

The Influence of Bias on the Outcome of Prospective *versus* Retrospective Cephalometric Twin-Block Trials

Michael John Trenouth*

Consultant Orthodontist (Retired), Royal Preston Hospital, Sharoe Green Lane, United Kingdom

*Corresponding author: Michael John Trenouth, Consultant Orthodontist (Retired), Royal Preston Hospital, Sharoe Green Lane, Preston, PR2 9HT, United Kingdom, Tel: +44/15395-3378; E-mail: michaeltrenouth@hotmail.co.uk

Received: 08 Feb, 2021 | Accepted: 16 Feb, 2021 | Published: 23 Feb, 2021

Citation: Trenouth MJ (2021) The Influence of Bias on the Outcome of Prospective *versus* Retrospective Cephalometric Twin-Block Trials. Int J Dent Oral Health 7(3): dx.doi.org/10.16966/2378-7090.352

Copyright: ©2021 Trenouth MJ. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Objective: This was to investigate whether a retrospective study inflates or deflates the treatment response compared to a prospective study.

Materials and Methods: The results of a prospective randomized trial were compared to those obtained from two retrospective trials all comparing the outcome of twin block treatment with either incisor torquing springs or passive labial bow.

Results: There were no statistically significant differences in outcome except for a lesser degree of skeletal correction, measured by reduction in angle ANB and increase in angle SNB in the passive labial bow group in one of the retrospective studies.

Conclusions: The retrospective study design diminished rather than magnified the treatment effect in the control group of only one study. This could be attributed to the sequential treatment of the control group followed by the test group introducing bias together with the Hawthorne effect.

Keywords: Twin-block appliance; Torquing clasp; Prospective; Retrospective; Trial bias

Introduction

Prospective randomized controlled clinical trials are generally acknowledged to provide the highest level of evidence [1].

Observational retrospective studies provide a lower level of evidence because of their potential to introduce bias [2]. Bias leads to distortion of the treatment effects such that the differences observed might be due to how the groups were selected. Elimination of bias reduces the number of variables to one enabling a causal association between intervention and outcome.

Consolidated Standards of Reporting Trials or CONSORT [3,4] provides a checklist of essential items for reporting randomized clinical trials and also a flow chart of study participants. There are three recognized sources of bias, allocation, assessment and attrition. Allocation bias can be overcome by randomization such that each group is as alike as possible. Assessment bias can be overcome by blinding the operator to which group the patient belongs when measuring the outcome. Attrition bias occurs due to the loss of cases to follow up as documented in the CONSORT flow chart. The patients who fail to complete treatment may differ from those who do complete, introducing bias to the results. This can only be assessed by comparing the patients who defaulted with those who completed treatment to see if there was a difference [5]. Other potential sources of bias are operator skills where more than one operator is involved in the treatment [6]. This is especially so in multicentre trials because

these also involve different patient populations which may introduce further bias. Normative data may show statistically significant differences between different patient populations [7,8].

In 2012 O'Brien stated that "retrospective clinical trials are likely to overestimate the treatment effect by approximately 30 %" [9]. Shaw WC, et al. [10] considered that retrospective studies inflate the treatment response as this has been found to occur in other fields of research such as psychiatry and medicine. However, the opposite effect was found in the early Streptomycin studies for the treatment of tuberculosis where retrospective studies deflated the treatment effect and prospective studies were necessary to demonstrate its effectiveness [11]. In a meta-analysis Papageorgiou SN, et al. [12] found that nonrandomized historical controls deflated the treatment effect.

Trenouth MJ, et al. [13] compared the treatment change, as measured by angle ANB, during Twin-block appliance therapy reported in four prospective studies with those of four retrospective studies. There was no statistically significant difference between outcomes for the two types of study. Also, control groups derived from growth study data in the retrospective studies did not differ significantly from untreated Class II Division 1 cases used in the prospective studies.

The literature search poses the research question as to whether a retrospective study inflates or deflates the treatment response compared to a prospective study.

In 2012 Trenouth MJ, et al. [5] performed a prospective randomized clinical trial of two alternative designs of Twin-block appliance. 52 patients were randomly allocated to two parallel groups, Southend and Non-Southend. The Twin-block appliances were identical except for the presence or absence of a Southend clasp on the upper and lower incisors. They found that the presence of a Southend clasp on the upper and lower incisors during Twin-block treatment limited their tipping which enhanced the skeletal correction. Harradine and Gale in 2000 [14] and Parkin NA, et al. in 2001 [15] found the same outcome with similar designs of torquing springs in their test groups and passive labial bows in their control groups. Both these studies were non-randomized and retrospective in design. Harradine and Gale selected their two groups from 200 patients, the first 90 had a conventional labial bow (CTB) and the next 110 had torquing spurs (STB). 30 cases were selected for each group based on good quality records and a final overjet of 4 mm or less. Parkin NA, et al. selected 36 patients consecutively treated between 1994 and 1996 for their labial bow group (TB1) which was reported by Lund DI, et al. in 1998 [16]. Between 1997 and 1999, 27 out of 49 patients on the waiting list met the criteria for inclusion in the torquing spring and high pull headgear group (TB2). Although the patients were recruited prospectively they did not meet the CONSORT criteria for a prospective trial because the two groups were selected at different periods. The present study was undertaken to compare the results of these two non-randomized retrospective studies with the prospective randomized study.

Materials and Methods

The results of the prospective randomized study reported by Trenouth MJ and Desmond S were compared to the retrospective studies of Parkin NA, et al. and Harradine NWT and Gale D. All studies compared two designs of a twin-block appliance, one group with some form of incisor torquing and the other as a control group. Trenouth MJ and Desmond S used a Southend clasp on both upper and lower central incisors in their Southend group compared to a passive labial bow in their Non-Southend group. Parkin NA, et al. [15] used torquing spurs on the upper central incisors together with high pull headgear in their TB2 group compared to a labial bow in their TB1 group. Harradine NWT and Gale D used torquing spurs on the upper central incisors in their STB group compared to a labial bow in their CBT group.

The cephalometric measurements analysed in the present study needed to be common to all three studies. The basic measurements reported were SNA (Sella-Nasion-Point A), SNB (Sella-Nasion-Point B), ANB (by subtraction of SNB from SNA), UI (Upper incisor long axis to the maxillary plane ANS-PNS), LI (Lower incisor long axis to the mandibular plane Go-Me).

The basic demographic data for all three studies are given in table 1. All three studies were tested for pre-treatment equivalence. This was verified by statistical testing which showed no significant differences

between the control and test groups before treatment for all three studies.

All studies also conducted an error analysis. Trenouth MJ and Desmond S assessed both the random and systematic error using the method of Bland JM and Altman DG [17] as did Parkin NA, et al. Harradine NWT and Gale D assessed random error using correlation coefficient and systematic error using a t-test after Houston [18].

Statistical Analysis

Ideally, the data from the different studies could be compared using confidence intervals to see if there was any overlap or not, but these were only reported in the Trenouth MJ and Desmond S study. The mean values, standard deviations and number of subjects in each group were reported in all three studies. For this reason, a two-sample t-test for means and standard deviations was chosen and analysis performed using NCSS (Number Cruncher Statistical System), UT, and the USA. The prospective data from the Trenouth MJ and Desmond S study was compared with the two sets of retrospective data from the Parkin NA, et al. and the Harradine NWT and Gale D studies to see if there were any statistically significant differences between the means of the three groups.

When multiple comparisons of data are made the probability of obtaining a significant result is increased leading to a type 1 error [19]. For this reason, a Bonferroni correction was applied to the analysis. Statistical significance was only assumed at a higher level of probability calculated as p divided by the number of comparisons which in the present study was 10.

Results

The raw data for the Trenouth MJ and Desmond S and Harradine NWT and Gale D studies with the results of the t-test with Bonferroni correction are shown in table 2. The control group with a passive labial bow for the Harradine NWT and Gale D study showed a statistically significant lesser skeletal change in ANB and SNB than the Trenouth MJ and Desmond S study. All other comparisons were non-significant.

The raw data for the Trenouth MJ and Desmond S and Parkin NA, et al., studies together with results of the t-test with Bonferroni correction are shown in table 3. All the comparisons were non-significant there being no differences in outcome between the two studies.

The raw data for the Harradine NWT, et al., and Parkin NA, et al., studies were also subject to a t-test with Bonferroni correction but all comparisons were statistically non-significant.

Discussion

The only statistically significant differences detected between the three trials were that between Trenouth MJ and Desmond S and Harradine NWT and Gale D for the control groups with passive labial

Table 1: Demographic Data.

Author	Trenouth and Desmond [5]		Parkin NA, et al. [15]		Harradine and Gale [14]	
	Southend	Non-Southend	TB1	TB2	CTB	STB
n=	20	21	36	27	30	30
M/F	8/12	11/10		12/15	14/16	14/16
Start age (years)	15.2 (4.4)	14.7 (3.3)				
Treatment time (months)	10.0 (3.0)	9.9 (4.0)			8.4 (3.6)	8.4 (3.6)
Overjet at start (mm)	11.0 (2.1)	11.0 (2.3)	8.17 (1.91)	8.75 (2.1)	10.6 (2.5)	9.9 (2.3)

Table 2: Cephalometric Data Comparison.

Author	Clasp	Trenouth and Desmond [5]			Harradine and Gale [14]			P=
		Mean	SD	n	Mean	SD	n	
SNA	TQ	-0.75	1.12	20	-0.9	2	30	ns
	LB	-0.43	0.81	21	-0.5	1.2	30	ns
SNB	TQ	3.2	1.51	20	2	1.2	30	ns
	LB	2.44	1.11	21	1.2	1.2	30	0.01
ANB	TQ	-3.45	1.05	20	-2.9	2.2	30	ns
	LB	-2.57	0.74	21	-1.6	0.9	30	0.001
UI	TQ	-6.05	6.03	20	-6.9	3.7	30	ns
	LB	-12	6.63	21	-14.1	5.7	30	ns
LI	TQ	3	3.6	20	4.7	4.7	30	ns
	LB	6.86	4.54	21	4.6	3.5	30	ns

Table 3: Cephalometric Data Comparison.

Author	Clasp	Trenouth and Desmond [5]			Parkin NA, et al. [15]			P=
		Mean	SD	n	Mean	SD	n	
SNA	TQ	-0.75	1.1	20	-1.4	1.1	27	ns
	LB	-0.43	0.8	21	-0.1	1.6	36	ns
SNB	TQ	3.2	1.5	20	2.4	1.7	27	ns
	LB	2.44	1.1	21	1.9	2	36	ns
ANB	TQ	-3.45	1.1	20	-3.8	2	27	ns
	LB	-2.57	0.7	21	-2	1.9	36	ns
UI	TQ	-6.05	6	20	-6.9	6.3	27	ns
	LB	-12	6.6	21	-11	7.6	36	ns
LI	TQ	3	3.6	20	6.4	6.3	27	ns
	LB	6.86	4.5	21	8.2	7.1	36	ns

bows. The retrospective Harradine NWT and Gale D study produced a lesser degree of skeletal correction as measured by a decrease in angle ANB and an increase in angle SNB than the prospective Trenouth MJ and Desmond S study. The finding in the present study is that the retrospective design diminished rather than magnified the treatment response. The possible sources of bias that could explain this difference were listed in the CONSORT guidelines.

Allocation bias in the Trenouth MJ and Desmond S study was reduced by random prospective assignment of patients to the control and test groups, based on random sequence generation and blinding of the operator into which group the patient was placed by the use of sealed envelopes. This produced two groups that were as alike as possible except for the factor under test. Statistical comparison of the two groups showed a non-significant difference. Selection factors were applied to ensure that only cases of adequate severity were included, this also applied to the Parkin NA, et al. study. In the retrospective study of Harradine NWT and Gale D 30 patients were selected to form the control group followed by 30 to form the test group. No selection criteria were applied which could result in milder cases being included at the start explaining the smaller response. However, when tested for differences at the start of treatment all cephalometric measurements were non-significantly suggesting an absence of selection bias.

Assessment bias was controlled in the Trenouth MJ and Desmond S study by having one tracer who analysed the cephalometric radiographs at a later date in the same random order that they entered the study and was blinded as to which group they were allocated. The Parkin NA, et al. and Harradine NWT and Gale D studies analysed two sequential groups, a control group followed by a test group and so the tracers could not, therefore, be blinded as to which group they were dealing with. However, in the Parkin NA, et al., study the radiographs were randomly digitized in each group which would help to reduce assessment bias. All three studies performed satisfactory error studies.

Attrition bias was controlled for in the prospective Trenouth MJ and Desmond S study using the CONSORT flow chart. The number of patients was accounted for and documented at each stage of the trial, enrolment, allocation to test or control group, follow up and analysis. Also, the patients who defaulted were compared with those who completed treatment and no differences were found. Harradine NWT and Gale D measured their default rate which was 17.5 %. This was comparable with that of Trenouth MJ and Desmond S which was 17.3 % [20] and below the mean of 23.7 % for 10 studies reported in the literature. Parkin NA, et al. did not report their default rate.

Operator skills can be an important factor in multicentre studies especially if clinicians are expected to use appliance techniques with which they are not readily familiar. However, in all three studies, the operators were skilled clinicians familiar with and committed to the techniques they used. Trenouth MJ and Desmond S and Harradine NWT and Gale D had two operators Parkin NA, et al. had two operators for the test and one for the control group together with a consultant common to both groups who prescribed the treatment.

Population differences were probably minimized because all three studies were performed in a single limited geographical area which should balance the control and test groups as these are drawn from the same population.

Appliance design is another consideration. The major difference was the use of high pull headgear as well as torquing springs in the Parkin NA, et al. study but this did not appear to have any effect as there were no significant differences between the test groups for the three studies. The only statistically significant difference was between Trenouth MJ and Desmond S and Harradine NWT and Gale D for the control groups with passive labial bows but there was little difference between the appliance designs.

The Hawthorne effect is where subjects modify their behavior simply because they are aware of being observed. It may also apply to the operators performing the treatment which is not often considered. In the present situation, it could influence the motivation of patients wearing their removable functional appliances. Madsen H, et al. [21] did not consider the Hawthorne effect of importance in orthodontics because the compliance rates of randomized controlled trials (RCT'S) did not exceed normal levels. Because the Hawthorne effect is not included in CONSORT guidelines it was only considered in 3.4 % of RCT'S [22]. Also, the Hawthorne effect wears off after time and thus has a lesser influence in longer studies. In all three studies, the operators were committed to using what they considered to be the functional appliance systems of their choice. Sandler J, et al. [23] attributed unusually high compliance with headgear to the Hawthorne effect but in the present study, there was no difference between Parkin NA, et al. torquing springs with headgear and Harradine NWT and Gale D torquing only and Trenouth MJ and Desmond S Southend clasp only.

The lesser skeletal response in the Harradine NWT and Gale D control group could be explained by the study design in that this group was treated first. It was then followed by the test group with torquing springs as an innovation. This may have resulted in the operators subconsciously devoting more attention to the patients resulting in higher levels of compliance through the Hawthorne effect. Meikle MC, et al. [24] pointed out that cephalometric outcome not only measured the ability of the appliance to alter the pattern of maxillary and/or mandibular growth but also patient compliance, the magnitude and/or timing of pubertal facial growth, and the skill and personality of the clinician.

Summary and Conclusions

- One randomized prospective trial was compared with two non-randomized retrospective trials. In each study, the cephalometric outcome of Twin-block treatment was measured.
- Each study had a test group with incisor torquing springs and a control group with a passive labial bow.
- No statistically significant differences in outcome were detected except for a lesser degree of skeletal correction in the control group of one retrospective study. The retrospective design diminished rather than magnified the treatment response in this study.

The findings that incisor torque control not only limited incisor tipping but enhanced the skeletal correction were the same for all three studies.

References

- Pandis N (2011) The evidence pyramid and introduction to randomised controlled trials. *Am J Orthod Dentofacial Orthop* 140: 446-447.
- Pandis N (2011) Sources of bias in clinical trials. *Am J Orthod Dentofacial Orthop* 140: 595-596.
- Begg C, Cho M, Eastwood S, Horton R, Moher D, et al. (1996) Improving the quality of reporting of randomised controlled trials. The CONSORT Statement. *JAMA* 276: 637-639.
- Turpin DL (2005) CONSORT and QUORUM guidelines for reporting randomized clinical trials and systematic reviews. *Am J Orthod Dentofacial Orthop* 128: 681-685.
- Trenouth MJ, Desmond S (2012) A randomized clinical trial of two alternative designs of Twin-block appliance. *J Orthod* 39: 17-24.
- Mew J (2002) Are random controlled trials appropriate for orthodontics? *Evid Based Dent* 3: 35-36.
- Trenouth MJ, Davies PHJ, Johnson JS (1985) A statistical comparison of three sets of normative data from which to derive standards for craniofacial measurement. *Eur J Orthod* 7: 193-200.
- Trenouth MJ, Laitung G, Naftel AJ (1999) Differences in cephalometric reference values between five centres: relevance to the Eurocleft Study. *Br J Oral Maxillofac Surg* 37: 19-24.
- Millett DT, Cunningham SJ, O'Brien KD, Benson PE, De Olivera CM (2012) Treatment and stability of Class II Division 2 malocclusion in children and adolescents: A systematic review. *Am J Orthod Dentofacial Orthop* 142: 159-169.
- Shaw WC (2000) How relevant is the evidence-based process to orthodontics. *Evid Based Dent* 2: 7-8.
- Mann H, Djulbegovic B (2013) Comparator bias: why comparisons must address genuine uncertainties. *J R Soc Med* 106: 30-33.
- Papageorgiou SN, Koretsi V, Jager A (2017) Bias from historical control groups used in orthodontic research: a meta-epidemiological study. *Eur J Orthod* 39: 98-105.
- Trenouth MJ (2009) Design flaws in some randomized controlled trials. *Am J Orthod Dentofacial Orthop* 135: 141.
- Harradine NWT, Gale D (2000) The effects of torque control spurs in twin-block appliances. *Clin Orthod Res* 3: 202-209.
- Parkin NA, McKeown HF, Sandler PJ (2001) Comparison of 2 modifications of the Twin-block appliance in matched Class II samples. *Am J Orthod Dentofacial Orthop* 119: 527-577.
- Lund DI, Sandler PJ (1998) The effects of Twin Blocks: a prospective controlled study. *Am J Orthod Dentofacial Orthop* 133: 104-110.
- Bland JM, Altman DG (1996) Statistical notes: measurement error. *Brit Med J* 313: 744.
- Houston WJ (1983) The analysis of errors in orthodontic measurements. *Am J Orthod* 83: 382-390.
- Mills JL (1993) Data torturing. *New Eng J Med* 329: 1196-1199.
- Trenouth MJ, Desmond SR (2013) The influence of clasp design on twin block appliance outcome-A prospective randomised, controlled clinical trial. *Cranio-View* 22: 18-34.
- Madsen H (2016) RCTs are here to stay. *Eur J Orthod* 38: 334.
- Abdulraheem S, Bondemark L (2018) Hawthorne effect reporting in orthodontic randomized controlled trials: truth or myth? Blessing or curse? *Eur J Orthod* 40: 475-479.
- Sandler J, Benson PE, Doyle P, Majumder A, O' Dwyer J, et al. (2008) Palatal implants are a good alternative to headgear: A randomized trial. *Amer J Orthod Dentofacial Orthop* 133: 51-57.
- Meikle MC (2005) What do prospective randomized clinical trials tell us about the treatment of Class II malocclusions? A personal viewpoint. *Eur J Orthod* 27: 105-114.