For better. . .or at least not worse

Papageorgiou, Spyridon N

DOI: https://doi.org/10.1177/14653125221136150

For better…or at least not worse

Spyridon N. Papageorgiou

Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Zurich, Switzerland

Corresponding author: Spyridon N. Papageorgiou, Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Plattenstrasse 11, Zurich, CH 8032, Switzerland. Email: snpapage@gmail.com

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Spyridon N. Papageorgiou https://orcid.org/0000-0003-1968-3326

Words in text: 1670

Blinded Manuscript

Theoretical scenario

In this theoretical scenario, a comparison of two different treatment approaches is made in terms of treatment duration. This could well also be a comparison between a gold standard treatment (treatment with traditional fixed appliances) and treatment with an additional use of an adjunct surgical procedure like corticotomy, piezocision, micro-osteoperforation, etc—this makes no difference for this piece.

Practically, a randomized clinical trial of adequate methodological robustness is assumed to have run and a clinical decision is to be made based on the trial's results. The absolute numbers for, let's say, the duration of the tooth alignment phase are not important, but rather how the two treatment approaches fare in relative terms. Based on the measured outcome data of the two treatment groups, a treatment effect estimate (here a difference in mean durations) with a measure of its imprecision (here a two-sided 95% confidence interval) should be calculated and reported by the authors of the trial in their paper– as is suggested by the Consolidated Standards of Reporting Trials (CONSORT) statement (Moher et al. 2010). The 95% confidence interval gives us a measure of how uncertain we are about where would the average treatment effect in reality lie and, usually, the larger the trial's sample size is, the narrower the 95% confidence interval is (even though sample size is not the sole determinant of the extent of the confidence interval).

The point estimate of the treatment effect (single value that is our best guess) and the 95% confidence interval can be used to gauge the relevant performance of a treatment, as has been shown in previous pieces. Here, we will also make a clinically informed choice about what would be an acceptable different between the two treatments that would make a difference for the patient (and the orthodontist). This could be either the minimum clinically-relevant alignment duration reduction that would make the new treatment preferable to the reference treatment (minimum clinical benefit) or it could be the maximum alignment duration increase that would not necessarily make the reference treatment clearly superior to the new treatment (maximum acceptable harm). Choosing this so called margin of non-inferiority

or margin of equivalence is not always easy and it might be associated with other possible advantages of the newer treatment like greater availability, reduced cost, less invasiveness, fewer adverse effects (harms), or greater ease of administration. However, this is a conscious decision that inadvertently includes some aspects of subjectivity. For example, starting from an average expected duration of tooth alignment of 8.8 months (263 days) (Wazwaz et al. 2022), one orthodontist might claim that a benefit of one month (30 days) might be clinically relevant to the patient. Another, maybe more conservative, orthodontist might claim that if the newer treatment is associated with increased invasiveness or potential side effects (let's say for example if corticotomies under complete periosteal flap are used), then a greater potential benefit of 40 or 50 days might be needed to justify the use of the newer treatment on a patient. The exact numbers here make no difference, but we need this notion of an equivalence margin (let's denote this as Δ) to relativise the two treatments.

Imagine we have established a priori (before we see the results of the trial) an almost universally acceptable equivalence margin (Δ) for tooth alignment duration. Then the results of the above-mentioned trial could fall into the different categories that are seen in Figure 1, which is similar to a forest plot, but for only a single study:

- Results are given as newer minus reference treatment in tooth alignment duration in days. This means that results on the left side of the vertical line (no-effect line) are beneficial for the new treatment, since they show a reduction in duration. Results on the right side indicate that the reference treatment is better than the newer treatment
- The potential results of the single trial are given as the difference in means (box) with their corresponding two-sided 95% confidence interval (horizontal black lines). Per convention, one can assume that if the horizontal line of the study crosses the vertical line (the 95% confidence interval contains zero) then the difference between treatments is not statistically significant at the 5% level.
- We can also visualise the margin of equivalence Δ, which is on both sides of the vertical line. Results that are on the grey regions with the pattern show either clinically relevant benefits (left grey region – region 1) or clinically relevant harms (right grey region –

region 4). Results falling in the middle white regions show either small benefits (region 2) or small harms (region 3), with which both the patient and the orthodontist can live with – meaning they wouldn't necessarily influence our clinical decision making.

Based on this information, we could have several scenarios about the relative performance of the two treatments, regarding of where the results of the trial might potentially fall.

Which of the following statements is correct, if any?

(A) The newer treatment can be claimed to be superior to the reference treatment for scenarios A & D, while the reference treatment is superior to the newer treatment for scenarios E & H.

(B) The new treatment can be claimed to be superior to the reference treatment to a clinically relevant extent in scenarios C & D.

(C) The new treatment can be claimed to be equivalent to the reference treatment in scenarios A, B, E, and F.

(D) The new treatment can be claimed to be non-inferior to the reference treatment in scenarios A-E.

(E) After testing and seeing that the new treatment is non-inferior to the reference treatment, there is no reason to check for superiority.

Discussion

As we mentioned earlier, an informal rule-of-thumb says that when the two-sided 95% confidence interval doesn't include the zero value, then a P value of < 0.05 is to be expected. This means, that given an error rate of 5%, the data derived from this specific trial on this specific setting might not be compatible with a null hypothesis of no difference between the two treatments. This is conventionally taken by clinicians to mean that we might expect a non-zero difference between the durations of the two treatments to exist. Therefore, all 4 scenarios A, D, E, and E have P values < 0.05 indicating differences between treatments. And as we set at the start the directions of the effects to be beneficial on the left for the new treatment and

beneficial for the reference treatment on the right, statement A is formally correct. The new treatment is superior to the reference treatment in scenarios A and D. The question however is, *how* superior is it against the reference?

As stated before, Δ (or even better –Δ) indicates a clinically relevant benefit that is universally accepted. Therefore, all studies with results on the grey region 1 (scenarios C and D) would be expected to show superiority of the newer treatment over the reference treatment. However, in scenario C the point estimate is on the grey area, but at the same time its 95% confidence interval crosses the line of no-effect (includes zero) and therefore in this scenario our data could potentially be compatible with a non-existing difference between the two treatments (we could anticipate either a zero difference or even a small positive difference). It can therefore be concluded that only scenarios D and not C can be used to claim a superiority of new over reference treatment to a clinically relevant extent.

It is logical to deduce that a newer treatment can be claimed to be equivalent to the older reference treatment when either (1) the newer treatment has in reality no difference to the reference treatment (difference of zero) or (2) the newer treatment is slightly more beneficial or detrimental, but within set margins of equivalence (absolute difference < Δ). The newer treatment can be expected to be slightly better than reference in scenario A and slightly better / no different / slightly worse in scenario B. In scenario E, the newer treatment is slightly worse than the reference, but to such a small extent that we can leave with it. Therefore, in A, B, and E, an equivalence can be claimed. However, in scenario F, even though the point estimate lies within +Δ, the 95% confidence interval includes also values greater than +Δ, which means that we might expect increased alignment durations being larger than our limit of acceptance (Δ). Therefore, no equivalence can be claimed for F, since at least one side of the 95% confidence interval includes the value of either –Δ or +Δ. Statement (C) is wrong.

The notion of non-inferiority is unique in the sense that it entails a one-sided direction, contrary to the notions of superiority (significant benefit or harm) and equivalence (either benefit or harm, but of small magnitude). For the newer treatment to be claimed non-inferior to the reference treatment, then we must have either scenarios of equivalence between

treatments or a clear superiority of the newer treatment. It can be therefore said that we might tolerate our 95% confidence interval having values of equivalence (–Δ to +Δ) or values of superiority (< –Δ), but not values over +Δ. It is easy then to conclude that non-inferiority of the newer treatment can be claimed when (i) we can expect reduced alignment durations (scenarios A & D), (ii) when we can expect reduced, zero, or slightly increased durations (scenarios B & C), or (iii) when we can expect slightly increased alignment duration (scenario E). Statement (D) is of course correct.

Finally, non-inferiority only tells us that the new treatment is not worse than the reference treatment (regions 1-3). A subset of this scenarios include also the possibility that the new treatment might be superior to the reference treatment (lie in region 1) and, therefore, testing for superiority might in some instance be a logical optional next step after the establishment of non-inferiority that gives us additional information. Statement (E) is correct.

References

Moher D, Hopewell S, Schulz KF, et al. (2010) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 340: c869.

Wazwaz F, Seehra J, Carpenter GH, et al. (2022) Duration of tooth alignment with fixed appliances: A systematic review and meta-analysis. *American Journal of Orthodontics & Dentofacial Orthopedics* 161: 20–36.

Figure 1. Plot depicting the results of a theoretical trial comparing two treatments in terms of tooth alignment duration.