Inclusion of non-inferiority analysis in superiority-based clinical trials with single-arm, two-stage Simon's design

Miguel Sampayo-Cordero (sampayo.mc@gmail.com)
MedSIR (Medica Scientia Innovative Research) https://orcid.org/0000-0003-1469-3410
Bernat Miguel-Huguet
Hospital Universitari de Bellvitge
José Pérez-García
IOB, Institute of Oncology, Quiron Salud Group
David Páez
Hospital de la Santa Creu i Sant Pau Institut de Recerca
Ángel Guerrero-Zotano
Vanderbilt-Ingram Cancer Center
Javier Garde Noguera
Hospital Universitari Arnau de Vilanova
Elena Aguirre
Hospital Quironsalud Zaragoza
Esther Holgado
Hospital Universitario Ramon y Cajal Servicio de Oncologia Medica
Elena López-Miranda
Hospital Universitario Ramon y Cajal Servicio de Oncologia Medica
Xin Huang
Pfizer Global Research and Development
Antonio Llombart-Cussac
FISABIO Hospital Arnau de Vilanova
Andrea Malfettone
MedSIR (Medica Scientia Innovative Research)
Javier Cortés
Vall d'Hebron Institute of Oncology (VHIO)

Research article

Keywords: Clinical trial; Non-inferiority; Switching to non-inferiority; Two-stage; Single-arm; Phase II; Early stopping; Group sequential designs; Adaptive designs; Non-comparative

Posted Date: August 31st, 2019

DOI: https://doi.org/10.21203/rs.2.13821/v1

Version of Record: A version of this preprint was published at Contemporary Clinical Trials Communications on December 1st, 2020. See the published version at https://doi.org/10.1016/j.conctc.2020.100678.

Abstract

Background The non-inferiority (NI) hypothesis is not usually considered in the early phases of clinical development. A proof-of-concept phase II study that allows for the analysis of NI, if superiority criteria cannot be met, may balance between multiple endpoints and additional parameters, which will result in more informed decisions in regard to the therapeutics' potential value. Methods A total of 12,768 two-stage Simon's design trials were constructed based on different assumptions of rejection response probability, minimum desired response probability, type I and II errors, and NI margins. P-value and type II error were calculated with stochastic ordering using Uniformly Minimum Variance Unbiased Estimator. Type I and II errors were simulated using the Monte Carlo method. Agreement between calculated and simulated values was analyzed with Bland-Altman plots. Results The level of agreement was equivalent between calculated and simulated values in two-stage superiority designs and ones in which switching to NI was allowed. Conclusion Switching to NI analysis, when superiority criteria are not achieved, may be useful for weighing additional factors such as clinical benefit duration, safety, cost, or biomarker strategy while assessing activity and early efficacy rate. Implementation of this strategy can be achieved through simple adaptations to existing designs for one-arm phase II clinical trials.

Background

A single-arm phase II trial is the proof-of-concept stage in drug development, and focuses on the evaluation of new therapeutic hypothesis[1] and strategies[2] in a clinical setting. Phase II studies in oncology are often multistage trials. Two-stage designs are becoming increasingly more common, allowing for early trial termination in cases with low response rates (RRs) towards avoiding wasting time resources on ineffective treatments.[3] These trials aim to determine whether the new regimen is superior to a pre-specified RR (often 5%)[4] or experience with the standard of care, whereas the alternative hypothesis is that RR is somewhat higher, say 20%.[4]

Nevertheless, the non-inferiority (NI) question might also be relevant in the phase II setting.[5] A typical scenario is one in which an experimental treatment is potentially less toxic, less costly, easier to administer or with longer duration of benefit than a conventional treatment, but these do not represent a reduction of efficacy or in percentage of patient with clinical benefit. [6]

In accordance with the European Medicines Agency guidelines,[7] in any superiority trial where NI may be an acceptable outcome, it is prudent to specify a NI margin in the protocol to avoid serious difficulties that can arise from later designation. Specification of a margin after viewing the results can produce an increase in the alpha error rate.[8] In 2016, the Food and Drug Administration published guidelines to establish effectiveness in NI trials. The statistical issues associated with NI studies, and procedures used to determine the NIM, have been extensively described.[6]

Among all available multistage designs, the most popular is a two-stage design with a futility stopping point based on Simon's minimax or optimal criterion.[9] The simplicity of Simon's design may account for its popularity. However, the inference procedures used in two-stage designs are often not corrected to account for these designs' adaptive nature.[10] For point estimation, previous authors have developed a method to calculate the Uniformly Minimum Variance Unbiased Estimator (UMVUE) for Simon's designs and to achieve optimal results.[11] What is more, p-values and type II errors can be calculated with stochastic ordering of the UMVUE.[9,12] These methods can be used when the realized sample size at the stopping stage is different from that specified in the initial design, and this property makes them very useful for designing and analyzing two-stage phase II trials.[9]

The aim of the present study is to assess the validity of the UMVUE-based calculation method is planning two-stage Simon's design phase II trials, where switching to NI is allowed if the superiority criteria cannot be met.

Methods

NI analysis

In a NI analysis, the goal of the study is to show that the effect of the test drug (p) is not inferior to the effect of the active control (p_0) by a specified amount, called NIM. The null and alternative hypotheses should be defined as follows:[6]

H_0: $p_0 - p \geq$ NIM (p is inferior to the control (p_0) by NIM or more); (1)

H_a: $p_0 - p <$ NIM (p is inferior to the control (p_0) by less than NIM). (2)

A challenging point in NI analysis is to distinguish an effective treatment from a less effective or ineffective treatment. The presence of assay sensitivity in a NI trial should be stated from: i) historical evidence of sensitivity to drug effects based on well controlled trials and a robust statistical and clinical judgment (e.g. A treatment cannot be used as a control arm if the superiority against placebo was inconclusive in historical studies); and ii) appropriate conduct of the trial that adheres closely to the design of the trials used to determine that historical evidence of sensitivity to drug effects exists.[13] The margin chosen for a NI trial should be defined prior to study initiation, taking into these historical evidences. Although the NI margin used in a trial can be no larger than the entire assumed effect of the active control against placebo (M1), it is generally desirable to choose a lower margin (M2) that reflects the largest loss of effect that would be clinically acceptable.[6] Showing NI to M1 provides assurance that the test drug had an effect greater than zero, but, in many cases, that is not sufficient to conclude that the test drug had a clinically acceptable effect.[6] In a fixed margin approach, the NIM could be considered as the risk ratio or risk difference, reflective of the average effect of the active control over placebo in previous studies [($p_{control}$ / $p_{placebo}$)> 1 or ($p_{control}$ - $p_{placebo}$)> 0], for example:

Relative risk = 2.64, 95% confidence interval (CI): (*1.72* to 3.56). (3)

Risk difference = 0.15, 95% CI: (*0.07* to 0.22). (4)

We selected the 95% CI lower bound (1.72 or 0.07) and adjusted to retain at least 50% of the historical effect of active control versus placebo arm ([1.72^(1−0.5) = 1.31] or [0.07*(1−0.5) = 0.035]). [6] Accordingly, the calculated NIM describes a ratio or a difference reflecting the largest loss of effect in control group RR (p_0) that would be clinically acceptable. Therefore, the null and alternative hypothesis of NI analysis can be defined as follows and depending on p_0/NIM:

H_0: $p \leq (p_0 / NIM_{as\ ratio})$ or H_0: $p \leq (p_0 - NIM_{as\ difference})$; (5)

$$H_1: p > (p_0 / NIM_{as\ ratio}) \text{ or } H_0: p > (p_0 - NIM_{as\ difference}); (6)$$

Risk ratio is preferred because it is less affected than risk differences by variability in the event rates in the placebo group.[6]

Switching between superiority to NI analysis in a single-arm design

In a superiority analysis design with tumor response as the primary endpoint, switching to NI analysis does not inflate the type I error rate when NI analysis and NIM are properly pre-specified.[7] Therefore, we can switch to NI analyses if the superiority criteria cannot be met, because the final number of responders is lower than the prespecified superiority boundary (a). We assumed the same number of patients as in superiority analysis (n); and a_{ni} (number of responding patients in NI analysis) is chosen as the lowest integer satisfying the type I error rate in NI analysis (α_{ni}) ≤ α:

$$B(a_{ni}|n, p_0 / NIM) \geq 1- \alpha. (7)$$

The power should be calculated:

$$1 - \beta_{ni} = B(a_{ni} -1|n, p_0/NIM); \text{ where } a_{ni} \leq a; 1 - \beta \leq 1 - \beta_{ni}. (8)$$

Accordingly, the study will achieve a positive finding when "p" is equal or higher than "p_0 / NIM" and significance level evaluated by binomial test in NI analysis is ≤ α. As the NI analysis has the same expected accrual and lower or equal number of responders needed to declare significance than superiority analysis ($a_{ni} \leq a$), power always will be equal or greater in NI than superiority criteria. Thus, this design can assess superiority and NI criteria with the same sample size, type I and type II error rates used in the superiority strategy (as outlined in the Supplementary Methods).[7]

Additionally, in a single-arm two-stage Simon's design, UMVUE-based calculation of p-value is still valid when the realized sample size and number of responders to achieve a positive finding are different at the stopping stage from that specified in the design (as outlined in the Supplementary Methods).[9] Thus, switching between superiority to NI strategy may be implemented in Phase II Simon's designs.

A numerical example has been proposed in Results section.

Implementation

A total of 12,768 two-stage, single-arm designs were computed based on different assumptions of p_0, p_1, α_1, 1 - β. The package "Clinfun" (function "ph2simon")[14] from R software[15] was used for computing these designs.

The NIMs selected to formulate the rejection proportion ranged between 1 to 1.45 in 0.05 increments.[8] P-values and type II errors in every design were calculated with stochastic ordering of UMVUE.[9] A user-defined function in the R software was used for calculating these P-values and type II errors. Alpha and beta errors were simulated with the Monte Carlo method. Number of random samples generated were based on the need to attain 95% confidence, so

that simulated values of alpha and beta errors were within 0.5% and 1% of the true values, respectively[16]. Agreement between calculated and simulated values were analyzed with Bland-Altman plots (as outlined in the Supplementary Methods) [17,18]

Results

The results showed a proportional bias between calculated p-values and simulated alpha levels. Higher levels of type I error related to greater absolute differences between the calculated p-values and simulated alpha errors. This is not surprising considering that high statistical error is likely reflective of small sample size and high imprecision. Moreover, we observed that the lower boundary of the 95% CI was crossed by more than 2.5% of the differences. This finding suggests that calculated p-values tended to be slightly lower than the alpha values (Figure 1).

However, it is important to consider that these two biases are common in designs where superiority was not switched (NIM = 1) and designs with switching to NI (NIM > 1). Additionally, the percentage of values crossing the 95% CI boundary is equivalent in both designs (Figure 1). Comparison of type II errors also reflected that simulated type II errors tended to be slightly lower than the calculated type II errors. However, superiority design and switching to NI design strategy displayed equivalent results with about 95% limits of agreement (Figure 2).

In the superiority scenarios (NIM = 1), we observed that the cloud of points was grouped in values of 0.01, 0.05, and 0.1 for the p-value, and 0.1 and 0.2 for the type II error. This is consistent with the pre-specified design constrains. However, we did not observe this behavior in scenarios where NIM > 1, because the designs' sample sizes were pre-specified for superiority analysis. So, type I and II errors that fulfilled design constraints for different values of NIM showed higher variability. Moreover, there were seven times more scenarios in NIM > 1 analysis than in superiority analysis (NIM = 1).

The maximum differences between calculated p-values and type I errors were 0.0015, 0.002, and 0.004 for 0.01, 0.05 and 0.1 type I errors, respectively (Figure 1). Therefore, in a study with a 0.01 significance level, the maximum simulated type I error ranged from 0.0085 to 0.0115. Studies with a 0.05 significance level had maximum simulated type I errors ranging from 0.048 to 0.052. Finally, studies with 0.1 significance levels had maximum simulated type I errors that ranged from 0.096 to 0.104 (Figure 1).

Regarding beta errors, the maximum bias in designs with 0.1 and 0.2 type II errors was 0.005, respectively. Therefore, calculated type II errors ranged from 0.195 to 0.205 and from 0.095 to 0.105 for designs with 80% and 90% of simulated power, respectively (Figure 2). These results suggest that differences between calculated and simulated scores are not relevant.

Collectively, our findings implied that the minimum and maximum differences for calculated and simulated values were equivalent in superiority (NIM = 1) and NI (NIM > 1) scenarios (Figure 1 and Figure 2).

Figure 1. Agreement between calculated p-value and simulated alpha errors in Simon's two-stage clinical designs.

Legend: Absolute differences have been plotted against average of calculated and simulated scores. The type I errors values considered were 0.1, 0.05 and 0.01, and type II errors values were 0.2 and 0.1. The NIMs selected to formulate the rejection proportion ($p_{0ni} = p_0 / NIM$) were (1, 1.15, 1.2, 1.25, 1.30, 1.35, 1.4, and 1.45). A maximum of 2.5% deviation defined the 95% limits of agreement.

Figure 2. Agreement between calculated and simulated type II error in Simon's two-stage clinical designs.

Legend: Absolute differences have been plotted against average of calculated and simulated scores. The type I values considered were 0.1, 0.05 and 0.01, and type II errors were 0.2 and 0.1. The NIMs selected to formulate the rejection proportion (p0ni = p0/NIM) were (1, 1.15, 1.2, 1.25, 1.30, 1.35, 1.4 and 1.45). A maximum of 2.5% deviation defined the 95% limits of agreement.

Numerical Example

We supposed a 65% historical RR and 6 months of median duration of response (DoR) with standard of care in a specific cancer population. Promised results were achieved with a novel, targeted therapy in preclinical models. Dose studies suggest that the new agent will be compatible with standard of care without safety issues. In addition, an overall RR around 80% was proposed in accordance with activity results in dose finding studies or investigators expect. We proposed an optimal two-stage Simon's design with the usual error constraints accepted for single-arm phase II trials ($\alpha = 0.1$ and $1 - \beta = 0.9$). For the design parameters:

$$(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.65, 0.8, 0.1, 0.9) \quad (9)$$

the optimal design is given by:

$$(a_1 / n_1, a / n) = (21 / 30, 46 / 63) \quad (10)$$

where a1 and n1 are the number of responders needed to move second stage and number of accrued patients at first stage, respectively. Calculations were implemented in the R "Clinfun" library (function "ph2simon").[14] In addition we also suppose a 9 months median DoR based on previous dose finding studies. We consider that achieving a response rate higher than conventional therapy (RR>65%) or a RR equivalent than conventional therapy (65%) with an increase in median DoR will justify the investigation of this strategy in further clinical trials. Additionally, achieving a non-inferior efficacy with a good safety profile could justify the combination of this strategy with standard-of-care; [19] or selection a subgroup of patients that could avoid the toxic effects of chemotherapy.[20] Therefore, the NI question is relevant.

In accordance, we propose to allow efficacy boundary switching to NI if superiority criteria are not achieved. Assuming a NIM of 1.1, the study is given as

$$(a1 / n1, a / n) = (21 / 30, 41 / 63) \quad (11)$$

under

$$(p0 / NIM, p1, \alpha \text{ and } 1 - \beta) = (0.65 / 1.1, 0.8, 0.1, \text{and } 0.9). \quad (12)$$

Therefore, at the study end, more than forty (RR≥65%) responding patients (lower than 46 -RR≥73%-) out of 63 patients will be required to achieve a positive finding.

Add more, if statistical significance is reached in RR analysis, we will to assess DoR based on a time-to-event analysis. [21] Despite multiple analyses, we assessed all approaches with a step-down methodology, where the statistical significance of the second analysis is dependent on the first analysis meeting the requirements for

significance.[22] Therefore, this design conserves the type I error. Although, alternative procedures that control type I error have been proposed.[23,24] Furthermore, since a study can have exploratory purposes, no formal adjustment for multiple comparisons may be required. Usually, a negative finding in the activity step alone will not end biomarker analysis or suspend the investigation of the therapeutic.[22,23] Although that decision involves an inflation of type I error. [24]

On the other hand, if statistical significance is reached in efficacy analysis, we will to assess and enrich our biomarker strategy in different ways.[25] We can evaluate clinical activity based on subgroups of patients classified according markers that has proved clinical validity.[26] We can explore biomarker activity according to previously published and applied statistical strategies to detect a significant effect in pharmacodynamic activity.[26,27] These strategies may be based in clinical proved biomarkers or proposing new ones to be validated in a prospective randomized superiority trial.[21]

Discussion

The use of biomarker information in clinical trials has great potential for efficiently identifying patients most helped by specific treatments.[26,28] These biomarkers must have proven their clinical validity in prospective randomized trial with a superiority design in the enriched population.[29] However, single-arm phase II trials are the first screenings for efficacy of new therapeutic agents in humans. These trials are important milestones towards testing and adapting the biomarker strategies validated in preclinical stages [27]. Although the NI question can be relevant in the phase II setting,[5] it is not usually considered on designing single-arm clinical trials in early clinical development.[30,31] Previous forays into precision medicine have shown that correctly identifying the target population and associated predictive biomarkers may be more critical for treatment success than simple demonstration of superior efficacy against an alternative.[22,23] Consequently, designing a proof-of-concept phase II study that permits NI analysis, if the superiority criteria cannot be met, will allow for more informed decisions that consider efficacy and other parameters, such as safety, cost, and biomarker strategy. Moreover, the analysis of NI provides additional information to the steering committee on the magnitude of the observed activity, without increasing target accrual.[7]

A prior systematic review evaluating the characteristics of phase II trials that best predict for phase III outcome, selected 270 single-arm phase II between 1981 to 2012. All these selected studies led to a phase III clinical trial. The meta-analysis showed that 168 single-arm trials were not positive; and 61 (36.3%) achieve a positive phase III result despite not having obtained a positive result in the proof-of-concept study.[22] Additionally, Jardin DL, et al. 2016, published that 10% of FDA anti-cancer drugs approvals, from 01/01/2009 to 06/30/2014 period, have a negative result in prior phase II clinical trials. [32] In some of these examples, despite the low response rate, investigators considered the response as not worse than conventional treatment; and they decided to continue with a Phase III trial based on other parameters as prolonged duration in clinical benefit or safety.[19,33–35] Therefore, investigators decide to switch their primary superiority criteria to a non-inferiority objective and weigh their decision with other relevant parameters. As a result, they achieve a positive Phase III trial, and FDA approves the therapies. [32] This suggest that the NI question could be relevant in the phase II setting and it is not so rare. However, the NI hypothesis was not preplanned, and it was only considered to deal with negative findings in proof-of-concept trials. This strategy leads to an increase in the probability of type I error and the number of false positives. [6] In accordance, both systematic reviews reported and high rate of negative Phase III trials after a negative Phase II superiority trial (between 64% and 85%).[22,32] However, if the decision to switch to non-inferiority had been preplanned and included in the statistical design, investigators would have additional information about magnitude of RR (non-inferior, or superior), clinical

benefit duration or other parameters without a type I error inflation. In addition, the non-inferiority comparison with an historical control has probed its validity identifying subgroups of patients in adjuvant setting who can avoid the toxic effects of chemotherapy.[20]

The most popular design for phase II cancer clinical trials is a single-arm two-stage Simon's study.[11] Numerous extensions have been proposed for Simon's design, including randomized multi-arm trials designed to select the winner among the proposed therapeutics (Pick-the-winner study design).[36] However, the inference procedures used in two-stage designs are often not corrected to account for the adaptive nature of these designs. A maximum likelihood estimator of the RR, the number of positive responses/total number of patients, is biased. CI and p-value should not be computed as if the data were obtained in a single stage due to the possibility of early termination.[10] Different methods have been proposed to obtain a proper inference from Simon's two-stage design.[10] The use of the UMVUE to estimate RR is recommended as it addresses situations when the actual number of patients recruited is equal to or different from preplanned values.[3] The calculation of RR, p-values and CIs has been incorporated in open-access statistical libraries (packages "clinfun" and "OneArmPhaseTwoStudy", R statistical software) based on previously published methods [9,12].

We observed some differences between calculated (p-value) and simulated values (type I error). However, they were not relevant to the most common design constraints used in phase II single-arm trials (0.01 to 0.1 type I errors and 0.1 to 0.2 type II errors). Using the UMVUE-based calculation method, we proved that the same level of agreement between calculated and simulated values (type I and II errors) results from both two-stage Simon's superiority designs and designs in which switching to NI was allowed. Importantly, our findings suggest that the proposed method for analyzing NI, when superiority cannot be met, does not introduce bias.

The major limitations of this method are based on bringing together the inherent complexities of a study with historical controls and a NI analysis. However, this limitations are common to comparative designs of NI, because NIM must be established based on historical controls evidence.[6] Additionally, some bias as selection of inappropriate patients, poor compliance and insufficient follow-up, that can lead to erroneously conclude that a treatment is not inferior to placebo in a comparative study, go against a positive achievement when the comparison is done among a theoretical rate of efficacy deduced from historical controls. As comparative designs of NI, to declare a therapy as non-inferior in a single arm trial, we need to demonstrate assay sensitivity based on an adequate trial design and conduct.[13] In accordance, lower sample sizes in phase II single arm trials are not more challenging for NI analyses than superiority ones if trial is properly preplanned and conducted. Our results suggested that we can conserve the same levels of alpha and beta errors after switching to NI analysis without increase sample size.

Switching to NI analysis with the UMVUE-based calculation method may be extended to two-stage designs with both futility and superiority boundaries (as outlined in the Supplementary Methods).[21] We limited our results to two-stage designs that are most popularly for phase II cancer clinical trials, but the methods discussed in this article could be extended to phase II trials with any number of stages.[9] Likewise, we can design a single-arm time-to-event study with switching to NI analysis based on the exponential maximum likelihood estimator, one-sample log-rank tests or other approximations to the Kaplan-Meyer estimations.[21,37] If we assume the same number of patients as in superiority analysis, we would formulate the NI rejection hazard rate (λ_{0ni}) from the hazard rate assumed under H_0 in the superiority analysis (λ_0) and the NIM estimated from historical studies[6] ($\lambda_{0NI} = \lambda_0 / NIM$).

Conclusions

Switching to NI analysis, when superiority criteria are not achieved, may be useful for weighing additional factors such as clinical benefit duration, safety, cost, or biomarker strategy while assessing activity and early efficacy rate. The results of previous single-arm designs leading to a successful Phase III trial or identifying subgroups of patients who can avoid the toxic effects of chemotherapy suggest that NI question is relevant in non-comparative studies. Implementation of this strategy can be achieved through simple adaptations to Simon's two-stage design and other existing designs for one-arm phase II clinical trials.

List Of Abbreviations

CI: Confidence interval

DoR:Duration of response

EMA:European Medical Agency

FDA: U.S Food and Drug administration

H0:Null hypothesis

H1:Alternative hypothesis

NI: Non-inferiority

NIM: Non-inferiority margin

RR: Response rate

UMVUE: Uniformly Minimum Variance Unbiased Estimator

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this works:

Miguel Sampayo-Cordero reports personal fees from Hospital Vall d'Hebron, Roche, Nestle Health Science, Laboratorios Leti, Medica Scientia Innovation Research (MedSIR), Syntax for Science, Ability Pharma and Scienco Klinico, outside the submitted work.

Bernat Miguel-Huguet declares no conflict of interest.

José Pérez-García has received consulting and advisor fees from: Roche and Eli Lilly.

David Páez declares a scientific advisory role for Amgen, Sanofi, Merck Serono, F. Hoffmann-La Roche Ltd, Lilly and Servier.

Ángel L. Guerrero-Zotano declares no conflict of interest.

Javier Garde Noguera declares no conflict of interest.

Elena Aguirre has received consulting and advisor fees from Pfizer and honorarias from: Roche, Novartis, Celgene, Eisai and Pfizer.

Esther Holgado declares no conflict of interest.

Elena López-Miranda declares no conflict of interest.

Xin Huang is a statistician at Pfizer Oncology.

Antonio Llombart-Cussac has received consulting and advisor fees from Roche, GlaxoSmithKline, Novartis, Celgene, Eisai, and AstraZeneca and has stock options, patents and intellectual property from MedSIR.

Andrea Malfettone declares no conflict of interest.

Javier Cortés declares the next conflict of interest:

Consulting/Advisor: Roche, Celgene, Cellestia, AstraZeneca, Biothera Pharmaceutical, Merus, Seattle Genetics, Daiichi Sankyo, Erytech, Athenex, Polyphor, Lilly, Servier, Merck Sharp&Dohme.

Honoraria: Roche, Novartis, Celgene, Eisai, Pfizer, Samsung, Lilly, Merck Sharp&Dohme.

Research funding to the Institution: Roche, Ariad pharmaceuticals, AstraZeneca, Baxalta GMBH/Servier Affaires, Bayer healthcare, Eisai, F.Hoffman-La Roche, Guardanth health, Merck Sharp&Dohme, Pfizer, Piqur Therapeutics, Puma C, Queen Mary University of London, Seagen.

Stock, patents and intellectual property: MedSIR

Funding

Not applicable

Authors' contributions

MSC, JC, AM and ALC conceived the idea. MSC, BMH, JC, JP, DP, AGZ, JGN, EA, EH, ELM, XH, ALC, AM and JC developed theoretical rational, clinical justification and discussed and interpreted the study results. MSC, BMH and XH developed, reviewed and reported statistical procedures. MSC, JC, ALC and AM coordinated the tasks and contributions of all the authors to the study. MSC, AM, JC and ALC drafted the manuscript. All authors critically reviewed and made important intellectual contributions to this manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank American Manuscript Editors for medical writing.

References

1. Quintela-Fandino M, Lluch A, Manso L, et al. 18F-fluoromisonidazole PET and Activity of Neoadjuvant Nintedanib in Early HER2-Negative Breast Cancer: A Window-of-Opportunity Randomized Trial. Clin Cancer Res 2017; 23: 1432–1441.
2. Llombart-Cussac A, Cortés J, Paré L, et al. HER2-enriched subtype as a predictor of pathological complete response following trastuzumab and lapatinib without chemotherapy in early-stage HER2-positive breast cancer (PAMELA): an open-label, single-group, multicentre, phase 2 trial. Lancet Oncol 2017; 18: 545–554.
3. Porcher R, Desseaux K. What inference for two-stage phase II trials? BMC Med Res Methodol 2012; 12: 117.
4. Rubinstein L, Leblanc M, Smith MA. More randomization in phase II trials: necessary but not sufficient. J Natl Cancer Inst 2011; 103: 1075–1077.
5. Neuenschwander B, Rouyrre N, Hollaender N, et al. A proof of concept phase II non-inferiority criterion. Stat Med 2011; 30: 1618–1627.
6. FDA. FDA GUIDANCE: Non-Inferiority Clinical Trials to Establish Effectiveness, https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf (2016).
7. Committee For Proprietary medicinal products (CPMP), The European Agency for the Evaluation of Medicinal Products. Points to consider on switching between superiority and non-inferiority. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003658.pdf.
8. Tanaka S, Kinjo Y, Kataoka Y, et al. Statistical issues and recommendations for noninferiority trials in oncology: a systematic review. Clin Cancer Res 2012; 18: 1837–1847.
9. Jung S-H. Statistical issues for design and analysis of single-arm multi-stage phase II cancer clinical trials. Contemp Clin Trials 2015; 42: 9–17.
10. Koyama T, Chen H. Proper inference from Simon's two-stage designs. Stat Med 2008; 27: 3145–3154.

11. Zhao J, Yu M, Feng X-P. Statistical inference for extended or shortened phase II studies based on Simon's two-stage designs. BMC Med Res Methodol 2015; 15: 48.
12. Jung S-H, Owzar K, George SL, et al. P-value calculation for multistage phase II cancer clinical trials. J Biopharm Stat 2006; 16: 765–775; discussion 777–783.
13. Choice of Control Group and Related Issues in Clinical Trials : ICH E10, http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/choice-of-control-group-and-related-issues-in-clinical-trials.html (accessed 15 October 2018).
14. Venkatraman E. Seshan. Package 'clinfun' - Clinical Trial Design and Data Analysis Functions. Version 1.0.14. https://cran.r-project.org/web/packages/clinfun/clinfun.pdf.
15. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, https://www.R-project.org/. 2016. (2016).
16. William Oberle. Weapons and Materials Research Directorate, ARL. Monte Carlo Simulations: Number of Iterations and Accuracy., Available from: file:///Users/becario1/Downloads/ADA621501%20(1).pdf (2015).
17. Potte H. Critical Review of Method Comparison Studies for the Evaluation of Estimating Glomerular Filtration Rate Equations| Nephrology and Kidney Failure Journal|SciForschen, https://www.sciforschenonline.org/journals/nephrology-kidney/IJNKF–1–102.php (accessed 15 October 2018).
18. Ludbrook J. Confidence in Altman-Bland plots: a critical review of the method of differences. Clin Exp Pharmacol Physiol 2010; 37: 143–149.
19. Tang PA, Cohen SJ, Kollmannsberger C, et al. Phase II clinical and pharmacokinetic study of aflibercept in patients with previously treated metastatic colorectal cancer. Clin Cancer Res 2012; 18: 6023–6031.
20. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. New England Journal of Medicine 2016; 375: 717–729.
21. Sin-Ho Jung. Randomized Phase II Cancer Clinical Trials. 2013.
22. Monzon JG, Hay AE, McDonald GT, et al. Correlation of single arm versus randomised phase 2 oncology trial characteristics with phase 3 outcome. Eur J Cancer 2015; 51: 2501–2507.
23. Schwaederle M, Zhao M, Lee JJ, et al. Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. JCO 2015; 33: 3817–3825.
24. U.S. Department of Health and Human Services Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Multiple Endpoints in Clinical Trials - Guidance for Industry. https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm536750.pdf.
25. No Biomarker, No Trial?, https://thetranslationalscientist.com/issues/0216/no-biomarker-no-trial/ (accessed 15 October 2018).
26. MONARCH 2: Abemaciclib in Combination With Fulvestrant in Women With HR+/HER2– Advanced Breast Cancer Who Had Progressed While Receiving Endocrine Therapy: Journal of Clinical Oncology: Vol 35, No 25, http://ascopubs.org/doi/full/10.1200/JCO.2017.73.7585 (accessed 15 October 2018).
27. Romero I, Rubio MJ, Medina M, et al. Preoperative olaparib in early-stage endometrial cancer (EC): A phase 0, window of opportunity trial to evaluate the PARP inhibition effect, targeting cell cycle-related proteins (POLEN study). JCO 2018; 36: 5598–5598.
28. Perez-Garcia JM, Saura C, Muñoz E, et al. Role of progesterone receptor status (PR) as predictive factor of pathologic complete response (pCR) to neoadjuvant chemotherapy (NACT) in breast cancer patients. JCO 2010;

28: 628–628.
29. Baselga J, Cortés J, Im S-A, et al. Biomarker Analyses in CLEOPATRA: A Phase III, Placebo-Controlled Study of Pertuzumab in Human Epidermal Growth Factor Receptor 2–Positive, First-Line Metastatic Breast Cancer. JCO 2014; 32: 3753–3761.
30. Mauri L, D'Agostino RB. Challenges in the Design and Interpretation of Noninferiority Trials. N Engl J Med 2017; 377: 1357–1367.
31. Althunian TA, de Boer A, Klungel OH, et al. Methods of defining the non-inferiority margin in randomized, double-blind controlled trials: a systematic review. Trials; 18. Epub ahead of print December 2017. DOI: 10.1186/s13063–017–1859-x.
32. Jardim DL, Groves ES, Breitfeld PP, et al. Factors associated with failure of oncology drugs in late-stage clinical development: A systematic review. Cancer Treat Rev 2017; 52: 12–21.
33. Vahdat LT, Pruitt B, Fabian CJ, et al. Phase II study of eribulin mesylate, a halichondrin B analog, in patients with metastatic breast cancer previously treated with an anthracycline and a taxane. J Clin Oncol 2009; 27: 2954–2961.
34. Cortes J, Vahdat L, Blum JL, et al. Phase II study of the halichondrin B analog eribulin mesylate in patients with locally advanced or metastatic breast cancer previously treated with an anthracycline, a taxane, and capecitabine. J Clin Oncol 2010; 28: 3922–3928.
35. Cortes J, O'Shaughnessy J, Loesch D, et al. Eribulin monotherapy versus treatment of physician's choice in patients with metastatic breast cancer (EMBRACE): a phase 3 open-label randomised study. Lancet 2011; 377: 914–923.
36. Jung S-H, George SL. Between-Arm Comparisons in Randomized Phase II Trials. J Biopharm Stat 2009; 19: 456–468.
37. Sample Size Calculator | Kengo Nagashima - The Institute of Statistical Mathematics, https://nshi.jp/en/js/ (accessed 18 June 2019).

Figures

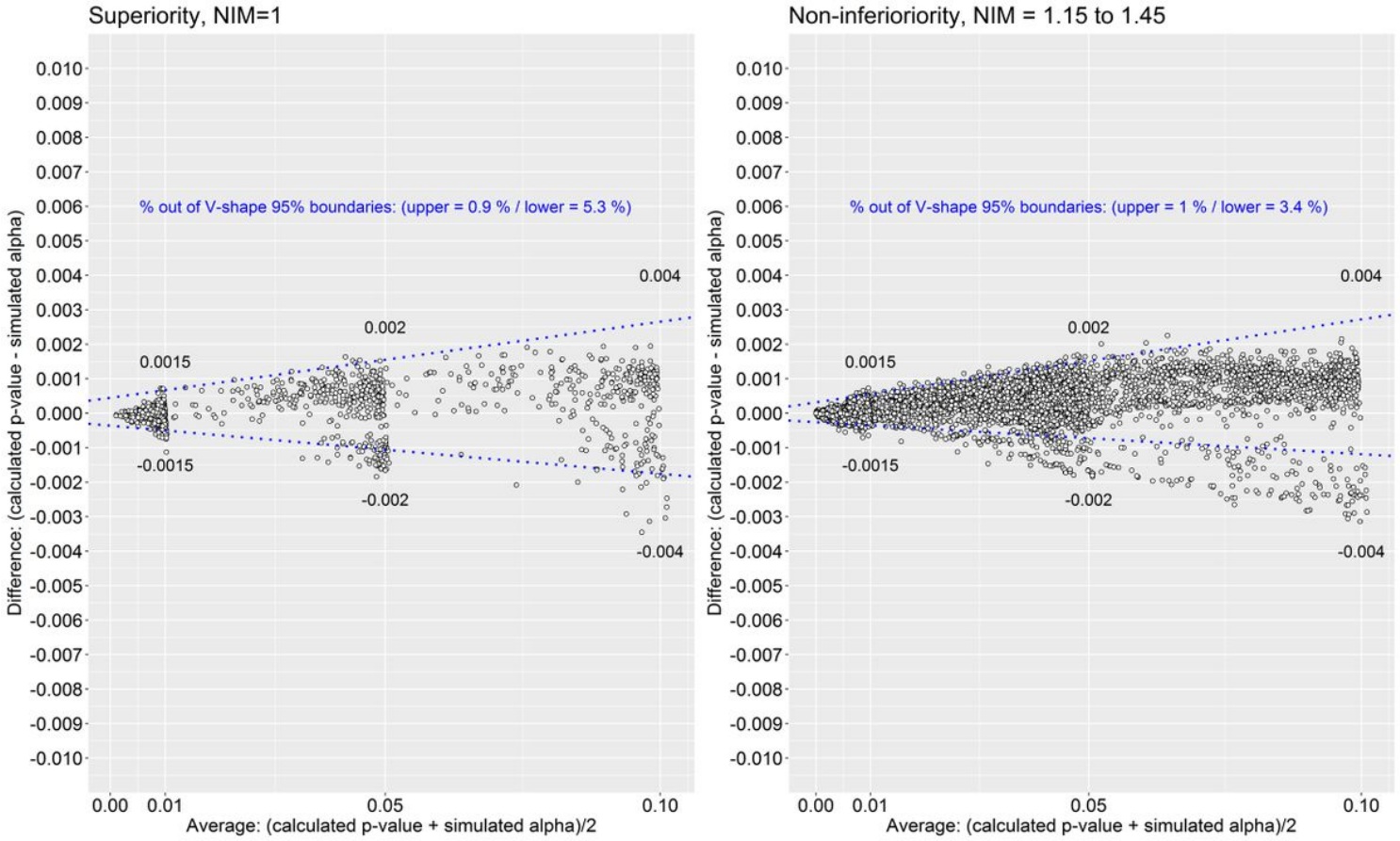

Figure 1

Agreement between calculated p-value and simulated alpha errors in Simon's two-stage clinical designs. Absolute differences have been plotted against average of calculated and simulated scores. The type I errors values considered were 0.1, 0.05 and 0.01, and type II errors values were 0.2 and 0.1. The NIMs selected to formulate the rejection proportion (p0ni = p0 / NIM) were (1, 1.15, 1.2, 1.25, 1.30, 1.35, 1.4, and 1.45). A maximum of 2.5% deviation defined the 95% limits of agreement.

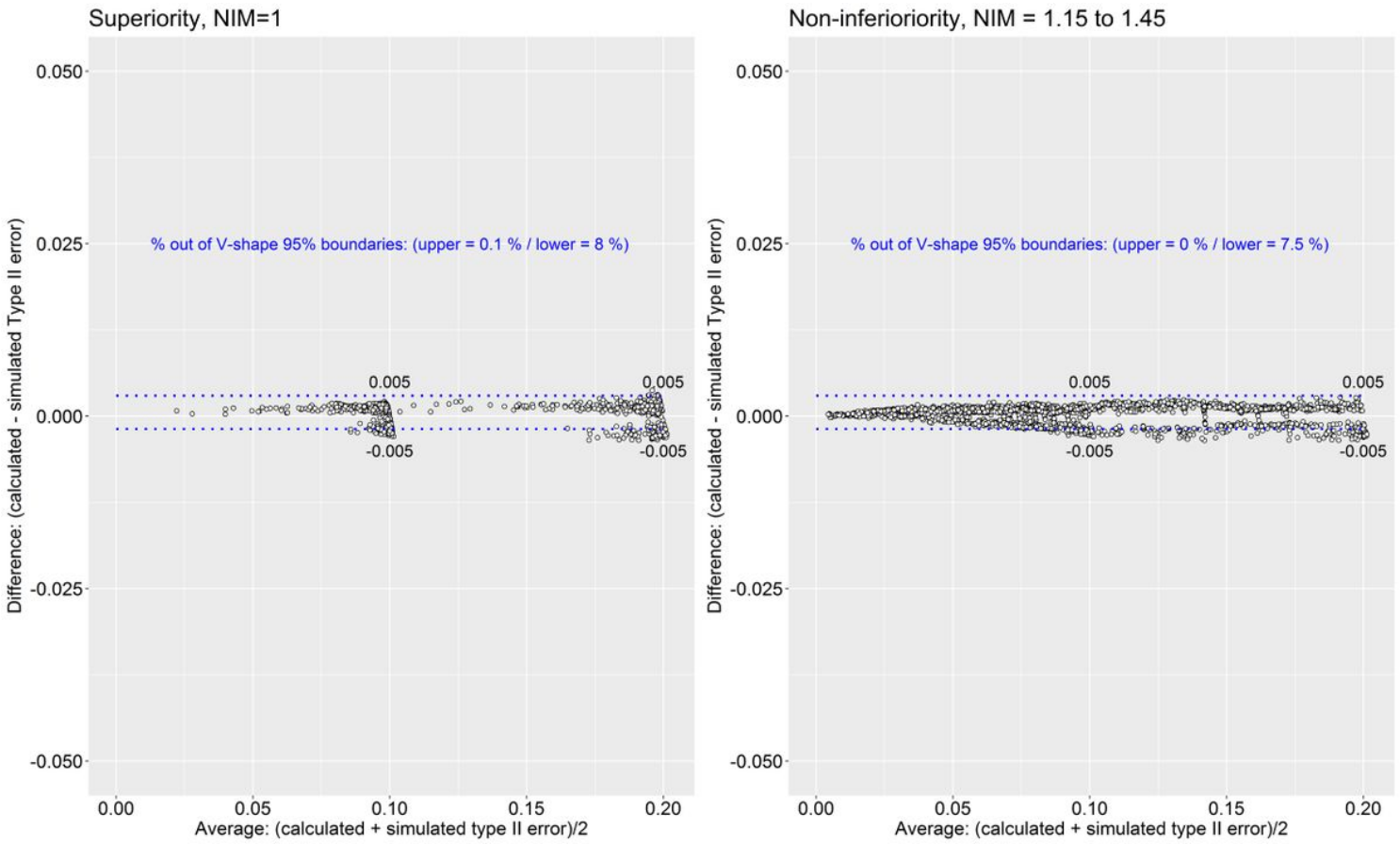

Figure 2

Agreement between calculated and simulated type II error in Simon's two-stage clinical designs. Absolute differences have been plotted against average of calculated and simulated scores. The type I values considered were 0.1, 0.05 and 0.01, and type II errors were 0.2 and 0.1. The NIMs selected to formulate the rejection proportion (p0ni = p0/NIM) were (1, 1.15, 1.2, 1.25, 1.30, 1.35, 1.4 and 1.45). A maximum of 2.5% deviation defined the 95% limits of agreement.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 3SampayoCorderoMSupplementalDataset1.csv
- 3SampayoCorderoMSupplementalDataset0.csv
- 3SampayoCorderoMSupplementalDataset2.csv
- 3SampayoCorderoMetalSupplementalFile2.txt
- 3SampayoCorderoMetalSupplementalFile1.txt
- 3SampayoCorderoMetalBMCMSupplementaryDataSummary.docx
- 2SampayoCorderoMetalBMCMSupplementaryMethods.docx