Journal Title
XX(X):1–9
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Choosing and changing the analysis scale in non-inferiority trials with a binary outcome

Zhong Li[1], Matteo Quartagno[2], Stefan Böhringer[3] and Nan van Geloven[3]*

Abstract

Background: The size of the margin strongly influences the required sample size in non-inferiority and equivalence trials. What is sometimes ignored however, is that for trials with binary outcomes the scale of the margin - absolute risk difference (ARD), risk ratio (RR) or odds ratio (OR)- has a large impact on power and thus on sample size requirement as well. When considering several scales at the design stage of the trial, these sample size consequences should be taken into account. Sometimes, changing the scale may be desirable at a later stage of the trial, e.g. when the event proportion in the control arm turns out different than expected. Also after completion of a trial a switch to another scale may be attractive, e.g. when using a regression model in a secondary analysis or when combining study results in a meta-analysis that requires unifying scales. The exact consequences of such switches are currently unknown.

Methods: This paper first outlines sample size consequences for different choices of analysis scale at the design stage of the study. We add a new result on sample size requirement comparing the ARD scale with the RR scale. Then we study two different approaches to changing the analysis scale after the trial has commenced: 1) mapping the original NI margin using the event proportion in the control arm that was anticipated at the design stage or 2) mapping the original NI margin using the observed event proportion in the control arm. We use simulations to illustrate consequences on type I and type II error. Methods are illustrated on the INES trial, a non-inferiority trial that compared single birth rates in subfertile couples after different fertility treatments.

Results: Our results demonstrate large differences in required sample size when choosing between ARD, RR and OR scales at the design stage of non-inferiority trials. In some cases the sample size requirement is twice as large on one scale compared to another. Changing the scale after commencing the trial using anticipated proportions mainly impacts type II error, whereas switching using observed proportions is not advised due to not maintaining type I error. Differences were more pronounced with larger margins.

Conclusions: Trialists should be aware that the analysis scale can have unexpectedly large impact on type I and type II error rates in non-inferiority trials.

Keywords

non-inferiority trial, non-inferiority margin, absolute risk difference, risk ratio, odds ratio, sample size calculation

Introduction

For ethical reasons, in several disease areas it is becoming increasingly difficult to justify testing the efficacy of new treatments against placebo. Instead, active controlled trials are being used to test whether a new treatment which may be cheaper, safer, less invasive or easier to use, has no worse efficacy than an already known effective treatment (Kaul and Diamond (2006)). No worse efficacy is defined as the difference between the new and the known effective

[1] Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, NL
[2] Institute for Clinical Trials and Methodology, University College London, 90 High Holborn, WC1V 6LJ, UK
[3] Department of Biomedical Data Sciences, Section Medical Statistics, Leiden University Medical Center, Leiden, NL

Corresponding author:
Nan van Geloven, Leiden University Medical Center, Building 2, Room S05-44, Einthovenweg 20, 2333 ZC Leiden, The Netherlands.
Email: N.van_Geloven@lumc.nl

treatment being bounded by a pre-specified margin that is still considered clinically acceptable (Snapinn (2000)). As pointed out by Mauri and D'Agostino (2017), the use of such non-inferiority (NI) trials has increased vastly over the last decades.

Choosing the non-inferiority margin, which defines what we consider 'not unacceptably worse', is a pivotal step in designing non-inferiority trials. It is well known that the size of the margin strongly influences the required sample size. What is sometimes ignored however, is that also the scale of the margin - for binary endpoints absolute risk difference (ARD), risk ratio (RR) or odds ratio (OR) – has a strong impact on the power of the trial and thus on the required sample size. Under seemingly equal assumptions, different scales for the analysis and the non-inferiority margin may lead to significantly different sample size requirements. Though this phenomenon has been pointed out in some statistical papers (Wellek (2005); Rousson and Seifert (2008); Hilton (2008)), it is not known to all trialists. Online sample size calculators often fail to offer the option of specifying the non-inferiority hypothesis on all three scales, in such instances typically only facilitating input on the risk difference scale. No comprehensive overview exists in which all three scales of choice are compared for different design settings. The aim of this paper is to provide such an overview.

Considering different analysis scales is common at the design stage of a trial. However, also after the trial has already commenced, there may be unforeseen situations that warrant reconsidering the scale. In the first place, when the observed risk in the control arm turns out lower or higher than expected, for example during a blinded review of the data, an initially defined absolute margin may no longer be deemed appropriate. In studies of bacterial pneumonia, an absolute non-inferiority margin of 10% is deemed acceptable by the FDA when studying all cause mortality (FDA (2020)). However, as shown in Talbot et al. (2019), if a certain trial was designed with such a ARD margin, but then observed that only 10-15% of the control patients died, the potential for loss of clinically acceptable efficacy with an absolute 10% margin may be judged too great. A smaller non-inferiority margin may be achieved by changing to a risk ratio or odds ratio scale. Authoritative trials in other disease areas faced similar challenges (Kaul et al. (2005); Schulz-Schüpke et al. (2015); Li et al. (2019)). A second situation where a scale switch may be considered is when a regression model is used in the analysis phase, e.g. for covariate adjustment in sensitivity analyses or in per protocol analyses (Zhang et al. (2008); Moore and Van der Laan (2009); Kahan et al. (2014)), or for clustering adjustment in cluster randomized trials. Though attempting to obtain results on the originally planned scale from such regression approaches may be better practice, for example through marginalization (Zhang et al. (2019)), sometimes a switch in the analysis scale is made. Lastly, when non-inferiority studies are combined in a meta-analysis as stated in Acuna et al. (2020), converting the scale of the analysis is necessary to allow pooling of study results.

Note that the decision to adjust the scale of the analysis should never be based on the observed comparative (between-arm) outcomes from the study, as this would invalidate results. In line with the potential reasons for switching analysis scale listed above, we assume in the remainder of this paper that the decision to change the scale is independent of the between-arm results.

We performed a search for non-inferiority trials with binary outcomes reported in the *New England Journal of Medicine* between 2016 and 2019. Of the 24 RCTs found, the majority (16) used an absolute difference to specify the NI margin. Two used risk ratio and six used odds ratio. In 9 papers a different scale than the scale of the main analysis was used to report trial results and/or make an additional analysis. In 2 papers the non-inferiority margin was changed related to observing higher or lower than expected event rates (Li et al. (2019) and Widmer et al. (2018)). Noticeably in the latter paper non-inferiority could be proven only on the ARD scale and not on the RR scale.

Our contribution in this paper is threefold. First, we describe sample size consequences when choosing between different scales at the design stage of a trial. We present a new result about how sample size changes when choosing the RR scale compared to the RD scale. Secondly, we describe changing the scale at a later stage during the trial. We provide a comprehensive overview of type I and type II error rates of two ways of mapping the NI margin using simulations. We provide intuition about our results by studying rejection regions. We illustrate the potential impact of the non-inferiority scale in a real trial (the INES trial). Our results can be used by trialists when choosing the NI scale at the design stage, and when considering performing an analysis on a different scale than the one chosen at the design stage.

Choosing between different scales at the design stage

Sample size calculation in the INES trial

As a case study, we consider the INES trial that compared two types of in vitro fertilization (IVF) to intrauterine insemination (IUI) treatment in couples with unexplained subfertility (Bensdorp et al. (2009)). A non-inferiority design was chosen since IVF was expected to prevent more risky twin pregnancies and a slightly lower single birth rate compared to IUI would be acceptable for that reason. The trial was designed anticipating a success rate of 40%, i.e., patients achieving a singleton pregnancy within 1 year, in the IUI control arm (with either no pregnancy or a non-singleton pregnancy counting as failure). A minimum success rate of 27.5% in the IVF treatment arm was considered clinically acceptable based on the anticipated advantage of lower multiple birth rates with IVF. Under the assumption of no real difference between treatments, planning for 80% power and 5% one-sided significance level, the study aimed to exclude an absolute risk difference of more than -12.5% (27.5% minus 40%), requiring 190 patients per arm. Had the study instead targeted the relative risk, aiming to exclude a risk ratio of 0.69 (27.5% divided by 40%), considerably less patients (133 per arm) would have been needed. Strikingly, if the same percentages were formulated as failure rates instead of success rates, i.e., the percentage of patients not achieving a singleton pregnancy within one year, excluding a risk ratio of 1.21 (72.5% divided by 60%) would require much more patients per arm (235). The fact that the two versions of RR require very different sample sizes may cause confusion to the trial designers. Triggered by this somewhat paradoxical finding, we aimed to systematically examine the effect of the analysis scale in a broad range of design settings. We use some of the design parameters of the INES study as starting point in our explorations.

Notation

We will focus on a two-arm trial with a binary outcome. The data collected in such a trial can be summarized by the success proportions in both arms, estimated from the observed frequencies in the treatment and control arms respectively: $\hat{p}_t = x_t/n_t$, $\hat{p}_c = x_c/n_c$. We will refer to success proportions throughout, but all arguments could also be made using the failure (or event) proportions as data summary. We will denote by p_c^* and p_t^* the anticipated success proportions during sample size planning (often $p_c^* = p_t^*$). With p_c and p_t we will denote the 'true', unknown success proportions in the control and treatment arm. The treatment effect can be evaluated on four different scales, with δ_{RD} the NI margin on the ARD scale, δ_{OR} the margin on the OR scale, δ_{RR} the margin on the RR scale using success rates and δ_{RR}^f the margin on the RR scale using failure rates, summarized in Supplementary Table 2.

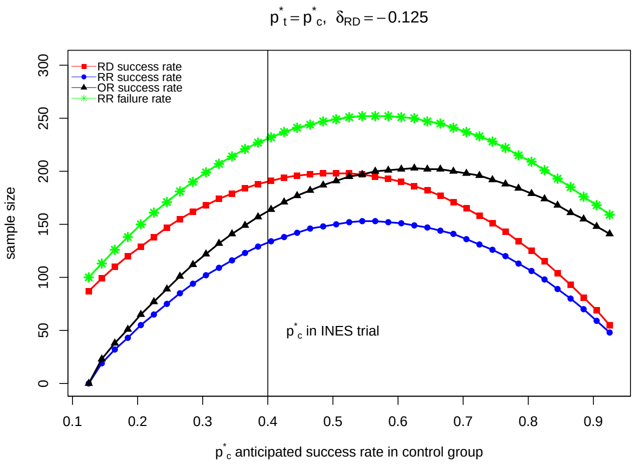

Figure 1. Comparison of sample size when considering different analysis scales at the design stage of the study assuming the boundary proportions for the success rate in treatment group is the same for each scale.

Structural comparison of sample sizes

We compared sample size requirements when considering the four analysis scales, mapping the NI margin in the way illustrated in the INES case study and described more generally in Appendix 5. We rely on the large sample approximation of the (unpooled) Z-test for the sample size calculations (Supplementary Table 2). The results were highly similar when using other sample size approaches relying on improved approximations (Farrington and Manning (1990)). The difference between the required sample sizes when considering different scales are shown in Figure 1 along a range of control proportions, using a NI margin of $\delta_{RD} = -0.125$, as was used in the INES trial. In Appendix 1, we show similar plots for smaller NI margins.

The results show that the differences in sample size needed for different scales as described for the INES trial (vertical line in Figure 1) is not an exception. Differences in required sample size when considering different analysis scales can be large, up to two times more when comparing the RR using success rates (bottom line in Figure 1) with the RR using failure rates (top line in Figure 1).

Analytical results

Some of the results we show in Figure 1 can be proven analytically. A comparison of sample size requirement for RD scale and OR scale was given in Rousson and Seifert (2008): under the assumptions that $p_c^* = p_t^*$ and $n_t = n_c$ and for some given value of δ_{OR}, one has that the power when using the ARD scale is larger than when using the OR scale as soon as $p_c^* \geq \frac{1}{1-\delta_{OR}} + \frac{1}{\ln(\delta_{OR})}$. This result coincides with

Figure 1. For $p_c^* = 0.40$ and the minimal acceptable success rate of 0.275 in the treatment arm, as in the INES trial, the δ_{OR} is 0.569. According to the result by Rousson and Seifert (2008), the sample size required for the ARD (line with red triangles in Figure 1) should be lower than that needed for the OR (line with black squares in Figure 1) for values of p_c^* greater than 0.547 and that is exactly where the lines cross.

We here add a proof of the sample size requirements when comparing the ARD scale to the RR scale with success proportions. Under the assumption that $p_c^* = p_t^*$ and $n_c = n_t$, one has that

$$\begin{aligned}\frac{n_{RD}}{n_{RR}} &= \frac{\frac{2(z_{1-\alpha}+z_{1-\beta})^2 p_c^*(1-p_c^*)}{(\delta_{RD})^2}}{\frac{2(z_{1-\alpha}+z_{1-\beta})^2 \frac{(1-p_c^*)}{p_c^*}}{(\ln(\delta_{RR}))^2}} \\ &= \frac{(p_c^*)^2}{(\delta_{RD})^2}\Big(\ln(p_c^* + \delta_{RD}) - \ln(p_c^*)\Big)^2 \\ &= \frac{1}{(\frac{\delta_{RD}}{p_c^*})^2}\Big(\ln\big(1+\frac{\delta_{RD}}{p_c^*}\big)\Big)^2 \\ &= \Big(\frac{\ln(1+x)}{x}\Big)^2, \text{with } \mathrm{x} = \frac{\delta_{\mathrm{RD}}}{\mathrm{p}_\mathrm{c}^*} \in (-1, 0) \\ &\in (1, +\infty)\end{aligned}$$

This shows that the sample size needed using the RR scale with success proportions (blue dotted line at the bottom in Figure 1) is always lower than the sample size needed using the ARD scale (line with red squares in Figure 1).

Changing the scale at the analysis stage

Re-analysis of the the INES trial

Based on observed single pregnancy rates in the 602 study participants (52% for the 201 patients allocated to IVF-SET, 43% for the 194 patients allocated to IVF-MNC and 47% for the 207 IUI patients respectively), both IVF-SET and IVF-MNC arms were concluded to be non-inferior to IUI (Bensdorp et al. (2015)). As pointed out in Van Geloven (2015), the trial reported results on the RR scale whereas the sample size calculation had been based on the ARD scale. A recalculation of the main study results using different scales shows that the trial could have reached a different conclusion had it been analysed on the ARD scale (Table 1, Van Geloven (2015)). As shown in Table 1, regardless of the scale used to report the results, IVF-SET can consistently be concluded to be non-inferior to IUI-COH. However, if one uses different scales to report the results of IVF-MNC vs IUI-COH, the conclusions are inconsistent. Specifically, when the OR or the RR with success rate is used, one can draw the conclusion that IVF-MNC is non-inferior to IUI. On the contrary, one cannot conclude that IVF-MNC is non-inferior to IUI when the ARD or the RR with failure rate is used. Particularly, the contradictory conclusions drawn by using the RR scale with success rate and failure rate respectively, may pose a dilemma for trialists as to whether the non-inferiority of IVF-MNC to IUI should be accepted.

Structural comparison of type I and type II error

When a change in the scale is made after the trial has commenced, sample size calculation has already been performed and is no longer of main interest. Therefore, for such switches we examined power, i.e., one minus type II error rate, and type I error rate, based on simulations assuming a fixed sample size. We consider two ways of mapping the non-inferiority margin to the new scale: either based on the anticipated control proportion (similar to what was done in the INES trial and in Widmer et al. (2018)), or based on the observed control proportion similar to what has been proposed in Quartagno et al. (2020). In the latter case, again starting with an absolute risk difference, this means that the NI margin is added to the observed success proportion in the control arm $\hat{p}_c$ to come to the minimum allowed success proportion in the treatment arm, $p_t^{inf,2}$. By comparing $p_t^{inf,2}$ and $\hat{p}_c$, the new margins on the risk ratio scale and the odds ratio scale can be chosen, see Supplementary Table 3.

a. Comparison of power We simulated the success proportions of 100000 trials similar in setup to the INES trial (sample size 190, one sided $\alpha = 0.05$, $\delta_{RD} = -0.125$) using binomial distributions according to the alternative hypothesis with anticipated proportions $p_t^* = p_c^* = 0.40$. Power was calculated as the proportion of trials in which H_0 was correctly rejected.

Results are presented in Figures 2(a) and 2(b). One can see that under these settings, the power for the RR scale using success rate is always the highest (top blue dotted line), while the power for the RR scale using failure rate is always the lowest (bottom green line with stars). In addition, the power on the RD scale and the OR scale lie between them, crossing at some point. This shows that power increases when switching from the RD scale to the RR scale, both when using the anticipated and when using the observed control proportion during mapping of the NI margin. The differences in power when switching using the anticipated control proportion are larger than when using observed control proportion according to our simulations.

b. Comparison of type I error rate In the simulation for type I error rate, we simulated 100000 trials with similar

Table 1. Recalculation of the INES trial main study results. Confidence intervals were calculated by score method. The risk difference margin is -12.5% (27.5-40%), the relative ratio margin with success rate is 0.69 (27.5%/40%), the odds ratio margin is 0.57 ((27.5%/72.5%)/(40%/60%), and the relative ratio margin with failure rate is 1.21 (72.5%/60%). Two sided p-values are presented.

Comparison	Margin Type	Estimate	95% Confidence Interval	P-value for NI	conclusion
IVF-SET vs IUI	RD	5%	(-5% to 14%)	<0.001	NI met
	RR success rate	1.11	(0.91 to 1.35)	<0.001	NI met
	OR	1.22	(0.82 to 1.79)	<0.001	NI met
	RR failure rate	0.91	(0.75 to 1.10)	0.003	NI met
IVF-MNC vs IUI	RD	-4%	(-14% to 6%)	0.090	NI failed
	RR success rate	0.91	(0.73 to 1.13)	0.012	NI met
	OR	0.85	(0.57 to 1.26)	0.048	NI met
	RR failure rate	1.08	(0.90 to 1.29)	0.195	NI failed

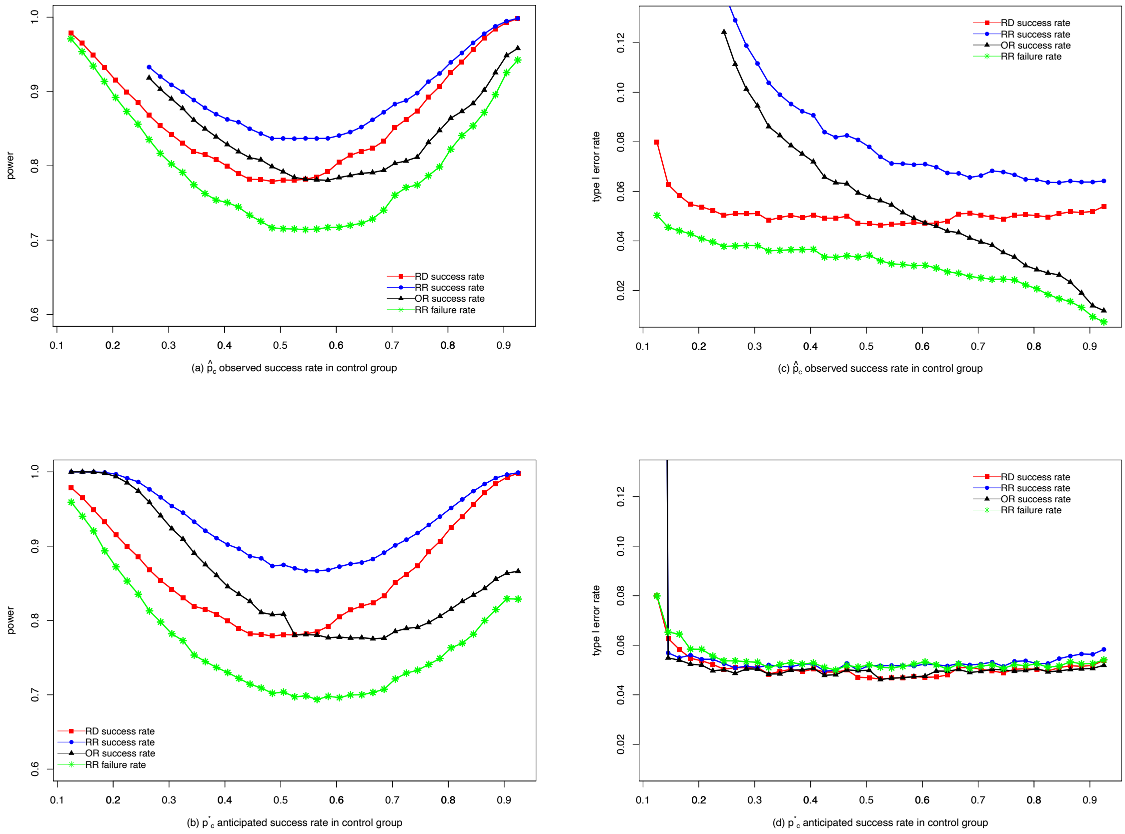

Figure 2. Comparisons of power and type I error rate. (a) comparison of power when mapping using the observed control proportion; (b) comparison of power when mapping using the anticipated control proportion; (c) comparison of type I error rate when mapping using the observed control proportion; (d) comparison of type I error rate when mapping using the anticipated control proportion

design but now with success proportions according to the null hypothesis $p_t = p_c + \delta_{RD}$, where the trial is originally designed on the risk difference scale. Type I error rate was calculated as the proportion of trials that incorrectly rejected H_0. When switching using the anticipated control proportion, the type I error rates on different scales were close to each other, all wiggling between 0.05 and 0.055 (Figure 2(d)). When the NI margin is mapped using the observed control proportion, it can be seen that the type I error rate on the RR scale with success proportion is unacceptably high on all occasions, whereas the type I error rate on the RR scale using failure proportion is too low. Moreover, the type I error rates on the RD scale and the OR scale are in-between and cross at around 60% observed control success rate. One can infer that the adaptive nature of this way of mapping fails to preserve type I error and should therefore not be advised.

c. Understanding the differences in type I and type II error rates through rejection regions The differences in type I and type II error rates that we found can best be explained by looking at rejection regions. We show these as region-plots of the results of the simulated trials in Appendix 2 for switching using anticipated proportions and in Appendix 3 for switching using observed proportions. The figures make clear that analyses on different scales will agree on rejecting the null hypothesis or not in trials where the observed rates ($\hat{p}_c$ and $\hat{p}_t$) are close to the anticipated proportions (p_c^* and p_t^*). However, due to chance variations, part of the trials will have a larger than expected success rate in the control arm and/or a lower than expected success rate in the treatment arm. In such trials, analyses on different scales will reach different conclusions on rejecting. We present rejection regions based on simulations for other designs, e.g., designs with unequal anticipated success rates in Appendix 4.

Discussion

We showed that unexpected differences in sample size requirements can occur when considering different analysis scales at the design stage of a non-inferiority trial. Changing the scale at the analysis stage using anticipated proportions for mapping the NI margin mainly impacts power. By studying rejection regions, we made clear that these results are not due to different inference (e.g. larger standard errors), but instead are caused by the fact that the choice of a particular scale plus non-inferiority margin defines a full rejection region. The regions of two scales only coincide when observed success rates are close to anticipated ones, but will differ when the observed proportions deviate from expectations. Moreover, even if we use the same scale to design and analyze a non-inferiority trial, using the RR scale with success rate and failure rate respectively, may lead to contradictory conclusions. This questions the appropriateness of using RR for non-inferiority trials. Mapping the non-inferiority margin relative to the observed success proportion in the control arm introduces problems as the evaluation criteria become too dependent on random low or high observed proportions. This is reflected in strongly in- or deflated type I error rates and matches the adaptive nature of the method. In general we advise against such data-dependent mapping. If it is considered, then a correction for type I error rate inflation must be used. Some advice for simulation-based correction methods are given in Quartagno et al. (2020).

The issues we describe are particularly important for non-inferiority trials since changing the analysis scale requires redefining the non-inferiority margin. In superiority trials the neutral comparison values (zero for the ARD and one for the RR or OR) do not change when switching the analysis scale such that no large differences between scales are expected. Analysing a trial in a different way than designed is considered bad practice in general. Whenever possible, we advise to keep the assessment of non-inferiority on the originally planned analysis scale. If the analysis (e.g. a regression model) is performed on another scale, marginalisation techniques can be used to report end results on the original scale. But as explained in the introduction section, changes may not be avoidable at times. A change in analysis scale should not be made lightly. Changing the scale means that trialists commit to a different boundary region of what they accept as clinically acceptable difference. It means that they realised that the original scale used was not correct. In fact, the results we describe about maintained type I error rates when mapping the margin based on anticipated event rate would not hold under the null of the original scale as soon as the true control event varies. It only holds if the relevant null situation is formulated according to the new scale. Our results should also not be read as encouragement to change the scale of a trial to gain power. Increased (or decreased) power can be a consequence of changing the scale but it should never be the reason for changing as the clinical judgement on what is an acceptable margin cannot be overruled by statistical arguments.

To avoid having to change the scale we recommend to consider at the design stage all clinical and trial size implication including scenarios where the anticipated event rates turn out higher or lower than expected and discuss whether the chosen margin would still suffice in such a situation. If a switch is unavoidable, we strongly recommend against switching based on the final observed event rates, but to use anticipated rates instead. Anticipated rates could potentially be updated based on blinded interim analysis but we did not study this in detail. Quartagno et al. (2020) recently proposed a more flexible way of defining the non-inferiority region, recommending the use of the arc-sine scale because of its power-stabilising properties.

We hope to have made clear that switching the scale in a non-inferiority trial is not without consequences and trialists should consider the impact on type I and type II error before such a switch is made.

Acknowledgements

We thank prof JGP Tijssen, prof AH Zwinderman and A Dane for fruitful discussions on earlier versions of this work.

Declaration of Conflicting Interests

The authors declare that there is no conflict of interest.

Funding

Matteo Quartagno's work in this paper was supported by the Medical Research Council [MC_UU_12023/29]. The other authors, i.e., Zhong Li, Stefan Böhringer and Nan van Geloven, received no financial support for the research, authorship, and/or publication of this article.

References

Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. *Ann Intern Med* 2006; 145(1): 62-69.

Snapinn SM. Noninferiority trials. *Trials* 2000; 1(1): 19.

Mauri L, D'Agostino RB. Challenges in the design and interpretation of noninferiority trials. *N Engl J Med* 2017; 377(14): 1357-1367.

Bensdorp AJ, Slappendel E, Koks C, et al. The INeS study: prevention of multiple pregnancies: a randomised controlled trial comparing IUI COH versus IVF e SET versus MNC IVF in couples with unexplained or mild male subfertility. *BMC Womens Health* 2009; 9(1): 1-8.

Wellek S. Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes. *Biom J* 2005; 47(1): 48-61.

Rousson V, Seifert B. A mixed approach for proving non-inferiority in clinical trials with binary endpoints. *Biom J* 2008; 50(2): 190-204.

Hilton JF. Noninferiority trial designs for odds ratios and risk differences. *Stat Med* 2010; 29(9): 982-993.

Food and Drug Administration, U.S. Department of Health and Human Services, Center for Drug Evaluation and Research. Guidance for Industry: Hospital-Acquired Bacterial Pneumonia and Ventilator Associated Bacterial Pneumonia: Developing Drugs for Treatment, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/hospital-acquired-bacterial-pneumonia-and-ventilator-associated-bacterial-pneumonia-developing-drugs (2020, accessed 10 July 2020).

Talbot GH, Das A, Cush S, et al. Evidence-based study design for hospital-acquired bacterial pneumonia and ventilator-associated bacterial pneumonia. *J Infect Dis* 2019; 219(10): 1536-1544.

Kaul S, Diamond GA and Weintraub WS. Trials and Tribulations of Non-Inferiority: The Ximelagatran Experience. *J Am Coll Cardiol* 2005; 46(11), 1986-1995.

Schulz-Schüpke S, Byrne RA, Ten Berg JM, et al. ISAR-SAFE: a randomized, double-blind, placebo-controlled trial of 6 vs. 12 months of clopidogrel therapy after drug-eluting stenting. *Eur Heart J* 2015; 36(20): 1252-1263.

Li HK, Rombach I, Zambellas R, et al. Oral versus Intravenous Antibiotics for Bone and Joint Infection. *N Engl J Med* 2019; 380(5): 425-436.

Zhang M, Tsiatis AA and Davidian M. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* 2008; 64(3): 707-715.

Moore KL, Van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Stat Med* 2009; 28(1): 39-64.

Kahan BC, Jairath V, Doré CJ, et al. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials* 2014; 15(1): 139.

Mohamed K, Embleton A and Cuffe RL. Adjusting for covariates in non-inferiority studies with margins defined as risk differences. *Pharm Stat* 2011; 10(5): 461-466.

Zhang Z, Tang L, Liu C, et al. Conditional estimation and inference to address observed covariate imbalance in randomized clinical trials. *Clin Trials* 2019; 16(2): 122-131.

Acuna SA, Dossa F and Baxter NN. Meta-analysis of Non-Inferiority and Equivalence Trials: Ignoring trial design leads to differing and possibly misleading conclusions. *J Clin Epidemio* 2020; 127: 134-141.

Widmer M, Piaggio G, Nguyen TMH, et al. Heat-stable carbetocin versus oxytocin to prevent hemorrhage after vaginal birth. *N Engl J Med* 2018; 379(8): 743-752.

Bensdorp AJ, Tjon-Kon-Fat RI, Bossuyt PMM, et al. Prevention of multiple pregnancies in couples with unexplained or mild male subfertility: randomised controlled trial of in vitro fertilisation with single embryo transfer or in vitro fertilisation in modified natural cycle compared with intrauterine insemination with controlled ovarian hyperstimulation. *BMJ* 2015; 350.

Van Geloven N. Non-inferiority or superiority? Letter to the editor/rapid response to BMJ, https://www.bmj.com/content/350/bmj.g7771/rr (2015, accessed 1 June 2020).

Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat Med* 1990; 9(12): 1447-1454.

Quartagno M, Walker AS, Babiker AG, et al. Handling an uncertain control arm event risk in non-inferiority trials: non-inferiority frontiers and the power-stabilising transformation. *Trials* 2020; 21(1): 1-12.

Appendix

In this part, we show supplementary results.

Appendix 1: Comparison of sample size when switching using the anticipated control proportion with smaller margins

In Figure 3, Figure 4 and Figure 5 of online supplemental materials, with one sided $\alpha = 0.05$ and power $= 0.80$, we present the results of comparison of sample size when switching using the anticipated control proportion, given $\delta_{RD} = -0.10$, $\delta_{RD} = -0.05$ and $\delta_{RD} = -0.01$, respectively.

Appendix 2: Illustrations of simulated rejection regions when switching using the anticipated control proportion and $p_t^* = p_c^*$

As shown in Figure 6 of online supplemental materials, we demonstrate the simulated rejection regions for power when switching using the anticipated control proportion (with 100000 simulations, $n = 190$, $p_c^* = p_t^* = 0.40$ and $\delta_{RD} = -0.125$). Specifically, when the trial is designed on the RD scale, regardless of the scales used to the report the trial results, all trial outcomes summarized by $(\dot{p}_t, \dot{p}_c)$ located in the area composed of green solid circles conclude that the treatment is non-inferior to the control (rejection of the null hypothesis of non-inferiority). For trials with outcomes $(\dot{p}_t, \dot{p}_c)$ located in the area composed of black circles analyses on all scales cannot conclude that the treatment is non-inferior to the control (no rejection of the null hypothesis). However, for trials with outcomes $(\dot{p}_t, \dot{p}_c)$ located in the area composed of blue solid circles, one can conclude that the treatment is non-inferior to the control when using the RR or OR scale to analyze the trial results, while one cannot draw this conclusion if the RD scale is used. Similarly, for trials with $(\dot{p}_t, \dot{p}_c)$ located in the area composed of red solid circles, one can conclude that the treatment is non-inferior to the control when using the RR scale to analyze the trial results, while one cannot draw this conclusion if the RD or OR scale is used. Therefore, this figure intuitively explains why power changes when different scales are used to analyze the trial results based on anticipated rate.

Considering that a type I error is the rejection of a true null hypothesis, the same method was used for Figure 7 of online supplemental materials, which intuitively explains why type I error rate changes when different scales are used to analyze the trial results and switching using the anticipated control proportion.

Appendix 3: Illustrations of simulated rejection regions when switching using the observed control proportion

As shown in Figure 8 of online supplemental materials, we demonstrate the simulated rejection regions for power when switching using the observed control proportion (with 100000 simulations, $n = 190$, $p_c^* = p_t^* = 0.40$ and $\delta_{RD} = -0.125$). Regarding the type I error, the same analysis method was used for Figure 9 of online supplemental materials. Conclusions similar to Appendix 2 can be drawn.

Appendix 4: Illustrations of simulated rejection regions when using anticipated rate and $p_t^* \neq p_c^*$

As shown in Figure 10 and Figure 11 of online supplemental materials, we demonstrate the simulated rejection regions for power when switching using the anticipated control proportion and $p_t^* \neq p_c^*$ (with 100000 simulations, $n = 190$, $p_c^* = 0.40$, $\delta_{RD} = -0.125$, $p_t^* = p_c^* + 0.05$ or $p_t^* = p_c^* - 0.05$, respectively). Conclusions similar to Appendix 2 can also be drawn.

Appendix 5: Supplementary Tables

In the sample size formulas presented in Supplementary Table 2, it is assumed that for each analysis scale an independent non-inferiority margin was chosen. When considering different analysis scales at the design stage of a study, it is not uncommon that researchers will first choose the ’boundary success rate’ that is still allowed in the treatment arm and then translate it to non-inferiority margins on different scales like was illustrated in the INES case study.

Sample size calculations use an anticipated success proportion in the control arm p_c^*. Now suppose that the margin of non-inferiority is initially defined on the risk difference scale, say it is δ_{RD}. Note that δ_{RD} is negative when using success proportion. Under the null hypothesis of inferiority, this results in the following boundary success proportion in the treatment arm $p_t^{inf,1}$ $=p_c^* + \delta_{RD}$. Based on the two proportions p_c^* and $p_t^{inf,1}$, margins on the other analyses scales may be chosen alternatively. Table 3 lists the margins when mapping non-inferiority margins in this way.

Table 2. Non-inferiority hypotheses and sample size formulas based on Z-tests for four different analysis scales. n represents the sample size required per arm. δ_{RD}, δ_{RR}, δ_{OR} and δ_{RR}^f represent the margin given on the RD,RR,OR scales with success rate and on the RR scale with failure rate, respectively. α denotes the type I error rate, β indicates the type II error rate and $z_{1-\alpha}$ ($z_{1-\beta}$) is the lower α-th (β-th) quantile of the standard normal distribution.

RD	RR (success rate)	OR	RR (failure rate)
$H_0 : p_t - p_c \leq \delta_{RD}$	$H_0 : p_t/p_c \leq \delta_{RR}$	$H_0 : (\frac{p_t}{1-p_t})/(\frac{p_c}{1-p_c}) \leq \delta_{OR}$	$H_0 : (1-p_t)/(1-p_c) \geq \delta_{RR}^f$
$H_a : p_t - p_c > \delta_{RD}$	$H_a : p_t/p_c > \delta_{RR}$	$H_a : (\frac{p_t}{1-p_t})/(\frac{p_c}{1-p_c}) > \delta_{OR}$	$H_a : (1-p_t)/(1-p_c) < \delta_{RR}^f$
$n_{RD} = \frac{2(z_{1-\alpha}+z_{1-\beta})^2 p_c^*(1-p_c^*)}{\delta_{RD}^2}$	$n_{RR} = \frac{2(z_{1-\alpha}+z_{1-\beta})^2 \frac{(1-p_c^*)}{p_c^*}}{(\ln(\delta_{RR}))^2}$	$n_{OR} = \frac{2(z_{1-\alpha}+z_{1-\beta})^2}{(\ln(\delta_{OR}))^2 p_c^*(1-p_c^*)}$	$n_{RR}^f = \frac{2(z_{1-\alpha}+z_{1-\beta})^2 \frac{p_c^*}{(1-p_c^*)}}{(\ln(\delta_{RR}^f))^2}$

Table 3. Given margin on the RD scale (δ_{RD} is negative here), mapping the NI margin to the RR, OR scales using anticipated control rate or observed control rate with success proportions (p_c^* and $\hat{p}_c$) or failure proportions ($(1-p_c^*)$ and $(1-\hat{p}_c)$), respectively. δ_{RR}, δ_{OR} and δ_{RR}^f represent the margin given on the RR,OR scales with success rate and on the RR scale with failure rate, respectively.

Target scale	Using anticipated control rate	Using observed control rate
RR	$\delta_{RR} = (p_c^* + \delta_{RD})/p_c^*$	$\delta_{RR} = (\hat{p}_c + \delta_{RD})/\hat{p}_c$
OR	$\delta_{OR} = 1 + \frac{\delta_{RD}}{p_c^*(1-p_c^*-\delta_{RD})}$	$\delta_{OR} = 1 + \frac{\delta_{RD}}{\hat{p}_c(1-\hat{p}_c-\delta_{RD})}$
RR^f	$\delta_{RR}^f = (1 - p_c^* - \delta_{RD})/(1-p_c^*)$	$\delta_{RR}^f = (1 - \hat{p}_c - \delta_{RD})/(1-\hat{p}_c)$