The difference between statistical significance and clinical relevance: the case of minimal important change, non-inferiority trials, and smallest worthwhile effect

Willigenburg, N.W.; Poolman, R.W.

Injury 54 (2023) 110764

Contents lists available at ScienceDirect

Injury

journal homepage: www.elsevier.com/locate/injury

The difference between statistical significance and clinical relevance. The case of minimal important change, non-inferiority trials, and smallest worthwhile effect☆

Nienke W. Willigenburg[a], Rudolf W. Poolman[a,b,*]

[a] Joint Research, Department of Orthopaedic Surgery, OLVG Hospital, Amsterdam, the Netherlands
[b] Department of Orthopaedic Surgery, Leiden University Medical Center, Leiden, the Netherlands

ARTICLE INFO

Keywords:
Clinical relevance
Statistical significance
Non-inferiority
Minimal important change
Minimal important difference
Smallest worthwhile effect

ABSTRACT

Clinical relevance and statistical significance are different concepts, linked via the sample size calculation. Threshold values for detecting a minimal important change over time are frequently (mis)interpreted as a threshold for the clinical relevance of a difference between groups. The magnitude of a difference between groups that is considered clinically relevant directly impacts the sample size calculation, and thereby the statistical significance in clinical study outcomes. Especially in non-inferiority trials the threshold for clinical relevance, i.e. the predefined margin for non-inferiority, is a crucial choice. A truly inferior treatment will be accepted as non-inferior when this margin is chosen too large. The magnitude of a clinically relevant difference between groups should be carefully considered, by determining the smallest effect for each specific study that is considered worthwhile. This means taking into account the (dis)advantages of both study interventions in terms of benefits, harms, costs, and potential side effects. This article clarifies common sources of confusion, illustrates the implications for clinical research with an example and provides specific suggestions to improve the design and interpretation of clinical research.

Background

Outcome measurement in clinical populations is often used to quantify health status, pain, and physical function in clinical settings or research settings. When patients quantify their outcome, for instance by completing a validated questionnaire, the resulting score is a patient-reported outcome. While measuring patient-reported outcomes is common, substantial challenges arise when interpreting these scores. For instance, what does a change of 12 points on a scale from 0 to 100 mean? Is it a ‘real’ change? Or could it be attributed to measurement error? And if the change is real, is it also important? These are relevant questions for patients, clinicians, and researchers.

In research, measuring outcomes is frequently used to compare outcomes between groups. Which of two (or more) treatment groups shows better improvement? Or ends up with a better absolute score? And what magnitude of a difference between treatment groups is clinically relevant? With the growing emphasis on evidence-based treatment, we need to know when differences in outcomes between groups of patients are real and important.

Definitions

An essential distinction to make is between a change and a difference. A *change* is defined as a change over time, *within* (groups of) patients. For instance, when the score was 6 on a 10-point pain scale before treatment and 3 after treatment, the *change* is 3 points. A *difference* is defined as a difference *between* (groups of) patients at a predefined time point. For instance, when an intervention group scores 6 on a 10-point pain scale after treatment and a control group scores 3 after placebo treatment, the *difference* is 3 points. Unfortunately, no consensus exists on the use of the terms change and difference. This lack of consensus is a widespread source of confusion and errors in methodological and clinical research.

This confusion can be attributed to the fact that the distinction between change and difference is not explicitly existent in the language of statistics. To detect a *change* in score after treatment, we statistically

☆ This paper is part of a Supplement supported by the Osteosynthesis and Trauma Care Foundation (OTCF) through a research grant from Stryker.

* Corresponding author at: OLVG, locatie Oost, Secretariaat Orthopedie, Postbus 95500, 1090 HM Amsterdam, the Netherlands.
E-mail address: r.w.poolman@lumc.nl (R.W. Poolman).

https://doi.org/10.1016/j.injury.2023.04.051
Accepted 28 April 2023

assess the *difference* between a pre-test score and a post-test score. The only statistical distinction between detecting a *change* over time and a *difference* between groups is the choice of *t*-test. A *change* over time is typically evaluated with a paired *t*-test, while a *difference* between groups is typically evaluated with an independent *t*-test.

To assess the *clinical relevance* of a *change* over time, it is essential to consider how reliable a measurement instrument can measure change, and what amount of change patients can actually feel. These concepts can be quantified as the Minimal *Detectable* Change (MDC) and Minimal *Important* Change (MIC). There needs to be more clarity about these concepts, which was nicely summarized in a recent conceptual clarification and systematic review [1].

The *MDC* reflects the smallest change in score that we can distinguish from measurement error (with 95% confidence) in individual patients. The MDC value should be measured in persons who have not changed in a test-retest reliability design. The statistical parameters to describe the MDC include the limits of agreement and the standard error of measurement. De Vet and colleagues (2006) provided a detailed discussion of this concept, including the mathematical equation [2]. The MDC is also called the Smallest Real Difference (SRD) or the Smallest Detectable Change (SDC).

The *MIC* is the smallest change in score that patients (on average) consider important. The MIC is not a measurement property but a parameter of interpretability. The MIC can be measured in a longitudinal design in persons who have changed, which requires an anchor question of perceived change. Terwee and colleagues formulated specific recommendations for calculating the MIC, including 1) a minimal correlation of 0.30 between the anchor question and the patient-reported outcome of interest; 2) the predictive modeling and receiver operating characteristic (ROC) method are preferred over the mean change method; and 3) when longitudinal data is not available, a vignette-based method can be used [1]. They also explain different methods to calculate the MIC, including references for readers who are interested in the underlying mathematical equations.

Following the terminology explained above, the clinical relevance of a *difference between groups* should be quantified as a Minimal Important Difference (MID). Unfortunately, the abbreviations MID and MCID (Minimal *Clinically* Important Difference) are frequently used as a synonym for the MIC [e.g. 3,4]. Consequently, threshold values based on perceived changes over time are often interpreted as thresholds for clinically relevant differences between groups. For instance, an instrument was recently developed to evaluate the credibility of anchor-based estimates of 'MIDs' [3], which are actually MICs according to the definition above. Application of this instrument to 585 studies reporting 5324 thresholds for a clinically relevant change for 526 distinct PROMs demonstrated severe credibility issues that hamper the interpretation of published threshold values [4]. Apparently, clinical and methodological researchers neither agree on the methods to determine thresholds for clinical relevance nor on the values of these thresholds. This is a threat for clinical research that warrants serious attention.

A concept that may be part of the solution to this problem is the *smallest worthwhile effect* (SWE). The idea with the SWE is that the clinical relevance of a difference between groups not only depends on the magnitude of the difference in a specific outcome, but also takes into account the costs, potential side effects and inconveniences associated with both interventions. The SWE value should be intervention-specific, formulated in terms of differences in outcomes with vs. without the intervention, and based on patients' perceptions. Ferreira and colleagues recommended a benefit-harm trade-off method for estimating the SWE [5]. Interestingly, such a benefit-harm trade-off method resulted in SWE estimates that did not differ across musculoskeletal pain sites in a population of patients referred for primary care physiotherapy [6]. Compared with natural recovery, people with neck, shoulder, and low back pain considered an additional improvement of 20% in disability and pain worthwhile, given the costs, potential side effects and inconveniences associated with physiotherapy. While relatively consistent across pain sites, SWE values were affected by age, work status, and use of pain medication. These are relevant findings for designing and interpreting studies that compare physiotherapy with no treatment in these populations. For other populations, interventions, and outcomes, SWE values are not so well established. Therefore, many clinical studies lack a solid basis in their design: which difference between groups really matters?

Isn't this why we do statistics?

A substantial proportion of clinicians and researchers is not sufficiently aware of the critical difference between statistical significance and clinical relevance. The majority of (bio)medical education programs include basic statistical teaching, which leaves students with the impression that a p-value below 0.05 demonstrates an effect (i.e. a change or a difference) that is statistically significant and (apparently also) clinically relevant. But this is not necessarily the case. Statistical p-values are valuable in discriminating between differences that likely occurred by chance and differences that likely reflect a 'true' difference. As mentioned earlier, a statistically significant effect can reflect a change (within groups of patients over time) or a difference (between groups). A combination of both (i.e. evaluating whether a between-group difference changed over time) can be evaluated by testing the *interaction effect* of time and group. Statistical analysis in itself does not say anything about the clinical relevance of an observed effect. If a study population is small, a clinically relevant difference can easily fail to reach statistical significance. And if a study population is large, (very) small differences can be detected as statistically significant, without being clinically relevant. This brings us to the methodological consideration that connects the concepts of clinical relevance and statistical significance: the sample size calculation.

Sample size calculation

When designing a prospective study, researchers need to perform a sample size calculation (also referred to as a power analysis), to estimate how many participants are needed. Two crucial assumptions determine the required sample size: 1) the standard deviation as a measure of variation within the study population, and 2) the magnitude of a between-groups difference that is considered clinically relevant.

The sample size calculation provides the minimum number of participants per group needed for a clinically relevant between-group difference to reach statistical significance. But both assumptions are challenging to estimate before the start of a new study reliably. Values are often derived from previous publications, but they vary widely. The standard deviation for instance typically depends on the size of a population. So when the population size is what you are trying to determine, it is questionable how to use the standard deviation as input for that calculation. Moreover, as stated above, the threshold for the clinical relevance of a between-group difference is frequently unknown. Consequently, reported values for *minimal detectable change* and *minimal important change* are often used as input. This is a prevalent practice, but it is highly questionable. These MDC and MIC values are based on changes within subjects over time rather than on differences between groups. And a rationale why the threshold for the clinical relevance of within-subject change would be the same as that for a between-group difference is lacking. Although literature and reference values for SWE are currently scarce, the smallest worthwhile effect seems a valuable alternative.

The case of non-inferiority trials

A non-inferiority trial aims to determine whether a new treatment is *not worse* than a reference treatment by more than an *acceptable* amount [7]. Because proof of exact equivalence is impossible, one could define a *margin of non-inferiority* for the treatment effect in a primary outcome

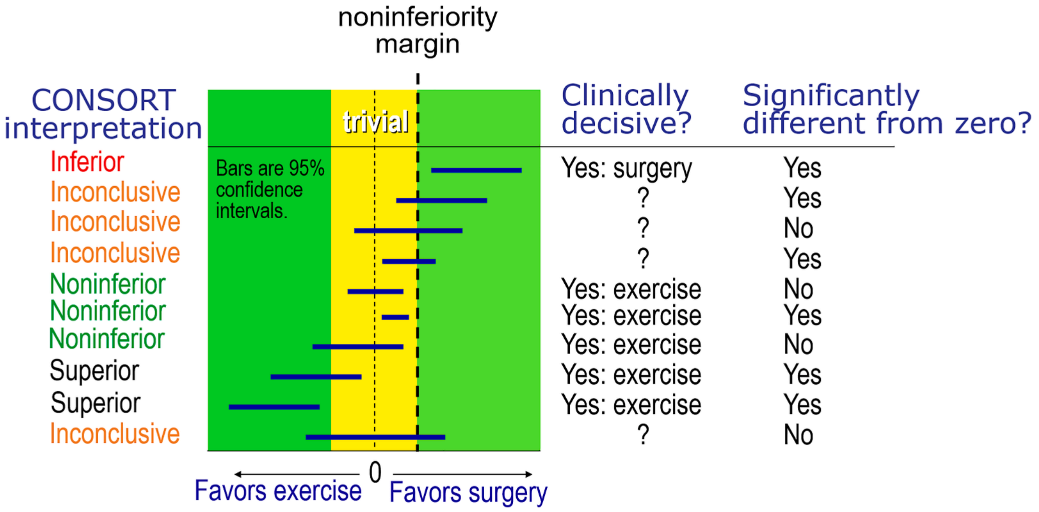

Fig. 1. Typical margin for non-inferiority (vertical dotted line) and 10 potential study outcomes as represented by 95% confidence intervals on an arbitrary scale (horizontal lines). Whether or not these 95% confidence intervals include or exclude the non-inferiority margin determines the interpretation based on the CONSORT guidelines and the clinical implication. The last column shows the difference in interpretation compared to superiority trials, in which statistical analyses test whether the 95% confidence intervals include or exclude the value 0.

that is relevant for patients. Non-inferiority of a new intervention with respect to the control treatment is particularly interesting when the new treatment has other advantages, i.e. better availability, lower costs, less invasiveness, lower risks of adverse events, or greater ease of administration. The margin of non-inferiority should be justified by a clinical rationale and be defined ‘a priori’. During data analysis, the observed difference between groups is not tested against the value zero but against the value of the predefined threshold for non-inferiority. Therefore, the interpretation of a non-inferiority analysis typically deviates from the interpretation of a superiority analysis. Fig. 1 shows 10 potential study outcomes and their interpretation with respect to the non-inferiority margin and the value zero. The control group in this generic example underwent surgical treatment and the intervention group participated in an exercise program.

If the non-inferiority margin is too large, a truly inferior treatment could be accepted as non-inferior. On the other hand, a very small non-inferiority margin is likely to yield inconclusive results, and requires an extremely large sample size. Fig. 2 visualizes potential study outcomes for the same generic study example, and interpretations with respect to a large (upper panel) and small (lower panel) margin for non-inferiority. A series of consecutive non-inferiority trials increases the risk of accepting a truly inferior treatment. For instance when treatment B is non-inferior to (just slightly worse than) treatment A, and treatment C in turn is non-inferior to (just slightly worse than) treatment B, the difference between treatment C and the original treatment A may actually be clinically relevant.

A systematic review of methods of defining the non-inferiority margin in randomized double-blind controlled trials included 273 studies and concluded that these methods are poorly reported and that this information is critical to allow for better judgment of non-inferiority trial results [8].

Example: the ESCAPE trial

Design and primary outcome

The ESCAPE trial aimed to answer the research question ‘Is physical therapy non-inferior to early surgery with arthroscopic partial meniscectomy for improving knee function among patients with non-obstructive meniscal tears?’ [9]. Because physical therapy has other substantial advantages over surgery in terms of costs, invasiveness and side effects, the non-inferiority design was considered adequate. Nine Dutch hospitals participated and a total of 321 patients were randomized to exercise therapy ($n = 162$) or surgery ($n = 159$). The primary outcome was the change in patient-reported knee function on the subjective knee form of the International Knee Documentation Committee (IKDC) from baseline over 24 months follow up.

A priori sample size calculation

When designing the ESCAPE non-inferiority trial [10] no estimate was available of the minimal clinically important difference for the IKDC in a population of patients with meniscal tears. Therefore, the non-inferiority margin was defined as the smallest detectable change of 8.8 points [11], rounded down to a margin of 8 points.

MIC in ESCAPE study population

During the study, the MIC was calculated in the ESCAPE study population, using an anchor-based MIC distribution method [12]. The receiver operating characteristics (ROC) curve indicated that a threshold value of 10.9 points on the IKDC best discriminated between patients who reported improvement ($n = 217$) and patients who reported no change ($n = 48$) on the external anchor question.

Interpretation of study results

Applying the MIC as determined in the ESCAPE study population as the margin for non-inferiority changes the interpretation of the results at 12 and 24 months. Based on the a priori set non-inferiority margin of 8, the outcome at these time points was inconclusive (Fig. 3, upper panel). When applying the non-inferiority margin based on the MIC in the ESCAPE trial population, these interpretations change to ‘non-inferior’ (Fig. 3, lower panel). So, which of these conclusions is ‘true’? Is any of these two values the ‘correct’ margin for non-inferiority? Who decides that and how? Which considerations are needed to determine why a between-group difference of for instance 3 points or 9 points would be relevant or not?

Discussion

Statistical significance and clinical relevance are two very different things, and both are important in clinical research. Complex considerations that underlie the sample size calculation strongly affect the conclusions of clinical studies. This is specifically the case in non-inferiority trials, where the magnitude of the observed difference between groups is tested against the non-inferiority margin rather than against zero. But also in superiority trials, because the sample size needed to detect a statistically significant difference directly depends on the assumption of the threshold for clinical relevance that is used as input for the sample

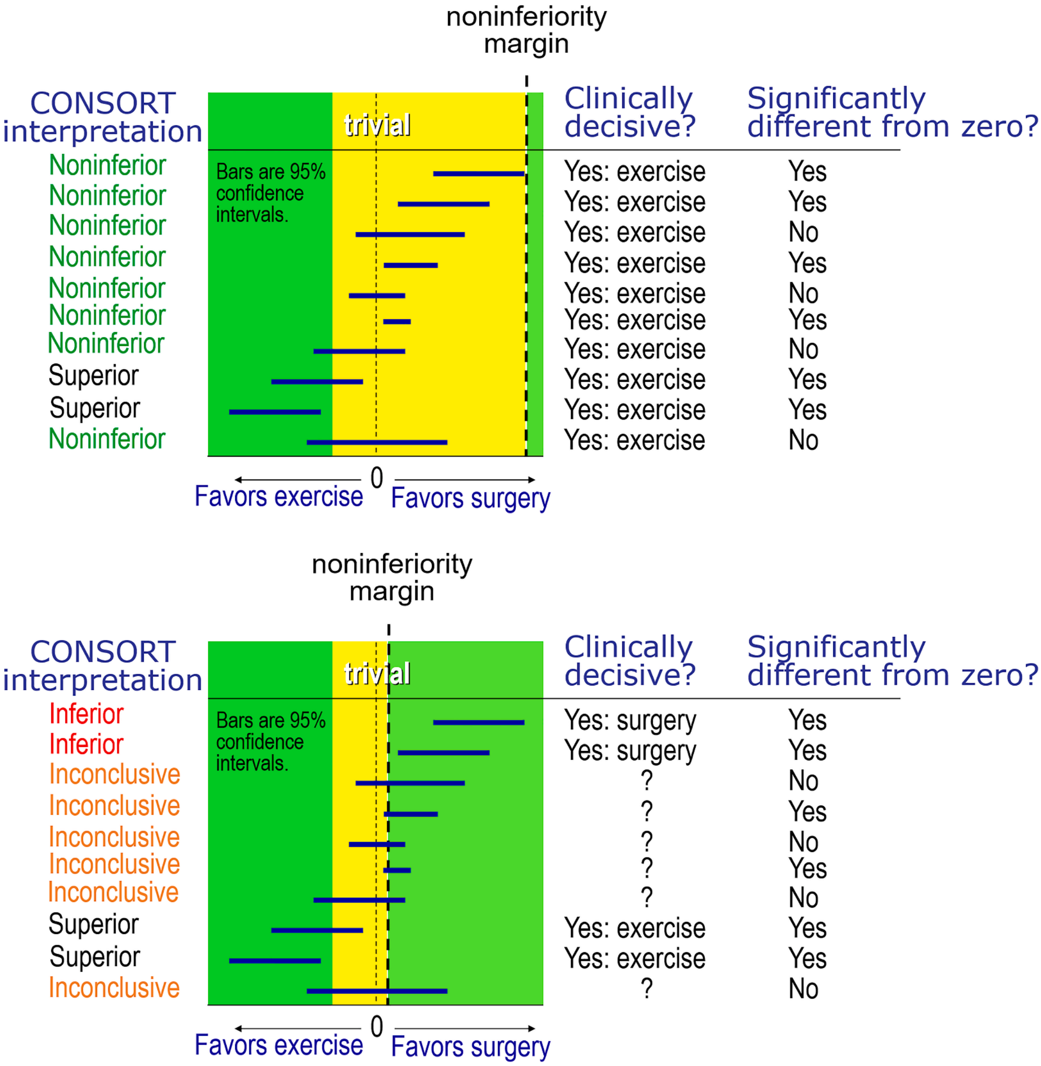

Fig. 2. Consequences for interpretation of the same 10 potential study outcomes, when the margin for non-inferiority is defined as larger (upper panel) or smaller (lower panel).

size calculation.

Threshold values for the clinical relevance of a within-group change over time are frequently used as a threshold for a between-group difference. Not only because many people do not understand the difference, but also because threshold values for clinical relevance of between group differences are difficult to quantify. On the one hand, one could argue that a difference between groups may not be clinically relevant if it cannot even be perceived as a change by an individual. On the other hand, if two treatments are very similar in terms of risks, invasiveness, and costs, one could argue that any small difference in outcome between treatment groups may be relevant, regardless of whether it can be perceived as a change over time by individuals. The additional risk with non-inferiority trials is that repeated non-inferiority trials can result in acceptance of a treatment that is actually inferior to the original treatment [13]. This disadvantage of repeated non-inferiority trials could be overcome by designing superiority trials instead, with a sample size calculation that is based on a thoughtfully established threshold value for clinical relevance that takes into account the potential risks and benefits of the interventions to be compared. If the apparent disadvantages of one of the treatments (i.e. higher costs, more serious risks or side effects) are already accounted for in the SWE, it could be argued that a well-powered superiority trial (using that SWE as threshold for clinical relevance in the sample size calculation), may be equally valuable for comparing the effectiveness of two treatments.

The example based on the ESCAPE non-inferiority trial illustrated how methodological choices affect conclusions. Interestingly, many journal editors and reviewers are critical on the presence of a sample size calculation in clinical research, but the choices and assumptions used as input for such a power analysis are hardly ever questioned. Moreover, these choices and assumptions are not always taken into consideration when interpreting study results. For instance, two randomized clinical trials comparing surgery with cast treatment in older adults with a distal radius fracture [14,15] had a very similar non-inferiority design. Both trials had the Patient-Reported Wrist Evaluation (PRWE) questionnaire as primary outcome and used a 14 point difference as a priori threshold for clinical relevance. This was referred to as a minimal clinically important *difference* (MCID), but the reported reference shows that this value was calculated as a minimal important *change* (MIC) [16]. Results from both trials indicated a difference between groups that was smaller than the predefined threshold for non-inferiority. Nevertheless, one trial concluded that surgery was better than cast treatment [14], while the other concluded that cast treatment was non-inferior [15].

How can we do better?

First, researchers and clinicians need to become aware of these

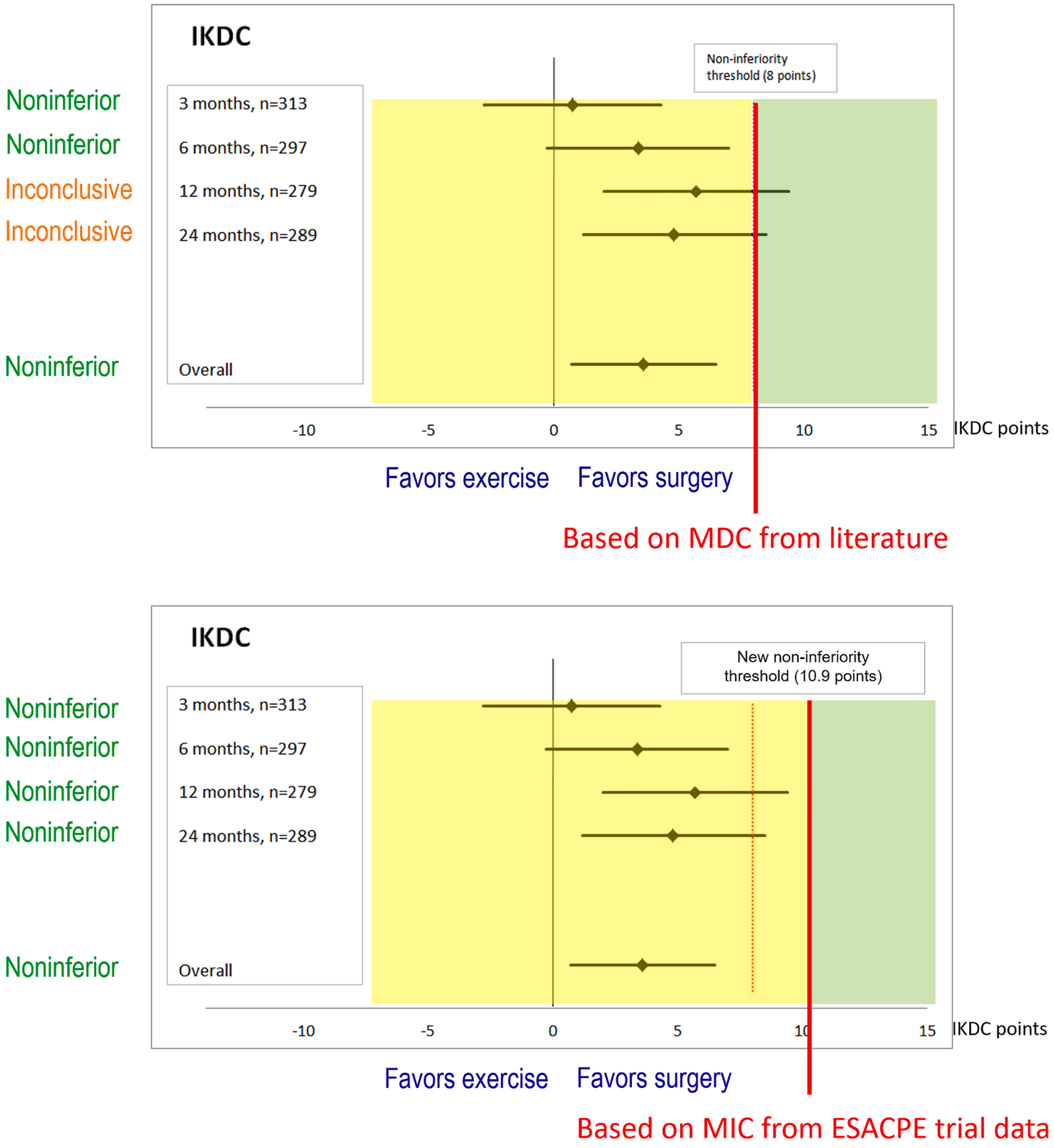

Fig. 3. The choice of non-inferiority threshold changes the interpretation of the ESCAPE trial results at 12 and 24 months.

issues, which are widespread in current literature. Second, we need an adequate and feasible methodology to quantify the relevance of a difference between groups. This should not (solely) depend on whether an individual can perceive a certain amount of change over time. One of the main determinants of the relevance of a between-group difference is likely the nature of the interventions that are to be compared. Characteristics of the population, such as the baseline score and expectations regarding the outcome of interest, will also affect the threshold for clinical relevance of a difference between two treatments.

To obtain a good estimate of the threshold for a clinically relevant effect, each clinical study should be preceded by a project to determine a study-specific SWE value. A benefit-harm trade-off project in patients undergoing total knee arthroplasty resulted in a smallest worthwhile effect on the Knee Osteoarthritis Outcome Score (KOOS) that was substantially larger than previously reported MCID and MIC values [17]. This impacts both the design and the interpretation of clinical studies. Specifically, with a larger threshold for clinical relevance, a smaller sample size is needed to detect a clinically relevant difference. Previous studies with sample sizes based on smaller MIC/MCID values could thus be considered 'overpowered' and detect differences that reach statistical significance but are not clinically relevant.

In addition to the benefit-harm trade-off methodology, SWE values can also be determined in a discrete choice experiment. Discrete choice methods give patients several hypothetical choices and the probability that an intervention will be considered worthwhile is calculated based on these choices. Ideally, both methods could be combined as nicely described in a study protocol for a fall prevention program for older adults [18]. Interestingly, 50% of the participants with the benefit-harm trade-off method and 82% with the discrete choice experiment did not consider the proposed exercise program worthwhile, even if it reduced their risk of falling to 0% [19].

Determining intervention-specific and population-specific SWE values is time-intensive and therefore costly. And even if time and costs would be no issue, it is not always feasible. The SWE is a difficult concept to explain, so not all study populations will be capable of participating in such a project. And quantifying the SWE with other stakeholders is not recommended, because the benefits, costs and harms are truly only experienced by the healthcare consumers [20]. However, the fall prevention example above illustrates that healthcare consumers may have very different ideas of worthwhile effects than other stakeholders.

Especially with the aging population and growing scarcity of healthcare resources, the question whether an intervention is worthwhile should also be asked from a societal perspective. Ideally, a consensus would be reached among representatives of different stakeholders (i.e. patients, clinicians, researchers, and policymakers). We therefore encourage all stakeholders, including funders for clinical research, to invest in this crucial topic. Only with widely supported threshold values for clinical relevance is it possible to adequately design, interpret, and grade clinical studies; and to successfully implement their results.

Conclusion

Clinical relevance and statistical significance are two different concepts, linked via the sample size calculation. Thresholds for the clinical relevance of a change over time are often (mis)interpreted as thresholds for the clinical relevance of a difference between groups. The clinical relevance of a difference between groups should not (only) be defined by a threshold of change that individual patients can perceive. Instead, it should incorporate potential harms, benefits, costs and side effects of the interventions of interest. Including an anchor question in clinical studies allows calculation of the MIC in the study population. Conducting an SWE study before or alongside a clinical study may help to interpret the subsequent results and reflect on their clinical relevance in the study population.

Declaration of Competing Interest

None.

References

[1] Terwee CB, Peipert JD, Chapman R, Lai JS, Terluin B, Cella D, et al. Minimal important change (MIC): a conceptual clarification and systematic review of MIC estimates of PROMIS measures. Qual Life Res 2021 Oct;30(10):2729–54. https://doi.org/10.1007/s11136-021-02925-y. Epub 2021 Jul 10. PMID: 34247326; PMCID: PMC8481206.

[2] de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006 Aug 22; 4:54. https://doi.org/10.1186/1477-7525-4-54. PMID: 16925807.

[3] Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020 Jun 4;369:m1714. https://doi.org/10.1136/bmj.m1714. PMID: 32499297; PMCID: PMC7270853.

[4] Carrasco-Labra A, Devji T, Qasim A, Phillips MR, Wang Y, Johnston BC, et al. Minimal important difference estimates for patient-reported outcomes: A systematic survey. J Clin Epidemiol 2021 May;133:61–71. https://doi.org/10.1016/j.jclinepi.2020.11.024. Epub 2020 Dec 13. PMID: 33321175.

[5] Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, Smeets RJ. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol 2012 Mar;65(3):253–61. https://doi.org/10.1016/j.jclinepi.2011.06.018. Epub 2011 Oct 19. PMID: 22014888.

[6] Christiansen DH, de Vos Andersen NB, Poulsen PH, Ostelo RW. The smallest worthwhile effect of primary care physiotherapy did not differ across musculoskeletal pain sites. J Clin Epidemiol 2018 Sep;101:44–52. https://doi.org/10.1016/j.jclinepi.2018.05.019. Epub 2018 May 29. PMID: 29852251.

[7] Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012 Dec 26;308(24):2594–604. https://doi.org/10.1001/jama.2012.87802. PMID: 23268518.

[8] Althunian TA, de Boer A, Klungel OH, Insani WN, Groenwold RH. Methods of defining the non-inferiority margin in randomized, double-blind controlled trials: a systematic review. Trials. 2017 Mar 7;18(1):107. https://doi.org/10.1186/s13063-017-1859-x. PMID: 28270184; PMCID: PMC5341347.

[9] van de Graaf VA, Noorduyn JCA, Willigenburg NW, Butter IK, de Gast A, Mol BW, et al. Effect of Early Surgery vs Physical Therapy on Knee Function Among Patients With Nonobstructive Meniscal Tears: The ESCAPE Randomized Clinical Trial. JAMA 2018 Oct 2;320(13):1328–37. https://doi.org/10.1001/jama.2018.13308. PMID: 30285177; PMCID: PMC6583004.

[10] van de Graaf VA, Scholtes VA, Wolterbeek N, Noorduyn JC, Neeter C, van Tulder MW, et al. Cost-effectiveness of Early Surgery versus Conservative Treatment with Optional Delayed Meniscectomy for Patients over 45 years with non-obstructive meniscal tears (ESCAPE study): protocol of a randomised controlled trial. BMJ Open 2016 Dec 21;6(12):e014381. https://doi.org/10.1136/bmjopen-2016-014381. PMID: 28003302; PMCID: PMC5223724.

[11] Crawford K, Briggs KK, Rodkey WG, Steadman JR. Reliability, validity, and responsiveness of the IKDC score for meniscus injuries of the knee. Arthroscopy 2007 Aug;23(8):839–44. https://doi.org/10.1016/j.arthro.2007.02.005. PMID: 17681205.

[12] Noorduyn JCA, van de Graaf VA, Mokkink LB, Willigenburg NW, Poolman RW, Research Group ESCAPE. Responsiveness and Minimal Important Change of the IKDC of Middle-Aged and Older Patients With a Meniscal Tear. Am J Sports Med 2019 Feb;47(2):364–71. https://doi.org/10.1177/0363546518812880. Epub 2019 Jan 4. PMID: 30608864.

[13] Fleming TR. Current issues in non-inferiority trials. Stat Med 2008 Feb 10;27(3): 317–32. https://doi.org/10.1002/sim.2855. PMID: 17340597.

[14] Martinez-Mendez D, Lizaur-Utrilla A, de-Juan-Herrero J. Intra-articular distal radius fractures in elderly patients: a randomized prospective study of casting versus volar plating. J Hand Surg Eur 2018 Feb;43(2):142–7. https://doi.org/10.1177/1753193417727139. Vol.Epub 2017 Sep 4. PMID: 28870129.

[15] CROSSFIRE Study Group Lawson A, Naylor JM, Buchbinder R, Ivers R, Balogh ZJ, et al. Surgical Plating vs Closed Reduction for Fractures in the Distal Radius in Older Patients: A Randomized Clinical Trial. JAMA Surg 2021 Mar 1;156(3): 229–37. https://doi.org/10.1001/jamasurg.2020.5672. PMID: 33439250; PMCID: PMC7807386.

[16] Sorensen AA, Howard D, Tan WH, Ketchersid J, Calfee RP. Minimal clinically important differences of 3 patient-rated outcomes instruments. J Hand Surg Am 2013 Apr;38(4). https://doi.org/10.1016/j.jhsa.2012.12.032. 641-9Epub 2013 Mar 6. PMID: 23481405; PMCID: PMC3640345.

[17] Henderson N, Riddle DL. The smallest worthwhile effect is superior to the MCID for estimating acceptable benefits of knee arthroplasty. J Clin Epidemiol 2022 Dec; 152:201–8. https://doi.org/10.1016/j.jclinepi.2022.10.019. Epub 2022 Oct 28. PMID: 36404574.

[18] Franco MR, Ferreira ML, Howard K, Sherrington C, Rose J, Haines TP, Ferreira P. How big does the effect of an intervention have to be? Application of two novel methods to determine the smallest worthwhile effect of a fall prevention programme: a study protocol. BMJ Open 2013 Feb 5;3(2):e002355. https://doi.org/10.1136/bmjopen-2012-002355. PMID: 23388197; PMCID: PMC3586108.

[19] Franco MR, Howard K, Sherrington C, Rose J, Ferreira PH, Ferreira ML. Smallest worthwhile effect of exercise programs to prevent falls among older people: estimates from benefit-harm trade-off and discrete choice methods. Age Ageing 2016 Nov;45(6):806–12. https://doi.org/10.1093/ageing/afw110. Epub 2016 Jun 27. PMID: 27496928.

[20] Ferreira M. Research Note: The smallest worthwhile effect of a health intervention. J Physiother 2018 Oct;64(4):272–4. https://doi.org/10.1016/j.jphys.2018.07.008. Epub 2018 Sep 3. PMID: 30190218.