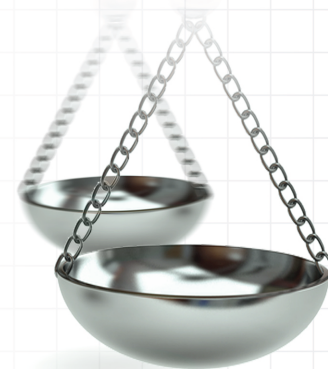




Some issues for the evaluation of noninferiority trials

Xuanqian Xie^{*,1,2}, Myra Wang¹, Vivian Ng¹ & Nancy Sikich¹

¹Health Quality Ontario, Toronto, ON M5S 1N5, Canada

²Technology Assessment Unit of the McGill University Health Centre, Montréal, QC H4A 3J1, Canada

*Author for correspondence: Tel.: +1 416 323 6868 ext. 559; Fax: +1 416 323 9261; shawn.xie@hqontario.ca

Journal of **Comparative Effectiveness Research**

Although published noninferiority trials (NITs) generally conclude that the experimental intervention being studied is noninferior compared with standard therapy or active control, NIT quality is often not satisfactory. We have proposed 14 questions to assist in evaluating the clinical evidence of the experimental versus standard therapy. The aim of these questions is to critically appraise NITs and support proper interpretation of study results. Readers should not only consider whether the confidence interval of the primary effect measure falls within the prespecified noninferiority margin (thus concluding noninferiority), but also assess the similarities between primary and secondary outcomes for the experimental and standard therapy. To conclude noninferiority conceptually is to synthesize evidence from both the current NIT comparing experimental therapy with standard therapy and historical data comparing standard therapy with placebo control. Therefore, readers should use external data sources (e.g., historical data) to validate the study design (e.g., selection of standard therapy, effect measure and the noninferiority margin), and assess the uncertainty of findings due to differences between the observed and expected incidence rates, follow-up time, effects of adjuvant therapy and the secondary outcomes of therapies. Following an explanation of the 14 questions, we then apply the questions to a NIT on intraoperative radiation therapy for early stage breast cancer, as an example.

First draft submitted: 22 April 2018; Accepted for publication: 15 June 2018; Published online: 7 September 2018

Keywords: assay sensitivity • noninferiority • noninferiority margin • noninferiority trial

A superiority trial aims to demonstrate that one treatment is better than another treatment or placebo. Although most clinical trials are superiority studies, a noninferiority trial (NIT) is more suitable in some situations. For example, since standard therapy protocols for some diseases (e.g., cancer and cardiovascular disease) have been well-established and are often supported by randomized controlled trials (RCTs), it is often not ethical to use placebo-controlled superiority trials to demonstrate the efficacy of new (experimental) therapies. Therefore, researchers can conduct a NIT to assess whether the new therapy is at least as effective as the standard therapy (active control). A positive conclusion of noninferiority in a NIT indicates that the confidence interval of the effect measure falls in a prespecified margin (i.e., rejection of the null hypothesis).

Designing a NIT can be quite challenging. Study features such as selecting effect measures and the noninferiority margin, the prevalence of disease, and the effectiveness of adjuvant therapy, may vary across therapies, therefore making them difficult to standardize [1]. Not surprisingly, the quality of NITs is often not satisfactory [2–4]. Two systematic reviews revealed that 83% (145/175 comparisons) and 90% (209/232 comparisons) of published NITs concluded noninferiority for the experimental therapy [3,5]. Although one explanation for this trend is potential publication bias, poor study conduct (e.g., high patient attrition or low patient adherence rate) may introduce biases that lead to an incorrect conclusion of noninferiority. Since the placebo control group is often not included in the NIT (e.g., due to ethical considerations), to determine the similarities between the clinical outcomes (both primary and secondary outcomes) of the experimental and standard therapy, we have to explore the appropriateness of the study design and conduct (e.g., selection of the standard therapy, effect measure and noninferiority margin), and the uncertainty of findings due to differences between the observed and expected incidence rates, follow-up time, effects of adjuvant therapy and the secondary outcomes of treatments.

Although publications exist on guidance for NIT methodology, study design and reporting [2,6,7], we are not aware of many articles that assess the quality or generalizability of NITs. We have proposed 14 questions to assess the noninferiority evidence between the experimental and standard therapy in NITs from the perspective of decision-makers, which have been mainly based on “Noninferiority Clinical Trials to Establish Effectiveness: Guidance for Industry” by the US FDA [8]. Instead of presenting guidance on the study design of NITs, the aim of these questions is to critically appraise NITs and support the proper interpretation of study results. These questions are used to evaluate both the primary and secondary outcomes of NITs to provide an overall understanding of treatment effects and the level of uncertainty of the findings. Although the developed questions were motivated by NITs in oncology, these questions may also be applicable for other disease conditions. While we focused on features of NITs, some questions (e.g., 3, 4, 8, 9, 11 and 14) are also applicable to superiority trials, but the implications of these questions (e.g., 4, 8 and 9) may be different between superiority trials and NITs. We also omitted some shared features found in superiority trials, such as the appropriateness of randomization and blinding. To conclude noninferiority conceptually is to synthesize evidence from both the current NIT comparing experimental therapy with standard therapy and historical data comparing standard therapy with placebo control [8]. Therefore, in principle, readers should explore external data sources to validate the current NIT (addressed in questions 1, 2, 3, 5, 6, 7, 8, 9, 12 and 14). Every question can be scored as ‘Yes’, ‘No’, or ‘Unable to determine’, with ‘No’ as the preferred answer for question 12 and ‘Yes’ as the preferred answer for the other 13 questions. Following an explanation of the 14 questions, we provide a discussion of how to review historical data for NITs. Finally, we demonstrate the use of the questions, using a NIT on intraoperative radiation therapy for early stage breast cancer [9], as an example.

14 questions for the evaluation of NITs

1. Was the rationale for conducting a NIT established?

The rationale for conducting a NIT should be well-established. Generally before conducting a NIT, there is evidence to show that the experimental therapy has a certain advantage over standard therapy [1]; the experimental therapy is better than placebo or best care in clinical studies (from RCTs or observational studies) [8]; the efficacy of the standard therapy has been demonstrated [8] and according to existing evidence, the efficacy of the experimental therapy is similar, if not better, than that of standard therapy. The placebo can be a commonly used monotherapy (i.e., background therapy), and the experimental and standard therapies are other treatments in addition to the monotherapy [10]. These criteria are likely not all explicitly stated in a single NIT [11], so readers may have to seek additional information from other sources to answer this question.

2. Was the standard therapy properly selected?

Sometimes there is more than one standard therapy. In this situation, the NIT should use the standard therapy with the most consistent evidence and ideally with the best efficacy [6]. If a suboptimal treatment is selected as the standard therapy, treatment efficacy can be diminished by repeatedly comparing the suboptimal treatment. If a NIT included multiple standard therapies, the experimental therapy should be compared with each standard therapy [12].

3. Did the study select the appropriate effect measures and end points?

There are a few commonly used clinical end points and effect measures for a disease. Not surprisingly, the selection of clinical end points and effect measures impact the determination of the noninferiority margin and potentially the conclusions of the NIT. Differences in the selection of end points, measures, noninferiority margin and the significance level may contribute to the very large proportion of NITs that show favorable results for the experimental therapy [5,13,14]. Although there is no consensus on these issues, selection of the most commonly used end points and effect measures within the primary analysis of RCTs for the disease of interest is recommended.

4. Was the study well-conducted?

Poorly conducted NITs are likely to bias the results toward the alternative hypothesis and may potentially lead to an incorrect conclusion of noninferiority [8]. For example, imprecise or poorly implemented entry criteria, poor compliance, use of concomitant treatments whose effects may overlap with the treatment(s) of interest, inadequate measurement techniques, errors in delivering assigned treatments, high patient attrition or poor patient follow-up are likely to reduce the differences between the experimental and standard therapy and produce similar results, thus concluding noninferiority [8].

5. Were the expected benefits of the experimental therapy confirmed in the NIT?

Compared with standard therapy, the experimental therapy is usually expected to have some benefit (e.g., improved health-related quality of life, reduced harm, convenient administration or cost savings). However, sometimes the expected benefit is based on lower levels of evidence or intuition and cannot be achieved in reality. For example, based on evidence from a retrospective study, it is expected that a lower dose of capecitabine plus docetaxel (XT) would have a significantly lower risk of adverse events, compared with standard dose XT [15]. However, the results of a subsequent randomized NIT showed similar serious adverse event rates and overall rates for lower dose XT and standard XT. Although the aim of the NIT was not to examine these benefits, the expected benefit of the experimental therapy should be reflected at least partially in a rigorously designed NIT. If not, the value of the experimental therapy should be questioned, even if noninferiority was successfully demonstrated based on the primary outcomes.

6. Was the magnitude of the noninferiority margin justifiable?

Both the European Medicines Agency and the FDA have issued guidelines for selecting the noninferiority margin [8,16]. Authors should report the method used to determine the noninferiority margin in the NIT publication, however this is often not done. A systematic review by Tanaka *et al.* [2] found that 63% of oncology NITs did not specify the method for selecting the noninferiority margin. The lack of justification for the choice of the noninferiority margin can likely be explained by the difficulty in selecting a margin that is supported by scientific evidence and the influence of the noninferiority margin on sample size [1]. Although a few approaches for determining the noninferiority margin are available, the FDA recommends using the preserved effect method and a fixed margin, which often offers a conservative estimate [1,8,17,18]. First, the margin of the conservative estimate of the entire effect of the standard therapy (M1) is calculated based on historical data (e.g., lower limit of the confidence interval of the effect) [8]. Then, a smaller margin (M2) is selected based on clinical judgment of how much of the standard therapy effect should be preserved. For example, M2 as 50% of M1 is usually selected for NITs for cardiovascular disease [8]. However, from a scientific point of view, the fixed margin approach suggested by the FDA does not have obvious advantages over the synthesis method, which combines data from the historical study and the current NIT to directly assess the noninferiority of the experimental therapy without specifying a fixed noninferiority margin [8,17]. Furthermore, when the experimental therapy has other important advantages over the standard therapy (e.g., lower risk of severe adverse events), it may be justifiable to accept a wider noninferiority margin for efficacy [8,16].

7. Was the assay sensitivity presented in the current NIT?

The standard therapy should present assay sensitivity (i.e., ability to distinguish an effective treatment from placebo) in the NIT [8,16]. The results of the NIT would be very difficult to interpret when the observed clinical response of the standard therapy is very different from the response of a historical population, which may violate the constancy assumption. If a NIT includes three groups (i.e., experimental therapy, standard therapy and placebo), the presence of assay sensitivity can be assessed directly within the trial. In addition, this three-arm NIT can assess the superiority of the experimental therapy versus placebo and the noninferiority of the experimental therapy versus control simultaneously. Furthermore, the performance of the placebo may impact the presence of assay sensitivity, since the placebo effect may not be stable in some diseases (e.g., depression) [12]. Therefore, the adoption of a placebo group is recommended when there are no severe, harmful consequences for patients treated with placebo [16]. However, in practice, most NITs do not have placebo groups, so we need to use external data to assess assay sensitivity. We may compare the similarities in various aspects between the current NIT and historical studies (e.g., patient characteristics and entry criteria, trial design, and clinical practice) [8,16]. In addition, we need to compare the performance of the standard therapy in the current NIT with that reported in historical studies [8]. We have elaborated on the use of external data in the section 'Review of historical studies for the evaluation of NITs'.

8. Was the observed risk in the standard therapy group close to the expected value?

Due to improved disease management (e.g., earlier diagnosis, improvements in treatment, etc.), the observed event rate in the standard therapy group is often less than the expected event rate that is usually estimated from historical data. A simulation study has demonstrated that NIT conclusions are highly sensitive to both effect measurements and the underlying risk in the standard therapy group [14]. In an extreme situation, such as when no event occurs in either study group, the efficacy of the experimental therapy versus the standard therapy should be interpreted as 'inconclusive' rather than 'identical'.

9. Was the follow-up duration adequate?

Authors should understand the risk profile of the disease being investigated, and the high-risk period for the event of interest should ideally be within the follow-up period. For example, in oncology, historical data with long follow-up may demonstrate that the standard therapy prevents cancer recurrence [19]. In contrast, a recent study may demonstrate that the experimental and standard therapies have a similar recurrence rate, simply because the follow-up period was too short to observe a difference between groups.

10. Did authors conduct both intention-to-treat and per-protocol analysis?

Generally, intention-to-treat analysis tends to minimize the differences between the experimental and standard therapies, whereas per-protocol analysis tends to reflect the maximum possible difference between the two groups, assuming random crossover and patient attrition [6]. The per-protocol analysis alone may introduce bias (e.g., different patient characteristics for those who completed the treatment originally allocated) [20,21]. It is easier to show noninferiority if patient attrition is high or patient adherence (for drug therapy) is low, which results in similar effects between groups. Demonstrating noninferiority in both an intention-to-treat and per-protocol analysis provides the necessary assurance of noninferiority [21].

11. Did the survival analysis adhere to the random censoring assumption?

In early stage cancer studies, it is not uncommon to select local recurrence as the primary outcome and to censor other more severe events, such as distant recurrence and mortality. For convenience, authors also often make the assumption of random censoring when analyzing time-to-event data. If the study censors oncological events, readers should compare the censored event rates between the two groups and check for adherence to the random censoring assumption.

12. Did the adjuvant therapy contribute to the clinical outcomes?

In addition to the treatments being investigated, most cancer patients also receive one or more adjuvant therapies during follow-up. The adjuvant therapy would reduce the effect difference between the experimental and standard therapies (justification is in the appendix). Since the effects of various therapies cannot be separated, when the effect of the adjuvant therapy is relatively large, it leads to more uncertainty in estimating the true difference between the two therapies.

13. Is the point estimate between the experimental and standard therapy close to no effect for the primary clinical outcomes?

Readers should assess the following: do the data allow you to make a conclusion of noninferiority?; does the confidence interval fall within the prespecified noninferiority margin?; is the point estimate close to no effect between treatments (e.g., hazard ratio of 1, risk difference of zero)? Clinical relevance should be judged using best estimates. For example, although a NIT did not statistically demonstrate noninferiority (e.g., noninferiority margin of hazard ratio 1.075 in the example study), the results showed that the experimental therapy was similar to the standard therapy with relatively small uncertainty (e.g., hazard ratio 1.03; 95% CI: 0.95–1.11) [22].

14. Other than the primary end point, were other benefits of the standard therapy also obtained by the experimental therapy?

Benefits of the standard therapy often extend beyond the primary outcomes. For example, based on studies with 15- and 20-year follow-ups, radiotherapy not only reduces local recurrence for patients with early stage breast cancer, but also mortality risk [23,24]. On the contrary, the evidence for secondary end points (e.g., survival) is often premature for experimental therapies. Thus, when the standard therapy has other important benefits in the long term, readers must be cautious of the potential uncertainties in the similarities of the secondary outcomes between the experimental and standard therapy groups.

Review of historical studies for the evaluation of NITs

Reviewing historical data is particularly important for evaluating NITs that do not have placebo groups. A targeted literature search should be conducted to identify historical studies that compare the standard therapy versus placebo and studies on the experimental therapy. Ideally, the targeted search would find the same (or at least largely overlapping) key articles as those used in designing the current NIT. If not, readers should understand why the

search results are different and may conduct further searches as necessary. Readers then need to examine the similarities between the historical studies and the current NIT, including aspects such as: patient characteristics and inclusion criteria, definitions of disease and end points, study methodology and the performance of the standard therapy [8]. In addition, readers need to understand whether disease management (e.g., disease detection and prognosis, and usual care or placebo or adjuvant therapy for the disease) has changed considerably from the time of the historical studies to the current NIT [8,16]. If these aspects are not substantially different between the historical studies and the current NIT, the constancy assumption likely has not been violated. In addition to historical data on the standard therapy versus placebo, it may be beneficial to review earlier studies on the experimental therapy to explore the potential benefits of the experimental therapy over standard therapy and the strength of evidence for the experimental therapy over placebo. Readers can then compare the performance of the experimental therapy in previous publications with the current NIT.

An example of the application of the evaluation list

Postoperative whole breast external beam radiotherapy (external beam) reduces the risk of tumor recurrence and improves breast cancer survival [23,25]. A woman typically receives 42.56–50 grays (Gy) of radiation in 16–25 fractions over 4–5 weeks when treated with external beam radiation therapy. Most women also receive an additional boost of 10–16 Gy in 4–8 fractions [24]. Though the risk of recurrence has drastically decreased over the years, 80–90% of recurrences still occur at the site of the original tumor, regardless of whether the patient had radiotherapy [24,26]. Newer treatment strategies aim to increase the precision at which radiation is delivered. This is typically achieved by reducing the volume of tissue irradiated from whole breast to partial breast, which also decreases the treatment duration. Intraoperative radiotherapy (IORT) using Intrabeam[®] is one of the several options for partial breast irradiation [24]. Delivered to the tumor bed during surgical excision, IORT treatment usually lasts 20–35 min. The radiation dose is about 20 Gy at the surface of the tumor bed and 5–7 Gy for the surrounding tissues. Based on an observational study by Vaidya *et al.*, the 5-year ipsilateral recurrence rate is low, at 1.73% [27]. In a Phase III NIT, Vaidya *et al.* applied a single dose of IORT using Intrabeam without postoperative whole breast external beam radiotherapy and a boost dose in select patients with early stage breast cancer [9]. We will use this NIT to illustrate a practical application of our 14 questions.

1. Was the rationale for conducting a NIT established?

Answer: Yes [×]; No []; Unable to determine []

Explanation: It is reasonable to conduct a NIT to compare Intrabeam and external beam since preliminary evidence has demonstrated the efficacy and safety of Intrabeam and standard therapy for early breast cancer is well-established [23,25,27].

2. Was the standard therapy properly selected?

Answer: Yes [×]; No []; Unable to determine []

Explanation: Postoperative whole breast external beam radiotherapy is the standard therapy strategy for early breast cancer patients [23,25].

3. Did the study select the appropriate effect measures and end points?

Answer: Yes [×]; No []; Unable to determine []

Explanation: For NITs on early stage cancer with low risk of recurrence, the absolute risk difference is often used as the effect measure and local recurrence is often considered as the primary outcome.

4. Was the study well-conducted?

Answer: Yes []; No [×]; Unable to determine [].

Explanation: The Intrabeam group had a greater proportion of patients cross over to the external beam group. In the Intrabeam group, 100 patients did not receive the allocated treatment, including 61 patients who were eventually treated with external beam therapy. In the external beam group, 66 did not receive the allocated treatment, including 10 patients who received Intrabeam alone. The follow-up time was also short (median follow-up time of 2 years).

5. Were the expected benefits of the experimental therapy confirmed in the NIT?

Answer: Yes [×]; No []; Unable to determine [].

Explanation: Although 14% of patients who received Intrabeam also received external beam therapy, a large proportion of patients in the Intrabeam group avoided the 20–30 sessions of external beam therapy. This substantially decreased the workload of healthcare staff and can potentially reduce the wait time for patients. However, complication rates between the Intrabeam and external beam group did not differ significantly, with the exception of significantly reduced toxicity in the Intrabeam group (six patients [0.5%] vs 23 patients [2.1%]) and significantly increased development of seromas requiring more than three aspirations in the Intrabeam group (23 patients [2.1%] vs nine patients [0.8%]). The expected benefits were at least partially achieved.

6. Was the magnitude of the noninferiority margin justifiable?

Answer: Yes [×]; No []; Unable to determine []

Explanation: The authors did not provide a detailed description of how an absolute difference of 2.5% in local recurrence was determined as the noninferiority margin. Authors estimated the background recurrence risk of external beam therapy based on a review by the Early Breast Cancer Trialists' Collaborative Group in 1995 [28]. The review reported a local recurrence rate of 6.7% (501/7473 patients) for radiotherapy versus 19.6% (1480/7570 patients) for control (no radiotherapy). When preserving 50 and 67% effect of the standard therapy [1], we calculated that the noninferiority margins for the absolute risk difference and hazard ratio were 6 and 4% and 1.63 and 1.38, respectively. From this we acknowledge that the authors selected a more conservative noninferiority margin.

7. Was the assay sensitivity presented in the current NIT?

Answer: Yes []; No []; Unable to determine [×].

Explanation: For standard therapy (i.e., the external beam group), the event rate was much lower than that found in historical data. In addition, disease management for early stage breast cancer has evolved worldwide (e.g., earlier detection due to breast cancer screening programs). Therefore, it is unclear whether the assay sensitivity was presented in the current NIT.

8. Was the observed risk in the standard therapy group close to the expected value?

Answer: Yes []; No [×]; Unable to determine []

Explanation: Authors predicted the 5-year recurrence rate following external beam to be 6%. However, the observed 4-year local recurrence rate was only 0.95% in the external beam group, or approximately 16% of the expected value. Therefore, the observed risk in the standard therapy group was not close to the expected value.

9. Was the follow-up duration adequate?

Answer: Yes []; No [×]; Unable to determine [].

Explanation: The median follow-up time was about 2 years and only 420 (20%) patients had a follow-up time of 4 years or longer. The median time to local recurrence was between 40 and 65 months following standard radiotherapy [19]. Thus, the follow-up duration for the study was inadequate and a follow-up time of 5 years or more is needed to capture additional local recurrence events.

10. Did authors conduct both intention-to-treat and per-protocol analysis?

Answer: Yes []; No [×]; Unable to determine [].

Explanation: Authors only conducted intention-to-treat analysis because 23.3% of the 1113 patients originally randomized to Intrabeam did not receive the assigned therapy: 17 patients (1.5%) dropped out, 100 patients (9%) did not receive Intrabeam and 142 (12.8%) patients received both Intrabeam and external beam radiotherapy. The absence of a significant difference from a per-protocol analysis would have provided stronger assurance of noninferiority.

11. Did the survival analysis adhere to the random censoring assumption?

Answer: Yes []; No []; Unable to determine [×].

Explanation: In this NIT, patients were censored for death, undergoing mastectomy for any reason, withdrawal and loss to follow-up. However, authors did not report additional details for censoring.

12. Did the adjuvant therapy contribute to the clinical outcomes?

Answer: Yes [☒]; No [☐]; Unable to determine [☐]

Explanation: Most patients received additional adjuvant therapies. Overall 66% of the patients received hormone therapy and 12% of the patients received additional chemotherapy, with approximately equal numbers of patients in each arm. Since both adjuvant treatments are associated with reduced risk of local recurrence [19], the intended comparison of Intrabeam versus external beam is in fact a comparison of Intrabeam and adjuvant therapy versus external beam and adjuvant therapy.

13. Is the point estimate between the experimental and standard therapy close to no effect for the primary clinical outcomes?

Answer: Yes [☐]; No [☐]; Unable to determine [☒]

Explanation: At 4-year follow-up, six (1.2%) local recurrences were identified in the Intrabeam group compared with five (0.95%) in the external beam group, with no statistically significant difference between the two groups (log rank test: $p > 0.05$). The difference in recurrence rates between the two groups was 0.25% (95% CI: -1.04 to 1.54%), thus falling within the predefined noninferiority margin of 2.5%. However, due to the extremely low event rate, it was difficult to conclude the similarity of the two therapies for local recurrence.

14. Other than the primary end point, were other benefits of the standard therapy also obtained by the experimental therapy?

Answer: Yes [☐]; No [☐]; Unable to determine [☒].

Explanation: In addition to reducing the risk of tumor recurrence, whole breast external beam therapy also improves survival. A meta-analysis conducted by Clarke *et al.* demonstrated that additional radiotherapy resulted in a 15-year breast cancer mortality reduction of 5.4% ($p = 0.0002$) and a 15-year overall mortality reduction of 5.3% ($p = 0.005$) [23]. In contrast to the evidence for whole breast external beam therapy, Intrabeam radiotherapy remains in the research domain, supported so far by only a single RCT with relatively short follow-up time. It is therefore unclear whether Intrabeam can achieve any of the long-term survival benefits of external beam therapy.

Based on the answers from the 14 questions, we may conclude that compared with external beam radiotherapy, the strength of the evidence supporting the efficacy and safety of Intrabeam is still very limited. This was demonstrated by: the short follow-up duration and the associated low recurrence risk (questions 8 and 9), no per-protocol analysis (question 10), unreported informative censoring (question 11), the effect of adjuvant therapy (question 12), and an unclear benefit in survival (question 14). An adequate number of patients would have to complete at least 5 years of follow-up before a definitive conclusion on the efficacy of Intrabeam can be drawn from this NIT.

Acknowledgements

The authors would like to thank I Dhalla from Health Quality Ontario, Toronto, Canada and two anonymous reviewers for their valuable comments.

Disclaimer

The opinions expressed in this publication do not necessarily represent the opinions of Health Quality Ontario. No endorsement is intended or should be inferred.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/full/10.2217/cer-2018-0035

Summary points

- 14 questions were proposed to critically appraise noninferiority trials (NITs) and support proper interpretation of study results.
- To conclude noninferiority conceptually is to synthesize evidence from both the current NIT comparing experimental therapy with standard therapy, and historical data comparing standard therapy with placebo control. Therefore, readers should explore external data sources to validate the current NIT.
- The rationale for conducting a NIT should be well-established before conducting a NIT.
- Poorly conducted NITs are likely to bias the results toward the alternative hypothesis and lead to a conclusion of noninferiority.
- To examine the constancy assumption, readers need to check the similarities between historical studies and the current NIT, including aspects such as: patient characteristics and inclusion criteria, definitions of disease and end points, study methodology and the performance of the standard therapy.
- The adjuvant therapy would reduce the effect difference between the experimental and standard therapy. Since the effects of various therapies cannot be separated, when the effect of the adjuvant therapy is relatively large, it leads to more uncertainty in estimating the true difference between the experimental therapy and the standard therapy.
- The conclusions of a NIT are highly sensitive to both effect measurements and the underlying risk in the standard therapy group. It is important to examine whether the observed risk in the standard therapy group in the NIT was close to the expected value used in designing the NIT.
- Demonstrating noninferiority in both an intention-to-treat and per-protocol analysis provides the necessary assurance of noninferiority.

Open access

This work is licensed under Crown copyright protection and licensed for use under the open government licence unless otherwise indicated. Where any of the Crown copyright information in this work is republished or copied to others, the source of the material must be identified and the copyright status under the open government licence acknowledged. www.nationalarchives.gov.uk/do/c/open-government-licence/version/3/

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Wangge G, Roes KC, De Boer A, Hoes AW, Knol MJ. The challenges of determining noninferiority margins: a case study of noninferiority randomized controlled trials of novel oral anticoagulants. *Can. Med. Assoc. J.* 185(3), 222–227 (2013).
2. Tanaka S, Kinjo Y, Kataoka Y, Yoshimura K, Teramukai S. Statistical issues and recommendations for noninferiority trials in oncology: a systematic review. *Clin. Cancer Res.* 18(7), 1837–1847 (2012).
3. Wangge G, Klungel OH, Roes KC, De Boer A, Hoes AW, Knol MJ. Room for improvement in conducting and reporting noninferiority randomized controlled trials on drugs: a systematic review. *PLoS ONE* 5(10), e13550 (2010).
4. Schiller P, Burchardi N, Niestroj M, Kieser M. Quality of reporting of clinical noninferiority and equivalence randomized trials: update and extension. *Trials* 13, 214 (2012).
5. Soonawala D, Middelburg RA, Egger, M Vandenbroucke JP, Dekkers OM. Efficacy of experimental treatments compared with standard treatments in noninferiority trials: a meta-analysis of randomized controlled trials. *Int. J. Epidemiol.* 39(6), 1567–1581 (2010).
6. D'agostinoRB, Sr Massaro JM, Sullivan LM. Noninferiority trials: design concepts and issues: the encounters of academic consultants in statistics. *Stat. Med.* 22(2), 169–186 (2003).
7. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *J. Am. Med. Assoc.* 308(24), 2594–2604 (2012).
8. USFDA. Noninferiority clinical trials to establish effectiveness: guidance for industry. (2016) [/www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf](http://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf)
- **Guidance document on designing noninferiority trials (NITs). Our proposed 14 questions to evaluate NITs are mainly based on this document.**
9. Vaidya JS, Joseph DJ, Tobias JS *et al.* Targeted intraoperative radiotherapy versus whole breast radiotherapy for breast cancer (TARGIT-A trial): an international, prospective, randomized, noninferiority Phase III trial. *Lancet* 376(9735), 91–102 (2010).
- **NIT on an intraoperative radiation therapy for early stage breast cancer, which was used as an example NIT to demonstrate the application of the 14 questions.**

10. Lavallo-Gonzalez FJ, Januszewicz A, Davidson J *et al.* Efficacy and safety of canagliflozin compared with placebo and sitagliptin in patients with type 2 diabetes on background metformin monotherapy: a randomized trial. *Diabetologia* 56(12), 2582–2592 (2013).
11. Riechelmann RP, Alex A, Cruz L, Bariani GM, Hoff PM. Noninferiority cancer clinical trials: scope and purposes underlying their design. *Ann. Oncol.* 24(7), 1942–1947 (2013).
12. Li-Ching H, Miin-Jye W, Hung CS, Shing KK. Noninferiority studies with multiple reference treatments. *Stat. Methods Med. Res.* 26(3), 1295–1307 (2017).
13. Saad ED, Buyse M. Noninferiority trials in breast and non-small-cell lung cancer: choice of noninferiority margins and other statistical aspects. *Acta Oncol.* 51(7), 890–896 (2012).
14. Xie X, Ye C, Mitsakakis N. The impact of the underlying risk in control group and effect measures in noninferiority trials with time-to-event data: a simulation study. *J. Clin. Med. Res.* 10(5), 376–383 (2018).
15. Buzdar AU, Xu B, Digumarti R *et al.* Randomized Phase II noninferiority study (NO16853) of two different doses of capecitabine in combination with docetaxel for locally advanced/metastatic breast cancer. *Ann. Oncol.* 23(3), 589–597 (2012).
16. European Medicines Agency. Guideline on the choice of the noninferiority margin. (2005)
www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf
- **Guidelines on selecting the noninferiority margin and other aspects of NIT design. This document contributed to the discussion of NITs with or without a placebo group.**
17. Huitfeldt B, Hummel J, European Federation of Statisticians in the Pharmaceutical Industry, the draft FDA guideline on noninferiority clinical trials: a critical review from European pharmaceutical industry statisticians. *Pharm. Stat.* 10(5), 414–419 (2011).
18. Wang SJ, Hung HM. TACT method for noninferiority testing in active controlled trials. *Stat. Med.* 22(2), 227–238 (2003).
19. Reitsamer R, Fastner G, Kopp M, Menzel C, Sedlmayer F. Intraoperative radiotherapy for early breast cancer. *Lancet* 376(9747), 1141–1144 (2010).
20. Mauri L, D’agostino RB Sr. Challenges in the design and interpretation of noninferiority trials. *N. Engl. J. Med.* 377(14), 1357–1367 (2017).
21. Shah PB. Intention-to-treat and per-protocol analysis. *Can. Med. Assoc. J.* 183(6), 696 (2011).
22. Sacco RL, Diener HC, Yusuf S *et al.* Aspirin and extended-release dipyridamole versus clopidogrel for recurrent stroke. *N. Engl. J. Med.* 359(12), 1238–1251 (2008).
23. Clarke M, Collins R, Darby S *et al.* Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomized trials. *Lancet* 366(9503), 2087–2106 (2005).
24. Mancias JD, Taghian AG. Accelerated partial breast irradiation using TARGIT: the pros, cons and the need for long-term results. *Expert Rev. Anticancer Ther.* 10(12), 1869–1875 (2010).
25. Blank E, Kraus-Tiefenbacher U, Welzel G *et al.* Single-center long-term follow-up after intraoperative radiotherapy as a boost during breast-conserving surgery using low-kilovoltage x-rays. *Ann. Surg. Oncol.* 17 (Suppl. 3), 352–358 (2010).
26. Vaidya JS, Baum M, Tobias JS *et al.* Targeted intraoperative radiotherapy (Targit): an innovative method of treatment for early breast cancer. *Ann. Oncol.* 12(8), 1075–1080 (2001).
27. Vaidya JS, Baum M, Tobias JS *et al.* Long-term results of targeted intraoperative radiotherapy (Targit) boost during breast-conserving surgery. *Int. J. Radiat. Oncol. Biol. Phys.* 81(4), 1091–1097 (2011).
28. Early Breast Cancer Trialists’ Collaborative Group. Effects of radiotherapy and surgery in early breast cancer: an overview of the randomized trials. *N. Engl. J. Med.* 333(22), 1444–1455 (1995).

