Journal of Dentistry 128 (2023) 104380

Contents lists available at ScienceDirect

Journal of Dentistry

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

Clinical trials evaluating desensitising agents. Some design and analysis issues.

Robert G. Newcombe[a,*], Joon Seong[b], Nicola X. West[a]

[a] Cardiff University, Cardiff CF10 3AT, UK
[b] Clinical Trials Group, School of Oral and Dental Sciences, University of Bristol, Lower Maudlin Street, Bristol BS2 1LY, UK

ARTICLE INFO

Keywords:
Dentine hypersensitivity
Clinical trials
Relative risk reduction
Confidence interval

ABSTRACT

Introduction: The purpose of this short communication is to draw attention to an efficient design for trials to evaluate desensitising agents, and an appropriate statistical analysis.
Methods: Two recent sensitivity trials conducted by the Bristol Dental School Clinical Trials Group are reviewed.
Results: The methodology used was effective to establish efficacy of the products evaluated.
Conclusions: This methodology is recommended for wider use.
Clinical Significance: Effective clinical trial methodology enables establishment of efficacy of desensitising products leading to patient benefit.

1. Introduction

Many desensitising agents active against dentine hypersensitivity (DH) are now available in toothpaste form. The commonest scenario is routine home use over a period to control sensitivity, achieving oral hygiene applying a desensitising paste by brushing in place of a regular one. Alternatively, a paste may be applied by a dental professional directly to the affected site(s), with the prospect of substantial instant relief. This short communication explains some of the methodological issues that pertain to trials to evaluate such agents, with reference to two trials recently published by the Bristol Dental School Clinical Trials Group.

2. Two recent sensitivity trials

Trial 1 [1] compared a toothpaste containing calcium silicate and sodium phosphate to a control paste for DH pain reduction after 14 and 28 days regular home use.

Trial 2 [2] compared a toothpaste containing aluminium lactate, potassium nitrate and hydroxylapatite to a control paste containing potassium nitrate for DH pain reduction immediately after a single supervised brushing and after 7 and 14 days regular home use.

In both studies, two teeth were selected which were sensitive at baseline with Schiff [3] scores 2 or 3, then re-scored after a period of either active or control product use. Both led to very clear conclusions. Other researchers may latch on to this being a very effective design – which it is. However, there is a substantial risk of failing to realise that special statistical methods are needed to obtain a statistically sound assessment of benefit in the form of the relative risk reduction. Running a simplistic analysis as if the two study teeth are independent can lead to seriously misleading inferences being drawn.

3. Basics of comparative trial design

In trial 1, in the control group the mean Schiff score decreased from 2.25 at screening and baseline to 1.89 and 1.81 after 14 and 28 days of product use. An even more marked decrease occurred in the control group in trial 2. This phenomenon is widely observed in trials of this kind. It is attributable to two effects. The **placebo effect** is well-known: participants may obtain subjective benefit knowing they may be on active treatment. But what is more important here is **regression to the mean**. Participants are required to have a specified level of sensitivity to be eligible for recruitment. An individual's level of sensitivity tends to fluctuate, for a variety of reasons. A patient whose sensitivity level hovers around the threshold for eligibility will be admitted to the trial if it is relatively high for that individual at the time of screening, not if it is relatively low. This is the principal reason why this decline is observed.

These two phenomena are commonly observed in trials of interventions in a wide range of dental and other health-related contexts. This is the main reason why definitive research consists of controlled

[*] Corresponding author.
E-mail address: Newcombe@cardiff.ac.uk (R.G. Newcombe).

https://doi.org/10.1016/j.jdent.2022.104380
Received 26 October 2022; Received in revised form 21 November 2022; Accepted 23 November 2022
Available online 29 November 2022

trials, in which the effect of an intervention applied to a group of participants is compared against an alternative regime applied to a different, similar group. We do not simply evaluate the benefit in a single group, as we would expect to observe some improvement anyway, for both reasons, even if the treatment under test was ineffective.

The most effective way to ensure the two groups start off similar is to randomly allocate eligible participants between the two regimes. This is why the **randomised controlled trial** is regarded as the gold standard in evaluating treatments.

In some other dental contexts, a crossover or split-unit design is effective and highly efficient. This applies particularly to designs that are experimental rather than therapeutic, such as plaque regrowth, salivary bacterial count and experimental gingivitis studies. For examples of such designs see references [4] [5] [6].

However, these designs are not applicable to trials of desensitising agents. Split-mouth designs do not work, simply because even if applied topically to one tooth, an agent has the potential to affect all other teeth. Crossover designs do not work, on account of the profound effect of regression to the mean.

4. Choice of outcome measure

The Schiff score [3] involves examiner rating of participant response to a DH challenge on a 4-point scale:

1 Subject does not respond to stimulus
2 Subject responds to stimulus but does not request discontinuation of stimulus
3 Subject responds to stimulus and requests discontinuation or moves from stimulus
4 Subject responds to stimulus, considers stimulus to be painful, and requests its discontinuation.

In study 1, the stimulus was a cold air blast. After shielding adjacent proximal teeth, a one-second blast of air was directed onto the exposed buccal root surface of the tooth from a distance of 1 cm, at 55-65 psi and 19-21°C. In study 2, a drop of iced water at 0°C was applied to the exposed dentine at the buccal cervical region of each identified tooth in turn. In both trials, this index was highly sensitive to detect the difference between active and control pastes.

In both trials, the response at tooth level was also assessed by a tooth-level VAS score. This yielded only mediocre discrimination between active and control pastes.

DH in response to tactile challenge [7] was determined by Yeaple probe which was calibrated at the start of every study day. Starting at a force of 10g and increasing in 10g increments the probe tip was passed over the exposed dentine on the buccal surface of the selected teeth, apical to the cement-enamel junction until the participant indicated that they were experiencing discomfort by providing a "yes" response. The force setting which elicited the “yes” response was repeated, and if a second "yes" was not obtained, the force setting was increased by 10g. Sensitivity was assessed until a force which elicited two consecutive "yes" responses was identified.

This measurement can be highly sensitive to detect difference if used expertly, but otherwise is less effective than Schiff scoring. The tactile and Schiff measures of sensitivity should be regarded as complementary. By no means the same teeth are identified as having benefitted from treatment, so correlations between these two scores are only moderate: in study 1, rank correlations between these measures ranged from -0.18 to -0.51 in various analyses, with a median of -0.26. The minus signs reflect the fact that the tactile score represents the force required to produce sensitivity, accordingly it increases, not decreases, when an effective agent is used.

The DHEQ15 quality of life questionnaire [8] was supplemented with 8 additional questions. All were scored on a 7-point Likert scale from strongly disagree to strongly agree. This was found to be of limited value to distinguish active vs. control paste in both studies.

5. Number of teeth per patient to evaluate

We believe that studying 2 teeth per mouth achieves virtually as much as if larger numbers of teeth were scored, because the responses to treatment from different teeth in the same mouth are far from independent. We settled for 2 teeth per mouth from different quadrants, avoiding molars and adjacent central incisors. It is important to avoid using adjoining or nearby teeth. In a population study of DH [9], the correlation between responses of pairs of teeth is high for adjoining teeth and declines as the distance apart increases: studying pairs of teeth from opposite quadrants of the same arch, the median rank correlation was 0.70 for pairs of incisors, in contrast to 0.24 for other pairs of teeth.

Using 2 designated teeth per mouth has the further advantage that a valid statistical analysis comparing proportions of teeth that remain sensitive after treatment is readily available, which duly takes into account the resulting non-independence.

6. Analysis – general principles

For trials generally with quantitative (scoring) outcome, and baseline scores available, the method of choice is analysis of covariance (ANCOVA). The ANCOVA model has two explanatory variables, the treatment allocated, and the baseline value for the variable in question. This analysis strategy is generally better than two alternatives sometimes encountered, (a) examining changes just on active by single-sample paired t-test or (b) comparing changes between 2 groups using a two-groups t-test.

Throughout the history of clinical investigation, it has been practically instinctive for researchers to start by examining changes in clinical outcomes from baseline to after active treatment. However, this is liable to give a grossly biased impression of efficacy, simply for the two reasons already discussed – placebo effect and regression towards the mean. As we have seen, this is why it is recognised that we need to do trials in which some participants get the treatment of interest, other similar patients a matching control treatment.

A simple way to get around this issue is to compare changes – such as reductions in Schiff score – between the two groups. However, this strategy tends to over-adjust for chance baseline differences between groups. The best way to see this is to explain just what ANCOVA does. ANCOVA essentially fits a regression model in which the y (outcome) variable is the value of the response variable after treatment, and the x (input) variable is the corresponding pre-treatment value. Normally, we expect that x and y will be positively correlated. If the response variable is measured immediately after the baseline, the correlation r, and closely linked to it, the regression coefficient b will be close to 1 – the shortfall merely reflecting imperfect reproducibility or objectivity of the score. As the time lag increases, the correlation and regression coefficient decrease markedly. A simple comparison of incremental changes in the two groups would be correct if b were 1. As b tends to be lower than 1, this is why comparing incremental changes between the two groups tends to over-adjust for baseline differences.

Also, it has been shown that the decision whether to adjust for baseline as covariate must not depend on whether there is a (significant) difference between groups at baseline, or whether there is a (significant) baseline-response correlation – these criteria are misleading [10]. No ifs, no buts - ANCOVA simply is the strategy of choice, and can go straight into the statistical analysis plan.

That is the general explanation. Though, in our two DH trials, participants were required to have Schiff score 2 or 3 at baseline for the two designated teeth, and in practice, scores of 2 predominated. Consequently, there is very little baseline variation here, and the effect of covariate adjustment is minimal. Indeed, the alternative analysis explained below treats the Schiff score as binary anyway, positive (2 or 3) or negative (0 or 1).

7. An alternative analysis – sensitivity treated as binary

For our 2 teeth per mouth design, most analyses are naturally based on the average of the scores of the two teeth, at each time point. In the two studies, we do report ANCOVAs for each outcome measure, on this basis.

But a neat alternative analysis leads directly to an estimate of the relative risk reduction (RRR) based on these two teeth, comparing active vs. control, with a 95% confidence interval (CI). The correct calculation for the CI for the RRR is as follows. We first obtain CIs for proportions of teeth remaining sensitive at follow-up for each product, using a method I developed that allows for responses of the two teeth not being statistically independent [11]. These are then post-processed to obtain an interval for their ratio [12].

Two Excel spreadsheets **MEAN012.xls** and **MOVER-R.xls** are available online accompanying this article, which perform the calculations described in these references. They enable post-processing of results obtained from SPSS or other statistical analysis software.

In trial 1, of the 125 participants using the active treatment, 66 (52.8%) had neither designated tooth still sensitive (Schiff 3 or 2) at 28 days. 30 (24.0%) had one of the two designated teeth sensitive, the remaining 29 (23.2%) had both designated teeth still sensitive.

When we open the spreadsheet **MEAN012.xls**, it shows the calculation [11] performed for these results. Here, $2 \times 29 + 30$ or 88 teeth are still sensitive. The mean number of designated teeth per participant that remain sensitive is 88 / 125 or 0.7040, with 95% confidence interval from 0.5687 to 0.8539.

Here, we are more interested in the probability that a tooth that was sensitive at baseline remains sensitive at 28 days. This is simply 88 out of $2 \times 125 = 250$ or equivalently half of 0.7040, i.e. 0.3520. The spreadsheet also displays confidence limits for this figure, which are obtained by halving 0.5687 and 0.8539, namely 0.2844 and 0.4270.

The corresponding figures for the 122 in the control group are very different. 23 (18.9%) had neither designated tooth still sensitive; 27 (22.1%) had one of the two designated teeth sensitive, the remaining 72 (59.0%) had both designated teeth still sensitive.

When we substitute these figures for n_0, n_1 and n_2 in place of 66, 30 and 29 in the spreadsheet **MEAN012.xls**, we obtain a mean of 1.4016 of the 2 teeth still sensitive, with 95% confidence limits 1.2529 to 1.5300. Just as in the active group, these figures are then halved, to get the proportion of designated teeth that remained sensitive, 0.7008, with 95% confidence interval from 0.6265 to 0.7650.

Thus the proportion of teeth that remain sensitive at 28 days on active, 35.2%, is just over half (50.2%) of the corresponding proportion for the control, 70.1%. In other words, the proportion of teeth that still have sensitivity at day 28 was 100 - 50.2 or 49.8% lower using active compared to control. We call this figure the **relative risk reduction**. We assess that active treatment reduces the risk of a tooth remaining sensitive at 28 days by 49.8% in relative terms.

The second spreadsheet, **MOVER-R.xls** obtains a confidence interval for the ratio of two quantities [12]. This is used to post-process the results of two applications of **MEAN012.xls** to obtain a confidence interval for the relative risk reduction. When we open **MOVER-R.xls**, it displays the calculations for trial 1 at 28 days. To use this for any other data, the estimated proportions of designated teeth remaining sensitive on the two treatments may be copied in from those produced by **MEAN012.xls** using Paste – Values.

We see that the ratio of the proportions of teeth that remain sensitive at 28 days on active and control, 0.5023, has a 95% confidence interval from 0.3991 to 0.6283. The relative risk reduction is the complement of this, 0.4977, with 95% confidence interval 0.3717 to 0.6009. The difference between the two treatments is statistically significant ($p<0.001$) in favour of active treatment.

In all these calculations, we can alter the confidence level if desired, to get say 90% or 99% limits. To be meaningful, this needs to be chosen carefully at the sample size planning phase and applied consistently choice throughout the analysis, of course.

An earlier version of **MEAN012.xls** (which displays different illustrative figures) is freely downloadable from the website associated with the first author's book https://www.routledge.com/Confidence-Intervals-for-Proportions-and-Related-Measures-of-Effect-Size/Newcombe/p/book/9780367576707# alongside several other related spreadsheets which are used in a similar way.

The table below shows these results alongside similar analyses for trial 1 at 14 and 28 days and trial 2 after a single treatment and at 7 and 14 days.

It is important to recognise that others may seek to copy this methodology, without realising that they need to use this special analysis for the results to be valid. The table also shows results obtained using a more naïve analysis based on the best CI method for an ordinary ratio of proportions [13], which treats the two teeth as if independent. 88 (35.2%) of the 250 designated teeth treated with the active paste remained sensitive at 28 days, compared to 171 (70.1%) of the 244 teeth on control. The relative risk reduction is exactly the same as above, 49.8%. But the calculated confidence interval is narrower, 39.7% to 58.5%.

—	Relative risk reduction	95% confidence interval	
		Correct	Naïve
Trial 1			
14 days	0.302	0.191 to 0.402	0.203 to 0.393
28 days	0.498	0.372 to 0.601	0.397 to 0.585
Trial 2			
Immediate	0.550	0.335 to 0.697	0.369 to 0.688
7 days	0.810	0.591 to 0.908	0.630 to 0.905
14 days	0.886	0.537 to 0.970	0.663 to 0.963

In all analyses, the naïve interval is somewhat too narrow, in comparison to the correct one that duly heeds non-independence. This applies particularly to the extreme RRR found at 14 days in trial 2. All of these comparisons are statistically significant ($p<0.001$) irrespective of which approach is used. But in studies with less clear evidence of benefit, attainment of statistical significance could be affected. This issue needs to be borne in mind if any studies using this design get submitted for publication.

8. Discussion

This situation pervades several health-related contexts involving paired or multiple organs. In the 1980s, many publications in the field of ophthalmology contained analyses which treated the two eyes of the same person as if they were statistically independent. Newcombe & Duff [14] published a simulation study based on bilateral inter-ocular pressure data from an actual trial. The aggregate results from all participants were repeatedly re-randomised into two groups, and statistical significance ($p<0.05$) or non-significance noted, based on the incorrect analysis method. This process should produce a statistically significant difference between groups in 5% of simulation runs. The false positive rate produced by doing so was found to be much higher, at 20%.

Considering the number of teeth per mouth, and sometimes also multiple surfaces or gingival sites per tooth, the possibility of incorrect analysis in some dental health contexts is mind-boggling! In our design, it is just 2 units per mouth – which as argued above is quite sufficient, so the impact is not so severe. But it does make a difference here also.

Concepts of independence and coincidence are central to statistical thinking, but tend to be poorly understood and little heeded in other contexts. The movie *Sully: Miracle on the Hudson* was based on the true story of Chesley 'Sully' Sullenberger's January 2009 emergency landing of US Airways Flight 1549 on the Hudson River, and the subsequent publicity and investigation. One of the issues used initially to discredit the pilot's account of events was the alleged implausibility of the impact of a flock of birds wrecking both starboard and port engines - losing one engine to a bird strike was an infrequent occurrence, therefore losing both ‘must’ be very infrequent indeed. But a moment’s reflection should convince us that this argument of an exceedingly rare coincidence is fundamentally flawed – it was simply the same flock of Canada geese.

Recommendations

- Desensitising agents need to be evaluated in parallel-groups randomised trials.
- Objective scoring methods involving provoking sensitivity yield much clearer conclusions than subjective ones.
- A very efficient design selects 2 initially sensitive teeth for treatment.
- Excel resources are available to perform appropriate data analyses for this design.

CRediT authorship contribution statement

Robert G. Newcombe: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Joon Seong:** Investigation, Resources, Writing – original draft. **Nicola X. West:** Investigation, Resources, Writing – original draft.

Conflict of interest statement

The authors declare that they have no competing interests with regard to this paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jdent.2022.104380.

References

[1] J. Seong, R.G. Newcombe, J.R. Matheson, L. Weddell, M. Edwards, N.X. West, A randomised controlled trial investigating efficacy of a novel toothpaste containing calcium silicate and sodium phosphate in dentine hypersensitivity pain reduction compared to a fluoride control toothpaste, J. Dent. (2020) 98, https://doi.org/10.1016/j.jdent.2020.103320.

[2] J. Seong, R.G. Newcombe, H.L. Foskett, M. Davies, N.X. West, A randomised controlled trial to compare the efficacy of an aluminium lactate/potassium nitrate/hydroxylapatite toothpaste with a control toothpaste for the prevention of dentine hypersensitivity, J. Dent. 108 (2021), https://doi.org/10.1016/j.jdent.2021.103619.

[3] T. Schiff, M. Dotson, S. Cohen, W. De Vizio, J. McCool, A. Volpe, Efficacy of a dentifrice containing potassium nitrate, soluble pyrophosphate, PVM/MA copolymer, and sodium fluoride on dentinal hypersensitivity: a twelve-week clinical study, J. Clin. Dent. 5 (1994) 87–92.

[4] M. Addy, J. Greenman, P. Renton-Harper, R.G. Newcombe, F. Doherty, Studies on stannous fluoride toothpaste and gel. 2. Effects on salivary bacterial counts and plaque regrowth in vivo, J. Clin. Periodontol. 24 (1997) 86–91, https://doi.org/10.1111/j.1600-051x.1997.tb00472.x.

[5] N.X. West, M. Addy, R.G. Newcombe, et al., A randomised crossover trial to compare the potential of stannous fluoride and essential oil mouth rinses to induce tooth and tongue staining, Clin. Oral. Invest. 16 (2012) 821–826, https://doi.org/10.1007/s00784-011-0560-9.

[6] S. Daly, J. Seong, R. Newcombe, M. Davies, J. Nicholson, M. Edwards, N. West, A randomised clinical trial to determine the effect of a toothpaste containig enzymes and proteins on gum health over 3 months, J. Dent. 90 (2019) S26–S32, https://doi.org/10.1016/j.jdent.2018.12.002.

[7] A.M. Polson, J.G. Caton, R.N. Yeaple, H.A. Zander, Histological determination of probe tip penetration into gingival sulcus of humans using an electronic pressure-sensitive probe, J. Clin. Periodontol 7 (1980) 479–488, https://doi.org/10.1111/j.1600-051X.1980.tb02154.x.

[8] C. Machuca, S.R. Baker, F. Sufi, S. Mason, A. Barlow, P.G. Robinson, Derivation of a short form of the dentine hypersensitivity experience questionnaire, J. Clin. Periodontol. 41 (2013) 46–51, https://doi.org/10.1111/jcpe.12175.

[9] J. Seong, D. Bartlett, R.G. Newcombe, N.C.A. Claydon, N. Hellin, N.X. West, Prevalence of gingival recession and study of associated related factors in young UK adults, J. Dent. 76 (2018) 58–67, https://doi.org/10.1016/j.jdent.2018.06.005.

[10] S. Senn, Change from baseline and analysis of covariance revisited, Stat. Med. 25 (2006) 4334–4344, https://doi.org/10.1002/sim.2682.

[11] R.G. Newcombe, Confidence intervals for the mean of a variable taking the values 0, 1 and 2, Stat. Med. 22 (2003) 2737–2750, https://doi.org/10.1002/sim.1479.

[12] R.G. Newcombe, MOVER-R confidence intervals for ratios and products of two independently estimated quantities, Stat. Meth. Med. Res. 25 (2016) 1774–1778, https://doi.org/10.1177/0962280213502144.

[13] O. Miettinen, M. Nurminen, Comparative analysis of two rates, Stat. Med. 4 (1985) 213–226, https://doi.org/10.1002/sim.4780040211.

[14] R.G. Newcombe, G.R. Duff, Eyes or patients? Traps for the unwary in the statistical analysis of ophthalmological studies, Br. J. Ophthalmol. 71 (1987) 645–646, https://doi.org/10.1136/bjo.71.9.645.