Issues with Specification of Margins and Analytic Methods on Potential Bias Towards the Alternative in Randomized Noninferiority Pragmatic Trials: Results of a Literature Survey

Maria Ciarleglio (maria.ciarleglio@yale.edu)  
Yale University  
https://orcid.org/0000-0002-6591-3674

Jiaxuan Li  
Yale University

Peter Peduzzi  
Yale University

Research Article

Keywords: Pragmatic trial, noninferiority trial, estimands

Posted Date: August 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1846419/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Issues with Specification of Margins and Analytic Methods on Potential Bias Towards the Alternative in Randomized Noninferiority Pragmatic Trials: Results of a Literature Survey

Maria Ciarleglio, Jiaxuan Li, Peter Peduzzi
Department of Biostatistics, Yale Center for Analytical Sciences, Yale School of Public Health, 60 College Street, New Haven, CT 06510

Corresponding author: Maria Ciarleglio, maria.ciarleglio@yale.edu

Keywords: Pragmatic trial; noninferiority trial; estimands

Words: 2399 (text)
Abstract (348 words)

Background: Issues with specification of margins and analytic methods can potentially bias results towards the alternative in randomized noninferiority pragmatic trials. To investigate this potential for bias we conducted a targeted search of the medical literature to examine how noninferiority pragmatic trials address these issues.

Methods: An Ovid MEDLINE database search was performed that identified any publications in New England Journal of Medicine, Journal of the American Medical Association, Lancet, or British Medical Journal published between 2015 and 2021 (inclusive) that included the words “pragmatic” or “comparative effectiveness” as well as “noninferiority” or “non-inferiority” in a multi-purpose search. Our search identified 14 potential trials of which 12 met our inclusion criteria.

Results: Of the 12 randomized pragmatic noninferiority trials, 11 were individually randomized trials and one was a cluster randomized trial. Ten of the 11 individually randomized trials met the criteria established for noninferiority as did the one cluster randomized trial. Noninferiority margins were prespecified for all the trials. The majority of margins (6) were based on either minimum clinically important differences, clinical experts, or consensus, while others were based on sample size, empirical data, or clinical decision. For two trials, no justification for the margin was provided. All trials conducted intent to treat or modified intent to treat analyses along with per protocol analyses and reached similar conclusions. Only two trials included all randomized participants in the primary analysis, one of which used multiple imputation to impute missing data. The percentage of participants excluded from the primary analysis ranged from about 2% to nearly 30% and sometimes differed between treatment arms.

Conclusions: Specification of margins and methods of analysis require careful consideration to prevent bias towards the alternative in noninferiority trials. Much of the guidance on these two issues has been developed for a regulatory environment and not for pragmatic noninferiority trials. Since many pragmatic trials generally follow the PRECIS criteria of little or no monitoring
of participant or practitioner adherence, it affects separation of treatments which in turn affects
both the setting of margins and analysis. More recent developments on estimands can address
the latter issue.
Introduction

Pragmatic trials are designed to produce more generalizable results than the typical efficacy trial. By their very nature, pragmatic trials look to assess effectiveness of the treatment under study. Hence, they are designed to mimic clinical practice and often follow the PRECIS or PRECIS-2 criteria in their design (Thorpe et al. 2009; Loudon et al. 2017). Briefly, these trials include heterogenous populations with minimal exclusion criteria, use of interventions that are easy to implement by health care personnel that require little to no monitoring of adherence, use of active controls or standard of care, minimal patient burden with follow-up schedules that mimic clinical practice, and analysis of all randomized participants, i.e., according to the intent to treat principle. Regarding adherence, the PRECIS criteria address both participant and practitioner adherence and state “There is unobtrusive (or no) measurement of participant [practitioner] compliance. No special strategies to maintain or improve compliance are used.”

Guidelines for conducting noninferiority (NI) trials have been published by The Consolidated Standards of Reporting Trials (CONSORT) group (Piaggio et al., 2012), US Food and Drug Administration (2016), and European Medicines Agency (2000). These guidelines emphasize the importance of quality control in the study design of noninferiority trials and focus on the regulatory setting; otherwise, there could be a bias towards claiming noninferiority of the treatment under study or bias towards the alternative. The recent article by Mo et al (2020), which focused on nonadherence in noninferiority trials and the problems with conducting intent to treat and per protocol analyses, also presented the results of a systematic review of five top-tier clinical journals [New England Journal of Medicine (NEJM), Lancet, Journal of the American Medical Association (JAMA), Annals of Internal Medicine, and British Medical Journal (BMJ)] from 1 January 2017 to 31 May 2019 showing that 86 of the 100 phase III/IV randomized trials reviewed met established criteria for noninferiority. Since there is less quality control in pragmatic trials with little or no monitoring of adherence, they present further analytic challenges when used in noninferiority designs.
There is sometimes a perception that pragmatic superiority trials may be biased towards the null because of the inherent variability introduced by trying to mimic general clinical practice creating issues with separation of treatments, i.e., making treatments look more alike than dissimilar. The true adherence pattern in a clinical setting may deviate greatly from full adherence expected in an efficacy study and this nonadherence has the potential to dilute the treatment effect. To investigate this perception, we conducted a limited literature search of four top clinical journals (NEJM, JAMA, BMJ, Lancet) for randomized phase III pragmatic trials published in 2020. Of the 34 pragmatic trials identified, 22 (65%) had null findings and, in particular, all three of the noninferiority trials identified demonstrated evidence of noninferiority. This led us to conduct a more targeted search of the literature to further examine this preliminary finding of potential bias towards the alternative in pragmatic noninferiority trials. In this paper, we present the results of the literature search with a particular emphasis on the criteria for the noninferiority margin and the methods of analysis.

Methods

The search criteria are outlined below. A broad multi-purpose (mp) Ovid MEDLINE database search was performed that identified any publications in NEJM, JAMA, Lancet, or BMJ published between 2015 and 2021 (inclusive) that included the words “pragmatic” or comparative effectiveness” as well as “noninferiority” or “non-inferiority”. The .mp search includes the title, abstract, subject heading, keyword heading, plus other fields (Ovid MEDLINE Database Guide, 2022). Table 1 reports the number of publications that met individual search criteria. Our search identified 14 potential trials. After excluding two trials that did not examine noninferiority for the primary outcome, we were left with 12 randomized pragmatic noninferiority trials.
Results

Our search identified 12 randomized pragmatic noninferiority trials of which 11 were individually randomized trials and one was a cluster randomized trial. Ten of the 11 individually randomized trials met the criteria established for noninferiority as did the one cluster randomized trial. The key features of these trials are displayed in Table 2. Noninferiority margins were prespecified for all the trials. In six trials, the margin was based on either the minimum clinically important difference (MCID), clinical experts or consensus (The CODA Collaborative, 2020; Corcoran et al., 2021; Drekonja et al., 2021; Li et al., 2019; Scott et al., 2015; Leuchs et al., 2015). In one trial, sample size was a consideration – “to set the margin closer to the expected 25% difference in effectiveness would have required unrealistically large sample size. Therefore, NI margin set at 37% (25% expected difference + further 12%)” (Williams et al., 2017). One trial stated that “Due to the lack of evidence defining NI, the margin of 0.90 was chosen to inform future physician-patient discussions about medication reduction; under these assumptions, if NI was demonstrated, it would suggest that for every 10 patients who have their medications reduced, 9 would still have controlled blood pressure at 12-week follow-up.” (Sheppard et al., 2020). Another trial based the margin on empirical data – “NI margin was specified before trial initiation as an absolute difference of 1.25 percentage points (13% relative difference, which corresponds to a noninferiority margin odds ratio of 1.15) based on examination of the empirical distribution of hospital-level 30-day rates of death or serious complications, intracluster correlations, and power calculations” (Billimoria et al., 2016). In another trial the margin “was selected based on the perceived advantages of outpatient treatment. These advantages include avoiding a general anaesthetic, being treated immediately after diagnosis, taking only half a day’s leave from work, not having to arrange for childcare, and fewer hospital appointments” (Kortekangas et al., 2019). Finally, for two trials, no justification for the margin was provided (Homan et al., 2017; Ross et al., 2019).
The results of these trials are summarized in Figure 1. Plotted are the point estimates and confidence intervals for the primary and secondary analyses along with the noninferiority margin. Eleven trials specified the primary analysis as being according to intent to treat (ITT or as-randomized), modified ITT or random allocation and one trial conducted the primary analysis in the per-protocol population (i.e., described as as-treated in Drekonja et al., 2021). Most secondary analyses were specified as per protocol, although one trial performed an as-randomized analysis as its secondary analysis. A few trials also performed as-treated analyses (treatment actually received regardless of study arm assignment) and one performed an additional worst-case analysis. All analyses showed consistent findings. Further details about the analysis populations are also provided in Table 2.

Although all but one trial specified ITT or modified ITT as the primary analysis, participants were excluded from the analyses of many of these trials (Table 2). Only two individually randomized trials included all randomized participants in the primary analysis (Holman et al., 2017 and Li et al., 2019) of which the latter used multiple imputation to impute missing data. The one cluster randomized trial by Bilimoria et al (2016) included all but one of the 118 clusters. The percentage of participants excluded from the primary analyses ranged from about 2% to nearly 30% and sometimes differed between treatment arms. The reasons for exclusion were varied and included randomized in error (e.g., not eligible, screen failure), non-adherence to protocol assigned treatment (e.g., did not receive full dose), did not receive assigned treatment (e.g., drug not available, surgery postponed), death, withdrawal, lost to follow up, and no outcome data or incomplete data.

Discussion

Two issues in noninferiority trials are specification of the margin and method of analysis, and pragmatic trials can complicate these issues. While guidelines have been developed in the regulatory setting for specifying the margin (e.g., U.S. FDA and EMA), they are lacking for
pragmatic noninferiority trials. There is no disagreement that the margin needs to be pre-
specified and justified in the protocol development stage. This was the case for all but two of the
trials identified in our search. Reasons for justification were based on expert clinical
opinion/decision, MCID, prior data and sample size. Sample size can be an issue in
noninferiority trials because requirements are higher than superiority trials because of the often
small noninferiority margin. Thus, there is frequently a tradeoff between sample size and size of
the margin to make a trial feasible to conduct. However, wide margins can make it easier to
demonstrate noninferiority. There is no simple answer to setting the margin, it is study
dependent. We echo the comment by Mo et al (2020), “Despite the development of many
objective methods to justify the non-inferiority margin, its determination remains a contentious
issue and highly context specific.”

A second issue concerns the method of analysis. The usual approach to the analysis of
noninferiority trials, particularly in the regulatory setting, is to conduct both ITT and per protocol
analyses. EMA (2000) has stated that “In a non-inferiority trial, the full analysis set [ITT] and the
per protocol analysis set have equal importance and their use should lead to similar conclusions
for a robust interpretation.” Because of the issue of bias towards the alternative in non-inferiority
trials, a per protocol analysis has been recommended in addition to an ITT analysis, which could
be problematic if these two analyses differ. The U.S. FDA is noncommittal on which method of
analysis to use (FDA, 2016). Both FDA and EMA have pointed out the importance of quality
control in noninferiority trials and that poorly conducted trials could lead to spurious claims of
noninferiority. In a recent article, Mo et al (2020) discusses the problem with nonadherence in
noninferiority trials and its effect on ITT analysis. They indicate that nonadherence can dilute
treatment effects and bias results towards noninferiority. They also recommend “When
investigators anticipate substantial non-adherence (that is, ≥5%) in a non-inferiority trial, an
adjusted per protocol analysis should be planned for and performed either as primary or
supplementary analysis, depending on the primary questions of interest.” They also propose
some causal inference methods of analysis “in the presence of non-adherence driven by confounders,” such as inverse probability weighting and G estimation. However, pragmatic trials by their nature test interventions that are easy to implement by health care personnel that require little to no monitoring of adherence. Thus, there is less quality control in these trials raising issues about appropriate analytic methods.

All 12 trials in our survey conducted both ITT or modified ITT and per protocol types of analyses and reached similar conclusions. The traditional definition of ITT is to include all participants as randomized regardless of adherence to the protocol. A modified version of ITT is often used in which some exclusions are allowed, such as randomizations in error and those without any follow-up data. Only two individually randomized trials included all randomized participants in the primary analysis (Holman et al., 2017 and Li et al., 2019). The other trials excluded individuals from both the ITT and modified ITT analyses for a variety of reasons related to study eligibility, administration of the intervention, withdrawal, death and lost to follow-up. Since these exclusions do not preserve the randomization, there may be issues with type I error and interpretation of p-values. In addition, per protocol analyses are known to be biased, since reasons for exclusion could differ between treatment arms. Therefore, given these issues, what are appropriate methods of analysis for pragmatic noninferiority trials.

Using estimands may provide an approach to the analytic problem and are becoming commonplace in the regulatory setting. International Conference on Harmonization (ICH) E9(R1) defines them as “a precise description of the treatment effect reflecting the clinical question posed by a given clinical trial objective. It summarizes at a population level what the outcomes would be in the same patients under different treatment conditions being compared. The targets of estimation are to be defined in advance of a clinical trial.” Estimands have five components: the population defined by inclusion/exclusion criteria, treatment, endpoint, handling of intercurrent events (e.g., death, withdrawal, treatment discontinuation/switch) and population summary variable for the analysis (e.g., mean, proportion, hazard rate, etc.). ICH
E9(R1) defines intercurrent events as “events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest.” These could include death, withdrawal, treatment discontinuation/switch, etc. Missing data due to lost to follow-up are not considered an intercurrent event and are handled in the analysis using appropriate methods for handling of missing data (National Research Council, 2010). Various methods of analysis have been proposed for estimands including treatment policy (ITT), composite strategy (intercurrent event is incorporated in the outcome, e.g., stroke or death), while on treatment (events after treatment discontinuation are not counted), hypothetical (assumes intercurrent event did not occur) and principal stratification. Leuchs et al. (2015) provide some guidance on choosing estimands during the planning phase of clinical trials; however, the application of estimands to pragmatic noninferiority trials has not to our knowledge been addressed.

There are several limitations to this work. Our search was narrow and only included four top tier clinical journals over a five-year period. There is also the issue of publication bias – null trials may not have been published, i.e., those in which noninferiority was not demonstrated. Despite these limitations, we found that all but one of the pragmatic trials demonstrated noninferiority and that noninferiority was confirmed by both the primary ITT analysis and secondary per protocol analysis – there was no inconsistency in the findings. Mo et al. (2020) also found that a high percentage (86%) of the 100 phase III/IV randomized trials reviewed met established criteria for noninferiority, but their search was not confined to pragmatic trials. The majority (82%) of the trials used ITT as the primary analysis, 15% used per protocol as the primary and only three used both. In contrast, all of the trials we reviewed conducted both an ITT/modified ITT and per protocol analysis.
Conclusions

In summary, the interface between pragmatic and noninferiority trials is complex. Specification of margins and methods of analysis require careful consideration to prevent bias towards noninferiority. Much of the guidance on these two issues has been developed for regulatory environment and not for pragmatic noninferiority trials. Since many pragmatic trials follow the PRECIS criteria in which there is "unobtrusive (or no) measurement of participant [practitioner] compliance" (Thorpe et al., 2009), it affects separation of treatments which in turn affects both the setting of margins and analysis. More development is needed in these two areas.
List of abbreviations

AT – As treated (treatment actually received regardless of study arm assignment)
BMJ – British Medical Journal
CI – Confidence interval
CONSORT – Consolidated Standards of Reporting Trials
EMA – European Medicines Agency
FDA – U.S. Food and Drug Administration
ICH – International Conference on Harmonization
ITT – Intent to treat (or as-randomized)
JAMA – Journal of the American Medical Association
MCID – Minimally clinically important difference
mITT – Modified intent to treat
NEJM – New England Journal of Medicine
NI – Noninferiority
PP – Per protocol (according to allocated treatment)
Declarations

- Ethics approval and consent to participate: Not applicable
- Consent for publication: Not applicable
- Availability of data and materials: All data analyzed during this study are included in this published article.
- Competing interests: The authors declare that they have no competing interests
- Funding: Not applicable
- Authors' contributions: MC performed MEDLINE search, performed data analysis, prepared manuscript; JL assisted with coding; PP designed the research project, summarized studies and synthesized results, drafted manuscript.
- Acknowledgements: The authors would like to thank Kate Nyhan, MLS, IPI PMC, who provided valuable suggestions and guidance when searching for relevant articles.
References


Tables and Figures

Table 1. Ovid MEDLINE search strategy and number of articles identified. Search performed on 28 June 2022.

<table>
<thead>
<tr>
<th>Query Number</th>
<th>Query</th>
<th>Articles Retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“0255562”.jc and (“2015&quot; or “2016&quot; or “2017&quot; or “2018&quot; or “2019&quot; or “2020&quot; or “2021”).yr. [NEJM]</td>
<td>10,865</td>
</tr>
<tr>
<td>2</td>
<td>“7501160”.jc and (“2015” or “2016” or “2017” or “2018” or “2019” or “2020” or “2021”).yr. [JAMA]</td>
<td>10,962</td>
</tr>
<tr>
<td>3</td>
<td>“2985213R”.jc and (“2015” or “2016” or “2017” or “2018” or “2019” or “2020” or “2021”).yr. [Lancet]</td>
<td>11,332</td>
</tr>
<tr>
<td>4</td>
<td>“8900488”.jc and (“2015” or “2016” or “2017” or “2018” or “2019” or “2020” or “2021”).yr. [BMJ]</td>
<td>20,190</td>
</tr>
<tr>
<td>5</td>
<td>or/1-4</td>
<td>53,349</td>
</tr>
<tr>
<td>6</td>
<td>5 and (pragmatic or comparative effectiveness).mp.</td>
<td>266</td>
</tr>
<tr>
<td>7</td>
<td>5 and (noninferiority or non-inferiority).mp.</td>
<td>366</td>
</tr>
<tr>
<td>8</td>
<td>6 and 7</td>
<td>14</td>
</tr>
</tbody>
</table>

1 <term>.jc. National Library of Medicine (NLM) journal code is an alpha-numeric code that uniquely identifies every journal indexed in the NLM databases; 0255562 = NEJM, 7501160 = JAMA, 2985213R = Lancet, 8900488 = BMJ.

2 <term>.yr. Year of publication field that contains the four-digit year in which an article or monograph was published.

3 <term>.mp. Multi-purpose search. The fields searched by a .mp. include title, original title, abstract, subject heading, keywork heading, plus other fields. (Ovid MEDLINE Database Guide, 2022)
### Table 2. Key features of randomized pragmatic noninferiority trials*

<table>
<thead>
<tr>
<th>Study (Author, year)</th>
<th>Primary Outcome(s)</th>
<th>Basis for Noninferiority Margin</th>
<th>Analysis Populations (primary listed first)</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Randomized Trial Comparing Antibiotics with Appendectomy for Appendicitis (The CODA Collaborative, 2020)</td>
<td>30-day EQ-5D score (range from 0 to 1, with higher scores indicating better health status)</td>
<td>Difference of -0.05 points (antibiotics minus surgery) based on a minimum clinically important difference (MCID)</td>
<td>mITT framework (i.e., mITT): Primary analysis included all those who completed all items on the 30-day EQ-5D: 686/776 (88%) assigned to antibiotics vs. 664/776 (86%) assigned to appendectomy. PP analysis: For the group assigned to antibiotics, per-protocol included the use of an IDSA/SIS recommended antibiotic, a prescription for the full course, and, if an appendectomy was performed at index, that it was not reported as a protocol violation. For those assigned to appendectomy, per protocol was defined as adherence to the randomization assignment. (N not specified for this analysis.) Sensitivity analysis: Used multivariate imputation by chained equations (MICE) approach to impute missing data on the primary outcome. To inform the imputation, they used demographic and clinical factors collected at baseline and week that were associated with either odds of having missing data or with the primary outcome itself. mITT mean difference (antibiotics vs. surgery) = 0.01 points; 95% CI (−0.001, 0.03) PP mean difference = 0.01 points; 95% CI (−0.002, 0.03) Sensitivity analysis mean difference = 0.01 points; 95% CI (−0.004, 0.02)</td>
<td></td>
</tr>
<tr>
<td>Dexamethasone and Surgical-Site Infection (Corcoran et al., 2021)</td>
<td>Proportion with surgical-site infection within 30 days after surgery</td>
<td>2% difference (dexamethasone minus placebo) determined with the assistance of clinical experts and was set based on a modified Delphi process and an anticipated incidence of surgical-site infection of 9%</td>
<td>mITT: Primary analysis excluded people who did not undergo eligible surgery (surgery with a total incision length of &gt; 5 cm when the patient was under general anesthesia), who withdrew consent or whose clinician withdrew the patient from the trial. 4350/4444 (97.9%) assigned to dexamethasone vs. 4328/4436 (97.6%) assigned to placebo at a primary outcome at 30 days were analyzed in the modified ITT population. PP analysis: Included the same criteria as mITT and further excluded patients who did not receive dexamethasone or placebo or who received open label dexamethasone (or other glucocorticoid). 4074 (91.7%) dexamethasone vs. 4042 (91.1%) placebo were included in the per protocol population. AT: Per protocol population minus those who did not receive dexamethasone or placebo or received open label mITT risk difference (dexamethasone vs. placebo) = −0.9 percentage points; 95.6% CI (−2.1, 0.3); p&lt;0.001 for NI PP risk difference = −0.9; 95.6% CI (−2.1, 0.3); p&lt;0.001 for NI AT risk difference = 0.04; 95.6% CI (−1.2, 1.2); p=0.001 for NI</td>
<td></td>
</tr>
</tbody>
</table>

*Table adapted from the original source for readability and clarity.*
<table>
<thead>
<tr>
<th>Study (Author, year)</th>
<th>Primary Outcome(s)</th>
<th>Basis for Noninferiority Margin</th>
<th>Analysis Populations (primary listed first)</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Doxycycline versus Prednisolone as an Initial Treatment Strategy for Bullous Pemphigoid</strong> (Williams et al., 2017)</td>
<td>NI effectiveness: Proportion classed as treatment success (3 or fewer significant blisters) at 6 weeks, regardless of whether treatment had been modified because of a poor response</td>
<td>&quot;The smaller the NI margin, the greater the sample size, so to set the margin closer to the expected 25% difference in effectiveness would have required unrealistically large sample size. Therefore, NI margin set at 37% (25% expected difference + further 12%).&quot;</td>
<td>NI Effectiveness mITT: Fulfilled eligibility criteria and had data on primary outcome. 112/140 (80%) assigned doxycycline vs. 101/138 (73%) assigned prednisolone included in mITT analysis. PP population for the NI outcomes at weeks 3 and 6: Defined as those participants who for reasons other than treatment success or failure (determined by the investigators) had not increased their dose of allocated treatment, changed treatment or added a new treatment to their allocated treatment, used topical steroids between visit weeks 3 and 6 (week 6 outcome only), missed more than 3 consecutive treatment days, or committed other deviations deemed to be violations by an independent adjudicator masked to treatment allocation. 91/140 (65%) doxycycline and 78/138 (57%) prednisolone included in PP analysis. Safety mITT: 121 (86%) who started doxycycline vs. 113 (82%) who started prednisolone.</td>
<td>mITT adjusted difference in proportions (prednisolone vs. doxycycline) = 18.6%; 90% CI (11.1, 26.1) PP adjusted difference in proportions = 18.7%; 90% CI (9.8, 27.6)</td>
</tr>
<tr>
<td><strong>Effect of 7 vs 14 Days of Antibiotic Therapy on Resolution of Symptoms Among Afebrile Men with Urinary Tract Infection</strong> (Drekonja et al., 2021)</td>
<td>Resolution of UTI symptoms by day 14 after completion of active antibiotic treatment (assessed on day 7 in placebo group and day 14 in antibiotic group)</td>
<td>&quot;Four infectious disease physicians, all with multiple publications regarding UTI management, were asked to provide a value for a noninferiority margin they would consider acceptable for a proposed trial of afebrile men with UTI. Based on their responses (range, 10%-20%, with no process to achieve consensus), a noninferiority margin of 10% was selected...&quot;</td>
<td>AT: &quot;The primary noninferiority analysis compared the proportion of participants with resolution of UTI symptoms by day 14 after completion of active antimicrobial treatment, limited to the as-treated population (participants who took ≥26 of 28 doses and missed no more than 2 consecutive doses).&quot; 131/136 (96%) placebo and 123/136 (90%) antibiotic included in as-treated analysis. (The AT analysis in this context is equivalent to a PP analysis). As-randomized: &quot;A secondary, as-randomized analysis was performed that included all randomized participants analyzed in the groups to which they were randomized.&quot; 136/136 (100%) placebo and 136/136 (100%) antibiotic included in as-randomized analysis. (The as-randomized analysis is equivalent to an ITT analysis). At 7 days after treatment, UTI symptoms had resolved for 122/131 (93%) in the placebo group vs. 111/123 (90%) in the antibiotic group at 14 days after treatment. AT (equivalent to PP) absolute difference in proportions (placebo vs. antibiotic) = 2.9%; 97.5% 1-sided CI (–5.2, ∞) As-randomized absolute difference in proportions (placebo vs. antibiotic) = 1.5%; 97.5% 1-sided CI (–5.8, ∞)</td>
<td></td>
</tr>
<tr>
<td>Study (Author, year)</td>
<td>Primary Outcome(s)</td>
<td>Basis for Noninferiority Margin</td>
<td>Analysis Populations (primary listed first)</td>
<td>Results</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Effect of Antihypertensive Medication Reduction vs Usual Care on Short-term Blood Pressure Control in Patients with Hypertension Aged 80 Years and Older (Sheppard et al., 2020)</td>
<td>Systolic blood pressure lower than 150 mm Hg at 12-week follow-up</td>
<td>&quot;Prespecified NI margin was a relative risk (RR) of 0.90. Due to the lack of evidence defining NI, the margin of 0.90 was chosen to inform future physician-patient discussions about medication reduction; under these assumptions, if NI was demonstrated, it would suggest that for every 10 patients who have their meds reduced, 9 would still have controlled blood pressure at 12-week follow-up.&quot;</td>
<td>mITT: Primary analysis population was defined as all participants for whom data were available, and participants were analyzed according to groups to which they were randomly allocated, regardless of deviation from protocol. 265/282 (94%) assigned to medication reduction vs. 269/287 (94%) assigned to usual care include in mITT. PP analysis of the primary outcome: Excluded patients from intervention group who did not reduce treatment or who had medication reinstated during follow-up (although this latter action was part of the medication reduction protocol). 185 (66%) medication vs. 169 (59%) usual care included in PP.</td>
<td>mITT adjusted RR (medication reduction vs. usual care) = 0.98; 97.5% 1-sided CI (0.92, ∞)</td>
</tr>
<tr>
<td>Effects of Once-Weekly Exenatide on Cardiovascular Outcomes in Type 2 Diabetes (Holman et al., 2017)</td>
<td>Primary composite outcome = first occurrence of death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke</td>
<td>Coprimary hypotheses were that exenatide, administered once weekly, would be NI to placebo for the primary outcome, with a noninferiority margin of 1.3 for the upper limit of the two-sided 95% confidence interval of the hazard ratio. Primary efficacy hypothesis was that exenatide, administered once weekly, would be superior to placebo, with a superiority margin of less than 1.0 for the upper limit of the two-sided 95% confidence interval.</td>
<td>ITT: Included all patients who underwent randomization. 14752/14752 (100%) analyzed for primary outcome. Sensitivity analyses of the primary efficacy outcome were performed in the PP population, which included all randomized patients who received at least one dose of the trial regimen and had no major protocol violations. 14565/14752 (98.7%) included in PP.</td>
<td>A primary composite outcome event occurred in 839 of 7356 patients (11.4%; 3.7 events per 100 person-years) in the exenatide group and in 905 of 7396 patients (12.2%; 4.0 events per 100 person years) in the placebo group (ITT hazard ratio = 0.91; 95% CI (0.83, 1.00); p&lt;0.001 for NI), with the intention-to-treat analysis indicating that exenatide, administered once weekly, was noninferior to placebo with respect to safety (p=0.06 for superiority) PP hazard ratio = 0.95; 95% CI (0.85, 1.07); p&lt;0.001 for noninferiority, p=0.39 for superiority</td>
</tr>
<tr>
<td>Gentamicin Compared with Ceftriaxone for the Treatment of Gonorrhoea (G-ToG) (Ross et al., 2019)</td>
<td>Clearance of gonorrhoeae at all initially infected sites, defined as a negative nucleic acid amplification test 2 weeks post treatment</td>
<td>NI margin was a lower confidence limit of −5% for the risk difference.</td>
<td>mITT approach: Primary approach was according to randomized allocation without imputation of missing outcome data and included only those with primary outcome data. Primary outcome data were available for 306 (85%) of 362 assigned to ceftriaxone vs. 292 (82%) of 358 assigned to gentamicin. A variety of sensitivity analysis were done including using multiple imputation.</td>
<td>At 2 weeks after treatment, infection had cleared for 299/306 (98%) in the ceftriaxone group vs. 267/292 (91%) in the gentamicin group. mITT adjusted risk difference = −6.4%, 95% CI (−10.4%, −2.4%) NI was not met. Sensitivity analyses were consistent with the primary analysis.</td>
</tr>
<tr>
<td>National Cluster-Randomized Trial of Duty-Hour Flexibility in Surgical Training</td>
<td>30-day rate of postoperative death or serious complications</td>
<td>&quot;NI margin was specified before trial initiation as an absolute difference of 1.25 percentage points (13% relative difference, which</td>
<td>ITT approach: Analyzed 117 of 118 randomized programs, 138,691 residents. ITT unadjusted odds ratio for the flexible-policy group vs. standard policy = 0.96; 92% CI (0.87, 1.06); p=0.44</td>
<td></td>
</tr>
<tr>
<td>Study (Author, year)</td>
<td>Primary Outcome(s)</td>
<td>Basis for Noninferiority Margin</td>
<td>Analysis Populations (primary listed first)</td>
<td>Results</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------</td>
<td>-----------------------------------</td>
<td>---------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>(Bilimoria et al., 2016)</td>
<td>Primary Outcome(s)</td>
<td>Basis for Noninferiority Margin</td>
<td>Analysis Populations (primary listed first)</td>
<td>Results</td>
</tr>
<tr>
<td>Oral versus Intravenous Antibiotics for Bone and Joint Infection (Li et al., 2019)</td>
<td>Definitive treatment failure within 1 year after randomization</td>
<td>&quot;Originally, a 5-percentage point NI margin was based on a consensus among a wide range of researchers, infectious-disease specialists, and orthopedic surgeons and balanced the potential risks and benefits of oral therapy. In February 2015, after 601 participants had undergone randomization in the multicenter trial, a planned interim analysis showed an overall failure rate of approximately 12.5%. The original 5-percentage-point margin was consequently considered too restrictive and therefore, by consensus, the investigators adjusted the NI margin to 7.5 percentage points on the absolute risk difference scale (corresponding to a 60% relative difference) with approval from the trial steering committee, data and safety monitoring committee, and ethics committee.&quot;</td>
<td>Primary ITT analysis: Used multiple imputation by chained equations for missing end-point data. All 1024 randomized participants (527 per group) were included in the ITT analysis. mITT analysis: Included only the participants with complete endpoint data, available for 1015 (96.3%). PP analysis: Included only participants who received at least 4 weeks of their randomly assigned treatment. 889 (87%) included in PP analysis. For worst case sensitivity analyses, missing endpoint data were replaced by treatment failure in the oral group and success in the intravenous group.</td>
<td>ITT difference in proportion with definitive treatment failure (oral vs. IV) = −1.4; 90% CI (−4.9, 2.2) mITT difference in proportion = −1.5; 90% CI (−5.0, 2.1) PP difference in proportion = −2.5; 90% CI (−6.3, 1.3) Worst-case sensitivity analysis difference in proportion = 2.1; 90% CI (−1.5, 5.7)</td>
</tr>
<tr>
<td>Outpatient versus Inpatient Uterine Polyp Treatment for Abnormal Uterine Bleeding (Cooper et al., 2015)</td>
<td>Successful treatment, determined by the women's assessment of bleeding at 6 months</td>
<td>&quot;The prespecified NI margin of 25% (relative reduction) was based on the assumptions that outpatient polypectomy would be more convenient for women and cheaper, permitting it to be considered the treatment of choice even if fewer women had alleviation of bleeding symptoms.&quot; Limitation: &quot;The NI level of 25% might be considered large by some and hence a limitation of the study; however, this was selected based on the perceived advantages of&quot;</td>
<td>Primary analyses were specified as ITT; however, 228/254 (90%) assigned to outpatient vs. 211/253 (83%) assigned to inpatient were available for analysis at 6 months. PP analysis: Included only those women who received their allocated treatment.</td>
<td>73% (166/228) of women in the outpatient group and 80% (168/211) in the inpatient group reported successful treatment at 6 months. ITT relative risk = 0.91, 95% CI (0.82, 1.02) PP relative risk = 0.92; 95% CI (0.82, 1.02)</td>
</tr>
<tr>
<td>Study (Author, year)</td>
<td>Primary Outcome(s)</td>
<td>Basis for Noninferiority Margin</td>
<td>Analysis Populations (primary listed first)</td>
<td>Results</td>
</tr>
<tr>
<td>------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Three Week versus Six Week Immobilisation for Stable Weber B Type Ankle Fractures</td>
<td>Olerud-Molander Ankle Score at 12 months (OMAS; range 0-100; higher scores indicate better outcomes and fewer symptoms).</td>
<td>The predefined NI margin for the primary outcome was -8.8 points. In the absence of better evidence, we organised a focus group discussion among experts to define the appropriate estimate for non-inferiority margin. The panel reached a consensus that a 10% difference in 0-100 OMAS scale would not be clinically significant, which was then used to derive our non-inferiority margin (10% equals 8.8 points on the OMAS scale, Cohen’s d = 0.215, indicating a small effect size).</td>
<td>Analyses were primarily conducted according to ITT. All participants 234/247 (95%) randomized participants completed the study and were included in the primary ITT analysis (per supplement 2). There were 2 primary comparisons: 3-week cast vs. 6 week cast and 3-week orthosis vs. 6-week cast. PP analyses included 231/247 (94%) randomized (per supplement 2).</td>
<td>3-week cast vs. 6-week cast difference in OMAS = 3.6 points; 95% CI (−1.9, 9.1), p=0.20 3-week orthosis vs. 6-week cast difference in OMAS = 1.7 points; 95% CI (−4.0, 7.3), p=0.56 3-week cast vs. 6-week cast difference in OMAS = 4.3; 95% CI (−1.2, 9.8) 3-week orthosis vs. 6-week cast = 2.4; 95% CI (−3.3, 8.1)</td>
</tr>
<tr>
<td>Tumour Necrosis Factor Inhibitors versus Combination Intensive Therapy with Conventional Disease Modifying Anti-Rheumatic Drugs in Established Rheumatoid Arthritis (Scott et al., 2015)</td>
<td>Reduction in disability at 12 months measured with patient recorded health assessment questionnaire (range 0.00-3.00)</td>
<td>“0.22 non-inferiority margin for combination treatment versus the biologic strategy. The minimal clinically important change in scores on the health assessment questionnaire is 0.22.”</td>
<td>Randomized patients who received treatment were assessed on an ITT basis. 107 were randomized to TNF and 101 started taking it; 107 were randomized to the combined drug strategy and 104 started taking the drugs; 205/214 (96%) patients were treated and analyzed. Complete case analyses evaluated patients who followed the protocol and received 12 months’ treatment – 147 (69%) complete cases, 72 in disease modifying drug strategy vs. 75 TNF.</td>
<td>ITT mean difference (combined drugs vs. TNF) = −0.14; 95% CI (−0.29, 0.01) Complete case mean difference = −0.14; 95% CI (−0.32, 0.03)</td>
</tr>
</tbody>
</table>

*AT = as treated (treatment actually received regardless of study arm assignment), CI = confidence interval, ITT = intent to treat (or as-randomized), mITT = modified intent to treat, NI = noninferiority, PP = per protocol (according to allocated treatment)
Figure 1. Reported treatment effects and confidence intervals

- **Antibiotics vs. Appendectomy for Appendicitis**
- **Dexamethasone and Surgical Site Infection**
- **Doxycycline vs. Prednisolone for Bullous Pemphigoid**
- **7 vs. 14 Days of Antibiotic Therapy for UTI**
(1) Per-protocol analysis denoted as “as treated” in manuscript

(2) ITT analysis denoted as “as randomized” in manuscript

(3) Excluding those without full required samples taken at baseline

(4) Excluding those who did not take trial medication

(5) Excluding those without positive samples at baseline

(6) Three week cast versus six week cast

(7) Three week orthosis versus six week cast
Supplementary Files

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

- EQUATORNetworkReportingChecklist20220707.pdf