide similar difference in treatment effect in general population.⁴⁸ Given the wide CIs and their close approximation to line of no difference in most of the included studies, there is an uncertainty regarding the effectiveness of such drugs after being marketed.

Nevertheless, as stated before, while analysing the results of the study, it is important to assess the relevance of difference between the means of two drugs. In NI trials, inclusion of NI margin addresses the issue of clinical relevance.⁴⁹ We have earlier proposed the need for superiority margin for superiority trials similar to NI trials.^{3 6} With this study, we emphasise fixing the clinically meaningful difference for that particular outcome (suppose say ‘10%’ for ‘composite of death from coronary heart disease, non-fatal MI, fatal or non-fatal ischaemic stroke or unstable angina requiring hospitalisation’) as a ‘superiority margin’. It makes sure that only when 95% CI is beyond this margin, the experimental drug will be deemed superior. The proposed margin can be contemplated based on research question and entire active control effect, similar to NI margin.

Whether there should be a guideline to the difference to be detected in a trial is debatable. Some authors argue that each research question is unique and deserves special consideration. They believe that guidelines encourage the same approach being used and enough attention may not be paid to individual differences towards all study designs.⁵ However, absence of guidelines

may invite carelessness and engage investigators in playing with sample size, CI and p-value.

This study has a few limitations: First, we included articles published in NEJM only, hence, we cannot generalise the study conclusion for trials published in other journals. However, we believe that NEJM is sufficiently representative of leading medical journals and these results should be quite robust. Second, since a single journal was selected for review, only one author did the initial screen followed by all three authors. Third, we did not register our study protocol in PROSPERO which is typically recommended for systematic reviews and meta-analyses. Fourth, we did not include effect size of four studies to calculate average clinical meaningful difference as the effect size in these studies was not mentioned. They have used either annual rate of events or NI margin to calculate sample size. Owing to similar primary efficacy outcome in all the included studies, we presume that the authors would have considered similar effect sizes ranging from 9.375% to 35% to calculate sample size.

This study also raises some important questions such as, ‘Is there a need of superiority margin for superiority trials? Which is the best parameter that can be used to derive superiority margin? and What are the pros and cons of choosing clinically significant difference as superiority margin? Or should it be modified?’

CONCLUSION

In the absence of a superiority margin, the current practice leads to a situation where overpowered superiority trials provide statistically significant results with effect size so small that they may be of little clinical relevance. This meta-analysis shows that this practice is widespread in cardiovascular diseases and diabetic trials published in NEJM. We hope to generate debate on such practice.

Main messages

- ▶ Overpowered cardiovascular trials provide statistical significance but may have little clinical relevance.
- ▶ Superiority margin must be taken into account while interpreting results of such trials in order to avoid scientific misinterpretation.
- ▶ Clinically significant difference used to calculate sample size may be used as the superiority margin.

Figure 3 Risk of bias summary.

Current research questions

- Is there a need of superiority margin for superiority trials?
- Which is the best parameter that can be used to derive superiority margin?
- What are the pros and cons of choosing clinically significant difference as superiority margin? Or should it be modified?

What is already known on the subject

- Superiority trials use traditional null hypothesis statistical testing to decide the statistical significance between one or more experimental arms. Such analysis has the potential to misinterpretations and scientific misuse of superiority trials.

Contributors NG and SM hand-searched the articles and independently reviewed the articles. Disagreements were discussed with NS. All three authors contributed in data analysis and manuscript preparation.

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 None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement There are no data in this work.

ORCID iD

Samir Malhotra <http://orcid.org/0000-0002-6049-1628>

REFERENCES

- Sackett DL. Superiority trials, noninferiority trials, and prisoners of the 2-sided null hypothesis. *ACP J Club* 2004;140:38–9.
- Bhardwaj SS, Camacho F, Derrow A, et al. Statistical significance and clinical relevance: the importance of power in clinical trials in dermatology. *Arch Dermatol* 2004;140:1520–3.
- Shafiq N, Malhotra S. Superiority trials: statistical trickery or mass blindness? *Postgrad Med J* 2016;92:118–9.
- Keefe RSE, Kraemer HC, Epstein RS, et al. Defining a clinically meaningful effect for the design and interpretation of randomized controlled trials. *Innov Clin Neurosci* 2013;10:45–19.
- Gayet-Ageron A, Agoritsas T, Combesure C, et al. What differences are detected by superiority trials or ruled out by noninferiority trials? A cross-sectional study on a random sample of two-hundred two-arms parallel group randomized clinical trials. *BMC Med Res Methodol* 2010;10:93.
- Shafiq N, Malhotra S. Superiority trials: raising the bar of null hypothesis statistical testing. *Evid Based Med* 2015;20:154–5.
- Gamad N, Shafiq N, Malhotra S. Effect size in CANTOS trial. *BMJ Evid Based Med* 2018;23:44.
- Vasileiou K, Barnett J, Thorpe S, et al. Characterising and justifying sample size sufficiency in Interview-Based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Med Res Methodol* 2018;18:148.
- Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. *JAMA* 2018;319:1429–30.
- Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates, 1988.
- Sedgwick P. Measuring the benefit of treatment: number needed to treat. *BMJ* 2015;350:h2206.
- Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;151:264–9.
- Robinson JG, Farnier M, Krempf M, et al. Odyssey long term Investigators. efficacy and safety of alirocumab in reducing lipids and cardiovascular events. *N Engl J Med* 2015;372:1489–99.
- Pi-Sunyer X, Astrup A, Fujioka K, et al. A randomized, controlled trial of 3.0 Mg of liraglutide in weight management. *N Engl J Med* 2015;373:11–22.
- Gaudet D, Alexander VJ, Baker BF, et al. Antisense inhibition of apolipoprotein C-III in patients with hypertriglyceridemia. *N Engl J Med* 2015;373:438–47.
- Thabit H, Tauschmann M, Allen JM, et al. Home use of an artificial beta cell in type 1 diabetes. *N Engl J Med* 2015;373:2129–40.
- Syngelaki A, Nicolaides KH, Balani J, et al. Metformin versus placebo in obese pregnant women without diabetes mellitus. *N Engl J Med* 2016;374:434–43.
- Stewart ZA, Wilinska ME, Hartnell S, et al. Closed-Loop insulin delivery during pregnancy in women with type 1 diabetes. *N Engl J Med* 2016;375:644–54.
- Ray KK, Landmesser U, Leiter LA, et al. Inclisiran in patients at high cardiovascular risk with elevated LDL cholesterol. *N Engl J Med* 2017;376:1430–40.
- Ridker PM, Tardif J-C, Amarenco P, et al. Lipid-reduction variability and antidrug-antibody formation with bococizumab. *N Engl J Med* 2017;376:1517–26.
- Calkins H, Willems S, Gerstenfeld EP, et al. RE-CIRCUIT Investigators. uninterrupted dabigatran versus warfarin for Ablation in atrial fibrillation. *N Engl J Med* 2017;376:1627–36.
- Garg SK, Henry RR, Banks P, et al. Effects of Sotagliflozin added to insulin in patients with type 1 diabetes. *N Engl J Med* 2017;377:2337–48.
- Graham MJ, Lee RG, Brandt TA, et al. Cardiovascular and metabolic effects of ANGPTL3 antisense oligonucleotides. *N Engl J Med* 2017;377:222–32.
- Perkins GD, Ji C, Deakin CD, et al. A randomized trial of epinephrine in out-of-hospital cardiac arrest. *N Engl J Med* 2018;379:711–21.
- Schwartz GG, Steg PG, Szarek M, et al. Alirocumab and cardiovascular outcomes after acute coronary syndrome. *N Engl J Med* 2018;379:2097–107.
- Bowman L, Hopewell JC, Chen F, et al. Effects of anacetrapib in patients with atherosclerotic vascular disease. *N Engl J Med* 2017;377:1217–27.
- Bowman L, Mafham M, Wallendszus K, et al. Effects of aspirin for primary prevention in persons with diabetes mellitus. *N Engl J Med* 2018;379:1529–39.
- Eikelboom JW, Connolly SJ, Bosch J, et al. Rivaroxaban with or without aspirin in stable cardiovascular disease. *N Engl J Med* 2017;377:1319–30.
- Ridker PM, Everett BM, Thuren T, et al. Antiinflammatory therapy with canakinumab for atherosclerotic disease. *N Engl J Med* 2017;377:1119–31.
- Neal B, Perkovic V, Mahaffey KW, et al. Canagliflozin and cardiovascular and renal events in type 2 diabetes. *N Engl J Med* 2017;377:644–57.
- Sabatine MS, Giugliano RP, Keech AC, et al. Evolocumab and clinical outcomes in patients with cardiovascular disease. *N Engl J Med* 2017;376:1713–22.
- Marso SP, Bain SC, Consoli A, et al. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *N Engl J Med* 2016;375:1834–44.
- Marso SP, Daniels GH, Brown-Frandsen K, et al. Liraglutide and cardiovascular outcomes in type 2 diabetes. *N Engl J Med* 2016;375:311–22.
- Yusuf S, Lonn E, Pais P, et al. Blood-Pressure and cholesterol lowering in persons without cardiovascular disease. *N Engl J Med* 2016;374:2032–43.
- Yusuf S, Bosch J, Dagenais G, et al. Cholesterol lowering in intermediate-risk persons without cardiovascular disease. *N Engl J Med* 2015;373:2101–11.
- Zinman B, Wanner C, Lachin JM, et al. Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. *N Engl J Med* 2015;373:2117–28.
- Wright JT, Williamson JD, Whelton PK, et al. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med* 2015;373:2103–16.
- Cannon CP, Blazing MA, Giugliano RP, et al. Ezetimibe added to statin therapy after acute coronary syndromes. *N Engl J Med* 2015;372:2387–97.
- Bonaca MP, Bhatt DL, Cohen M, et al. Long-Term use of ticagrelor in patients with prior myocardial infarction. *N Engl J Med* 2015;372:1791–800.
- Cannon CP, Bhatt DL, Oldgren J, et al. Dual antithrombotic therapy with dabigatran after PCI in atrial fibrillation. *N Engl J Med* 2017;377:1513–24.
- Higgins JPT, Green S. Cochrane Handbook for systematic reviews of interventions version 5.1.0. The Cochrane collaboration, 2011. Available: www.handbook.cochrane.org [Accessed Mar 2011].
- Tierney JF, Stewart LA, Ghersi D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;8:16.
- Rehal S, Morris TP, Fielding K, et al. Non-Inferiority trials: are they inferior? A systematic review of reporting in major medical journals. *BMJ Open* 2016;6:e012594.
- McAlister FA. The "number needed to treat" turns 20—and continues to be used and misused. *CMAJ* 2008;179:549–53.
- Dahiru T. P-value, a true test of statistical significance? A cautionary note. *Ann Ib Postgrad Med* 2008;6:21–6.
- Hochster HS. The power of "p": on overpowered clinical trials and "positive" results. *Gastrointest Cancer Res* 2008;2:108–9.
- Tavernier E, Giraudeau B. Sample size calculation: inaccurate a priori assumptions for nuisance parameters can greatly affect the power of a randomized controlled trial. *PLoS One* 2015;10:e0132578.
- Landow L. Current issues in clinical trial design: superiority versus equivalency studies. *Anesthesiology* 2000;92:1814–20.
- Lesaffre E. Superiority, equivalence, and non-inferiority trials. *Bull NYU Hosp Jt Dis* 2008;66:150–4.