GOOD SCIENCE
BETTER MEDICINE
BEST PRACTICE

Annals of Oncology 29: 573–577, 2018
doi:10.1093/annonc/mdx788
Published online 18 December 2017

REVIEW

Interim monitoring for non-inferiority trials: minimizing patient exposure to inferior therapies

E. L. Korn* & B. Freidlin

Biometric Research Program, National Cancer Institute, Bethesda, USA

**Correspondence to*: Prof. Edward L. Korn, Biometric Research Program, MSC 9735, National Cancer Institute, 9609 Medical Center Drive, Bethesda, MD 20892, USA.
Tel:+1-240-276-6029; E-mail: korne@ctep.nci.nih.gov

The goal of a non-inferiority randomized trial is to demonstrate that an experimental treatment is not unacceptably worse than a standard treatment. The experimental treatment is known to have less toxicity or other quality-of-life benefits when compared with the standard treatment, so that a small decrement in efficacy would be acceptable. Interim monitoring of randomized trials is used to stop trials early if the conclusions of the trial become definitive early. In the context of a non-inferiority trial, of special interest is stopping a trial early when the experimental treatment is inferior to the standard treatment. Methods for performing interim monitoring of non-inferiority trials are reviewed for their ability to minimize patient exposure to inferior experimental treatments. Examples of trials from the literature are discussed along with a computer simulation of a simple non-inferiority monitoring rule. Interim monitoring for non-inferiority trials is shown to substantially reduce the exposure of patients to inferior therapies when, in fact, the experimental treatment is inferior to the standard treatment. Interim monitoring rules typically used in superiority trials may be sub-optimal for non-inferiority trials, and may unnecessarily expose patients to inferior therapies. Examples of trials with inferior experimental arms and trials with sub-optimal monitoring rules are given. Appropriate interim monitoring of non-inferiority trials can reduce the exposure of patients to inferior therapies when the experimental treatment is inferior to the standard treatment.

Key words: futility monitoring, interim monitoring guidelines, non-inferiority randomized trials

Introduction

Non-inferiority randomized trials are motivated by the knowledge or assumption that an experimental therapy is less toxic and/or results in a better quality of life than a standard therapy [1, 2]. These trials are designed to provide conclusive evidence that substituting the experimental therapy for the standard therapy will not decrease efficacy (e.g. survival) by an unacceptable amount. Interim monitoring is an important feature of clinical trials: it allows stopping a trial if accruing data are sufficiently compelling to answer the study question early. There are two directions for early stopping: a trial might be (i) stopped because non-inferiority is clear, or (ii) stopped when the experimental treatment is unacceptably inferior to the standard treatment (harm) or unlikely to be shown to be non-inferior to the standard therapy (futility). The latter direction has particular importance because it protects patients from exposure to an inferior therapy (instead of receiving the standard of care with proven benefit). However, Tanaka et al. [3] reported in a review of 72 oncology non-inferiority trials published in 2000–2010 that only 36% of trials reported having a planned interim analysis. In addition, even when non-inferiority trials include formal interim monitoring, the specified plans may frequently be sub-optimal for protecting patients from inferior experimental therapies.

Non-inferiority trial design

The key design parameter in a non-inferiority trial is the unacceptable decrement in efficacy, defined by the non-inferiority margin, e.g. a hazard ratio (HR) (experimental treatment over standard treatment) whose value or larger would be considered clinically unacceptable. The typical decision rule at the end of a non-inferiority trial is based on a two-sided 95% (or 90%) confidence interval for the HR: if the confidence interval is completely below the non-inferiority margin, then non-inferiority of the experimental treatment is declared. With this decision rule, the probability of erroneously concluding non-inferiority when the experimental arm is inferior by at least the non-inferiority margin (type 1 error) will be low, 2.5% (or 5%). Choosing an appropriate non-inferiority margin for a given setting can be challenging; it involves the known benefit of the standard treatment over a (previous) reference (or no) treatment, the presumed toxicity reduction (or other quality-of-life benefits) of the experimental treatment, the nature of the efficacy outcome variable (e.g. overall survival, OS, versus progression-free survival, PFS), and the clinical setting [1, 4–9]. Of special concern is choosing a non-inferiority margin that is so large that a positive trial may not rule out an unacceptable loss of efficacy [10]. The sample size of a non-inferiority trial is chosen to ensure that the trial has a high probability (power, typically 90% or 80%) of declaring non-inferiority when the experimental and standard treatments are equally efficacious.

Interim monitoring for harm (inferiority) or futility

A simple intuitive approach to monitoring a non-inferiority trial for harm or futility is to stop the trial when half the expected events have occurred (50% information time) if the observed HR is equal or worse than the non-inferiority margin. This rule, which is an adaption of a commonly used futility rule for superiority trials [11, 12], can be further generalized to other information times by adapting a method described in Anderson and High [13] for superiority trials: at each monitoring time, the one-sided *P*-value is calculated for testing the hypothesis HR = 1 versus the alternative HR > 1 (meaning the experimental treatment is doing worse than the standard treatment). If the *P*-value is <0.0110 at a monitoring time, then the trial would stop with the conclusion that non-inferiority cannot be claimed. The cut-off *P*-value 0.0110 corresponds to the observed HR being exactly the non-inferiority margin (at 50% information) in a trial with 90% power and 2.5% type 1 error [13] (see footnote c of Table 1 for *P*-value cut-offs for trials with other design parameters).

The benefit of harm/futility monitoring in terms of reducing trial duration and the number of patients exposed to inferior therapy is illustrated in Table 1 by simulating the properties of a trial with non-inferiority margin HR = 1.2 with no interim analyses, one interim analysis (at 50% information), or two interim analyses (at 25% and 50% information). The numbers not in parentheses represent interim analyses carried out when the total number of events on both arms (pooled events) reach 25% or 50% of the required total for the final analysis; this is the most commonly used approach. Note that when the true HR is equal to the non-inferiority margin, the months and patients saved by using one or two interim analyses are quite large, and become even larger when the HR is larger than the non-inferiority margin. For example, if the true HR = 1.4, then the 1 or 2 interim analyses would reduce the average length of the trial from 77.3 to 48.9 and 36.9 months, respectively, and the number of patients on the experimental-treatment arm from 1000 to 808 and 609, respectively. As these numbers are averages, there can be larger savings for some trials. For example, 25% of the trials with two interim analyses would have less than 532 patients accrued to the experimental arm (compared with 1000 patients without interim monitoring) when the true HR = 1.4. Finally, it should be noted that these simulation results potentially underestimate the benefit due to the interim analyses, since the patients who are sufficiently early in their course of experimental therapy may be able so switch to the standard treatment.

Why are the increased reductions in the study duration under HRs larger than the non-inferiority margin relevant? Non-inferiority trials with poorly performing experimental arms do occur. Examples are given in Table 2. As these examples demonstrate, experimental treatments are sometimes much worse than the standard treatment in non-inferiority trials. In these situations, appropriate interim monitoring rules can dramatically reduce patient exposure to inferior therapies by stopping accrual of new patients (if accrual is still ongoing) and potentially discontinuing treatment of already enrolled patients who are still receiving the experimental treatment. (Note that the trials in Table 2 correctly detected the inferiority of the experimental therapies.)

When the experimental therapy is inferior (and thus events are accumulating faster on the experimental arm than on the control arm), a modest improvement in the timing of interim analyses can be obtained by conducting the interim analysis when one-half the required number of events occur in the experimental-treatment arm, if this happens before the required number of events is seen in both arms. This is known as Earliest Information Time (EIT) [14]. For example, instead of performing the 50% interim analysis when 632 (=1264/2) events in total have occurred in both treatment arms (Table 1), the 50% analysis is carried out when either 632 total events or 316 (=632/2) experimental-arm events have occurred, whichever occurs first. The simulated average duration and size of trials using EIT are given in parentheses in Table 1, and show an improvement over using the standard pooled-events approach when the experimental arm is doing much worse than standard arm. (Planned timing of analyses based solely on the occurrence of events in the standard treatment arm, which is sometimes done in multi-arm trials [14, 15], is a mistake as it may delay analysis times when the experimental arm(s) are doing poorly when compared with the standard-treatment arm.) When using EIT (or the timing of interim analyses bases on standard-treatment arm events) it is important to ensure that the total number of events that has occurred in an ongoing trial is not made public to avoid unintentional release of information about between-arm efficacy.

The cost of using futility interim monitoring is a slight loss of <2% in power for the trial [11]. There is also a very slight additional loss of power in using EIT over the standard pooled-events timing approach, <0.08% for the examples in Table 1(footnote f). It is worth noting a potential complication with futility monitoring that may occur if the observed survival in the standard-treatment arm is much better than expected and the non-inferiority margin was specified as a HR rather than an absolute difference in survival. For example, consider a trial designed to detect a non-inferiority HR margin of 1.3, corresponding to an 8.5% reduction from the expected control-arm 3-year OS rate of 60%. If the 3-year OS rate on the control arm turned out to be 90%, then the futility monitoring rule would suggest stopping the study for futility if the observed 3-year OS on the

Table 1. Average duration (months) and number of patients accrued on the experimental-treatment arm of a non-inferiority trial[a] with no, one, or two interim analyses for harm or futility

True hazard ratio[b]	Interim analyses[c]					
	None		One at 50% information based on pooled (EIT) events[d]		Two at 25% and 50% information based on pooled (EIT) events[e]	
	Duration	No. of patients	Duration	No. of patients	Duration	No. of patients
1.0[f]	84.8	1000	84.4 (84.4)	998 (998)	83.9 (83.9)	994 (994)
1.2	80.5	1000	65.1 (64.7)	915 (900)	60.1 (59.8)	842 (838)
1.4	77.3	1000	48.9 (46.1)	808 (758)	36.9 (35.8)	609 (590)
1.6	74.7	1000	46.7 (42.7)	780 (712)	31.7 (29.2)	529 (487)
2	70.9	1000	44.5 (39.1)	743 (653)	29.6 (25.8)	493 (431)

[a] The trial design is 1000 patients uniformly accrued per arm over 60 months, with assumed median survival of 36 months for the standard treatment arm, generated with an exponential distribution. The non-inferiority hazard ratio margin is taken to be 1.2. The type 1 error (probability of declaring non-inferiority when the true hazard ratio is equal to the non-inferiority margin) is taken to be 2.5%, and the power (probability of declaring non-inferiority when the hazard ratio equals 1) is taken to be 90% (with no interim analyses). The final analysis is done when a total of 1264 events occur in both treatment arms.

[b] The hazard ratio (HR) is the hazard of the experimental treatment over the hazard of the standard treatment, so that values >1 represent the experimental treatment being worse than the standard treatment.

[c] At an interim analysis, the trial would be stopped for harm or futility if the (one-sided) P-value for testing $\mathrm{HR} = 1$ versus $\mathrm{HR} > 1$ is less than $P = 0.0110$. If the trial were designed with type 1 error of 5% instead of 2.5%, the P-value cut-off would be $P = 0.0193$. If the trial had 80% instead of 90% power, the cut-offs would be $P = 0.0238$ and $P = 0.0394$ for 0.025 and 0.05 type 1 error designs, respectively [13].

[d] A single interim analysis is carried out at 50% information for stopping the trial if the experimental (reduced) treatment appears sufficiently worse than the standard treatment (see footnote c). The analysis occurs with pooled-events timing when there are 632 events in total in both arms; the analysis occurs with EIT timing when there are 632 events in both arms or 316 events in the experimental-treatment arm, whichever occurs first.

[e] Interim analyses are carried out at 25% and 50% information for stopping the trial if the experimental (reduced) treatment appears worse than the standard treatment (see footnote c). The analysis occurs with pooled-events timing when there are 316 events (25% information) and 632 events (50% information) in total in both arms; the analysis occurs with EIT timing when there are at 316 events in both arms or 158 events in the experimental arm, whichever occurs first (for the first interim analysis), and when there are 632 events in both arms or 316 events in the experimental-treatment arm, whichever occurs first (for the second interim analysis).

[f] The (simulated) powers are 90.04%, 89.80%, and 89.34% for no, one and two interim analyses using pooled events timing. With EIT timing for the interim analyses, the powers are 89.75% and 89.26% for one and two interim analyses.

Table 2. Examples of non-inferiority trials with poorly performing experimental-treatment arms

Disease setting	Reference number	End point	Non-inferiority margin (HR)	Observed HR	95% CI for observed HR
Early-stage breast cancer	16	Relapse-free survival	1.24	2.09	(1.38, 3.17)
Extensive-stage small-cell lung cancer	17	Overall survival	1.176	1.56	(1.27, 1.92)
Locally advanced prostate cancer	18	Overall survival	1.35	1.42	(1.09, 1.84)
Advanced breast cancer	19	Progression-free survival	1.25	1.37	(1.13, 1.65)
Multiple myeloma	20	Progression-free survival	1.43	2.51	(1.60, 3.94)

experimental arm was 87.2% (corresponding to a HR of 1.3); this may not make clinical sense, and the protocol would need to be amended to reflect the observed control-arm survival.

Sub-optimal interim monitoring plans in practice

Randomized clinical trials that are not small should have a data monitoring committee to monitor accruing outcome data. Additionally, pre-specification of formal monitoring guidelines allows the trial investigators to have input on the stopping rules and their statistical operating characteristics. Moreover, prospective specification of monitoring rules may improve transparency and the statistical validity of the trial conclusions. Even trials that have formal interim monitoring sometimes specify suboptimal plans in their protocols. For example, non-inferiority trials frequently use monitoring developed for superiority trials, e.g. the Haybittle–Peto [21] or O'Brien–Fleming [22] monitoring plans. However, these plans may offer insufficient protection in the non-inferiority setting when the experimental therapy is inferior to the standard treatment. For example, in a trial to assess the non-inferiority (in terms of biochemical failure) of a shorter duration of radiation therapy for localized prostate cancer with a non-inferiority margin $\text{HR} = 1.32$, the interim monitoring was specified (using Haybittle–Peto) as stopping if the *P*-value was <0.001 for testing whether the $\text{HR} = 1$ [23]. This rule would require an observed $\text{HR} > 1.57$ to stop at 50% information, as opposed to our rule that would only require $\text{HR} > 1.32$. (The experimental therapy was, in fact, non-inferior in this trial [23].)

Interim monitoring for declaring non-inferiority early

If the early-trial results allow one to rule out the non-inferiority margin then it is theoretically possible to stop a trial and conclude that the experimental treatment is non-inferior. However, this is generally not recommended [4, 8]. As the choice of non-inferiority margin involves some subjectivity, obtaining more information on the HR (and a smaller confidence interval for it) will be beneficial even if one could formally rule out the non-inferiority margin earlier. An exception to this would be when the experimental treatment is shown to be superior (in terms of the primary efficacy end point) than the standard treatment, e.g. statistically significantly better. In this case, it would be appropriate to stop the trial early (and one could include a pre-specified superiority monitoring guideline for that possibility).

Non-inferiority trials where modest superiority is expected

In some cases it is expected that the experimental therapy will be marginally better than the standard therapy, e.g. with a new less-toxic agent (as opposed to a reduction in therapy). In these settings, a hybrid non-inferiority approach that is designed to distinguish between the non-inferiority margin and a small improvement Δ ($\Delta < 1$) is sometimes used [7, 24, 25]. An example is given by the MA.31 trial [19], which assessed the PFS non-inferiority of lapatinib versus trastuzumab combined with taxanes for HER2-positive advanced breast cancer patients. The trial was designed with a non-inferiority margin $\text{HR} = 1.25$, but the power was calculated under the $\text{HR} = 0.9$ (Δ). The interim monitoring for harm or futility described previously can easily be modified for these trial designs: for a trial with 90% power and 2.5% type 1 error, rather than calculating the *P*-value for testing the $\text{HR} = 1$ and checking if it is <0.0110, one calculates the *P*-value for testing the $\text{HR} = \Delta$ and checks if it is <0.0110. (At 50% information, this corresponds to stopping if the observed $\text{HR} > 1.25$.) There is a slight power loss in using interim monitoring in this setting: 0.3% when the $\text{HR} = 0.9$, and 0.4% when the $\text{HR} = 1$. The MA.31 protocol used an O'Brien–Fleming boundary and specified stopping at 50% information if the (one-sided) *P*-value was <0.0015 for testing the $\text{HR} = 1$; this is a very conservative rule, e.g. requiring an observed $\text{HR} > 1.53$ to stop. (The data monitoring committee did, in fact, recommend early disclosure of inferiority for this trial at 67% information [19].)

Preliminary results reporting

Non-inferiority trials often involve long follow-up to allow the time-to-event outcomes to mature. Since many of these trials compare therapies that are already widely used, the clinical community may benefit from access to preliminary trial results. In such cases, in very special situations, it may be possible to report preliminary results of an ongoing non-inferiority trial (with final results reported later) even though an interim-analysis boundary has not been crossed. To preserve the study integrity, one should only consider early reporting when patients are off treatment-arm-specific therapies and the preliminary release of the results is very unlikely or impossible to influence the final trial results [26]. In particular, it should be unlikely or impossible, based on knowledge of the preliminary results, (i) that patients on the trial would modify their subsequent treatment and (ii) that the intensity of follow-up would be modified. (In general, it is not a good idea to release results of an ongoing trial [27, 28].) An example is given by the report of preliminary results of a trial to assess the RFS non-inferiority of 2 years of oral uracil-tegafur compared with six cycles of CMF as adjuvant treatment of node-negative high-risk breast cancer, after all patients had been followed for at least 5 years [29]. The valid theoretical concern [30] that reporting preliminary data from a trial may have a negative impact on other ongoing trials is probably not a major practical concern for non-inferiority trials, where it is unlikely that multiple trials would be simultaneously addressing the same non-inferiority question.

Discussion

Unless early evidence strongly suggests that the experimental arm is superior to the standard of care, early stopping with the conclusion of non-inferiority may be counterproductive. On the other hand, a formal plan for interim analyses for harm or futility, which we recommend as a guide for the deliberations of a data monitoring committee, is important to minimize patient exposure to inferior therapies, and therefore should practically always be included in the trial design. If not included, justification for its omission should be given in the protocol or report of the trial [31]. In some special circumstances, it may be possible to report preliminary results of a non-inferiority trial while awaiting sufficient events to occur for the final analysis.

Funding

None declared.

Disclosure

The authors have declared no conflicts of interest.

References

1. Food and Drug Administration. Non-inferiority clinical trials to establish effectiveness: guidance for industry 2016; https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf.

2. Riechelmann RP, Alex A, Cruz L. Non-inferiority cancer clinical trials: scope and purposes underlying their design. Ann Oncol 2013; 24(7): 1942–1947.
3. 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(7): 1837–1847.
4. D’Agostino RB, Sr., Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. Stat Med 2003; 22: 169–186.
5. Rothmann M, Li N, Chen G et al. Design and analysis of non-inferiority trials in oncology. Stat Med 2003; 22(2): 239–264.
6. Committee for Medicinal Products for Human Use (CHMP). Guideline on the choice of the non-inferiority margin. Stat Med 2006; 25: 1628–1638.
7. Fleming TR. Current issues in non-inferiority designs. Statist Med 2008; 27: 17: 317–332.
8. Saad ED, Buyse M. Non-inferiority trials in breast and non-small cell lung cancer: choice of non-inferiority margins and other statistical aspects. Acta Oncol 2012; 51(7): 890–896.
9. Burotto M, Prasad V, Fojo T. Non-inferiority trials: why oncologists must remain wary. Lancet Oncol 2015; 16(4): 364–366.
10. Schmidinger M, Wittes J. First-line treatment of metastatic renal cell carcinoma after COMPARZ and PISCES. Curr Opin Urol 2015; 25(5): 395–401.
11. Wieand S, Schroeder G, O’Fallon JR. Stopping when the experimental regime does not appear to help. Stat Med 1994; 13(13–14): 1453–1458.
12. Freidlin B, Korn EL, Gray R. A general inefficacy interim monitoring rule for randomized clinical trials. Clin Trials 2010; 7(3): 197–208.
13. Anderson JR, High R. Alternatives to the standard Fleming, Harrington, and O’Brien futility boundary. Clin Trials 2011; 8(3): 270–276.
14. Freidlin B, Othus M, Korn EL. Information time scales for interim analyses of randomized clinical trials. Clin Trials 2016; 13(4): 391–399.
15. Dearnaley D, Syndikus I, Mossop H et al. Conventional versus hypofractionated high-dose intensity-modulated radiotherapy for prostate cancer: 5-year outcomes of the randomised, non-inferiority, phase 3 CHHiP trial. Lancet Oncol 2016; 17(8): 1047–1060.
16. Muss HB, Berry DA, Cirrincione CT et al. Adjuvant chemotherapy in older women with early-stage breast cancer. N Engl J Med 2009; 360: 2055–2065.
17. Socinski MA, Smit EF, Lorigan P et al. Phase III study of pemetrexed plus carboplatin compared with etoposide plus carboplatin in chemotherapy-naive patients with extensive-stgae small-cell lung cancer. J Clin Oncol 2009; 27(28): 4787–4792.
18. Bolla M, de Reijke TM, Van Tienhoven G et al. Duration of androgen suppression in the treatment of prostate cancer. N Engl J Med 2009; 360: 2516–2527.
19. Gelmon KA, Boyle FM, Kaufman B et al. Lapatinib or trastuzumab plus taxane therapy for human epidermal growth factor receptor 2-positve advanced breast cancer: final results of NCIC CTG MA.31. J Clin Oncol 2015; 33(14): 1574–1583.
20. Gay F, Oliva S, Petrucci MT et al. Chemotherapy plus lenalidomide versus autologous transplantation, followed by lenalidomide plus prednisone versus lenalidomide maintenance, in patients with multiple myeloma: a randomized, multicentre, phase 3 trial. Lancet Oncol 2015; 16(16): 1617–1629.
21. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol 1971; 44(526): 793–797.
22. O’Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35: 549–4556.
23. Catton CN, Lukka H, Gu C-S et al. Randomized trial of a hypofractionated radiation regimen for the treatment of localized prostate cancer. JCO 2017; 35: 184–1890.
24. Freidlin B, Korn EL, George SL et al. Randomized clinical trial design for assessing noninferiority when superiority is expected. JCO 2007; 25: 5019–5023.
25. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. Trials 2011; 12: 106.
26. Korn EL, Hunsberger S, Freidlin B et al. Preliminary data release for randomized clinical trials of noninferiority: a new proposal. J Clin Oncol 2005; 23: 5831–5835.
27. Fleming TR, Sharples K, McCall J et al. Maintaining confidentiality of interim data to enhance trial integrity and credibility. Clin Trials 2008; 5(2): 151–167.
28. Korn EL, Hunsberger S, Freidlin B et al. Comments on ‘Maintaining confidentiality of interim data to enhance trial integrity and credibility’ by TR Fleming et al. Clin Trials 2008; 5(4): 364–366.
29. Watanabe T, Sano M, Takashima S et al. Oral uracil and tegafur compared with classic cyclophosphamide, methotrexate, fluorouracil as postoperative chemotherapy in patients with node-negative, high-risk breast cancer: National Surgical Adjuvant Study for Breast Cancer 01 Trial. JCO 2009; 27: 1368–1374.
30. Dignam JJ. Early viewing of noninferiority trials in progress. J Clin Oncol 2005; 23(24): 5461–5463.
31. Korn EL, Freidlin B. Inefficacy interim monitoring procedures in randomized clinical trials: the need to report. Am J Bioethics 2011; 11(3): 2–10.